DOCTORADO EN INFORMÁTICA

Universidad de Oviedo
Universidá d'Uviéu
University of Oviedo

TESIS DOCTORAL

"UDLearn: Modelo de aprendizaje de máquina que facilita la toma de decisiones académicas en las instituciones de educación superior"

Presentado por
Yuri Vanessa Nieto Acevedo

TESIS DOCTORAL

“UDLearn: Modelo de aprendizaje de máquina que facilita la toma de decisiones académicas en las instituciones de educación superior”

Presentado por
Yuri Vanessa Nieto Acevedo
Para la obtención del título de Doctor en Informática

Dirigido por
Doctor D. Vicente García Díaz
Doctor D. Carlos Enrique Montenegro Marín

Oviedo, 2019

Universidad de Oviedo
Universidá d'Uviéu
University of Oviedo

RESUMEN DEL CONTENIDO DE TESIS DOCTORAL

1.- Título de la Tesis	
Español/Otro Idioma: UDLearn: Modelo de aprendizaje de máquina que facilita la toma de decisiones académicas en las instituciones de educación superior	Inglés: UDLearn: Learning machine model to facilitate academic decision making at Higher Education Institutions

2.- Autor	
Nombre: YURI VANESSA NIETO ACEVEDO	DNI/Pasaporte/NIE:
Programa de Doctorado: Doctorado en Informática	
Órgano responsable: Centro Internacional de Postgrado	

RESUMEN (en español)

Las decisiones que se llevan a cabo a nivel estratégico en las Instituciones de Educación Superior (IES), involucran las políticas, las estrategias y las acciones que afectan a toda la comunidad académica. Esta alta jerarquía administrativa enfrenta diferentes retos durante la toma decisiones, las cuales en su mayoría se basan en intuiciones o experiencias de administraciones pasadas.

La separación que existe entre los *stakeholders* y la carencia en el uso de algoritmos computacionales eficientes en las instituciones ha conllevado que el proceso de toma de decisiones: A) Sea demorado, B) No tenga en cuenta toda la información necesaria, por lo que no existe una visión completa de los diversos escenarios, C) El impacto académico sea reducido, entre otros.

En esta investigación nos enfocamos en apoyar la toma de decisiones académicas a nivel estratégico, siendo las preocupaciones académicas de los decanos respecto a la tasa de graduados, la misión preminente que vamos a apoyar. Dicho aspecto es el motivo por el cual diseñamos un modelo de toma de decisiones en las IES. Comparamos cinco algoritmos supervisados de clasificación, de lo cual planteamos una metodología de pronósticos académicos impulsado por datos. Analizamos diversas métricas de evaluación de algoritmos y finalmente propusimos UDLearn, un modelo de aprendizaje de máquina que facilita la toma de decisiones académicas a la alta jerarquía administrativa de las IES.

RESUMEN (en Inglés)

Decisions made at the strategic level of Higher Educational Institutions (HEIs) affect policies, strategies, and actions that the institutions make as a whole. The high hierarchical management faces diverse challenges during the decision-making process, that sometimes are rely on intuition or past experiences.

The disengagement of the stakeholders and the lack of using efficient computational algorithms lead to: A) The decision process takes longer, B) The “whole picture” is not involved along with all data necessary and C) Small academic impact is produced by the decision, among others.

This research focuses on supporting decision-making at strategic level, being deans’ concerns about graduation rates the preeminent mission to bolster. Therefore, we design a decision-making model at Higher Educational Institutions. We compared five supervised classification algorithms, from where we posed a methodology for data-driven academic prognosis. We analyze diverse algorithms evaluation metrics and finally, we propose UDLearn, a machine learning model to facilitate decision-making at the high hierarchy of the HEIs.

SR. PRESIDENTE DE LA COMISIÓN ACADÉMICA DEL PROGRAMA DE DOCTORADO EN INFORMÁTICA

Universidad de Oviedo
Universidá d'Uviéu
University of Oviedo

FORMULARIO RESUMEN DE TESIS POR COMPENDIO

1.- Datos personales solicitante	
Apellidos: Nieto Acevedo	Nombre: Yuri Vanessa

Curso de inicio de los estudios de doctorado	2016/2017

	SI	NO
Acompaña acreditación por el Director de la Tesis de la aportación significativa del doctorando	x	
Acompaña memoria que incluye		
Introducción justificativa de la unidad temática y objetivos	x	
Copia completa de los trabajos *	x	
Resultados/discusión y conclusiones	x	
Informe con el factor de impacto de la publicaciones	x	
Se acompaña aceptación de todos y cada uno de los coautores a presentar el trabajo como tesis por compendio	x	
Se acompaña renuncia de todos y cada uno de los coautores a presentar el trabajo como parte de otra tesis de compendio	x	

* Ha de constar el nombre y adscripción del autor y de todos los coautores asi como la referencia completa de la revista o editorial en la que los trabajos hayan sido publicados o aceptados en cuyo caso se aportará justificante de la aceptación por parte de la revista o editorial

Artículos, Capítulos, Trabajos

Trabajo, Artículo 1

Titulo (o título abreviado)	Supporting academic decision making at higher educational institutions using Machine Learning–based algorithms
Fecha de publicación	Junio de 2019
Fecha de aceptación	12 de Febrero de 2018
Inclusión en Science Citation Index o bases relacionadas por la CNEAI (indíquese)	SCI
Factor de impacto	JCR 2.784 40/106 Q2

Coautor2 X Doctor □ No doctor . Indique nombre y apellidos	Vicente García Díaz
Coautor3 X Doctor □ No doctor . Indique nombre y apellidos	Carlos Enrique Montenegro
Coautor4 X Doctor □ No doctor . Indique nombre y apellidos	Rubén González Crespo
Coautor5 □ Doctor □ No doctor . Indique nombre y apellidos	

Coautor6 □ Doctor □ No doctor . Indique nombre y apellidos	
Coautor7 □ Doctor □ No doctor . Indique nombre y apellidos	

Trabajo, Artículo 2

Titulo (o título abreviado)	Usage of machine learning for strategic decision making at higher educational institutions
Fecha de publicación	Mayo 27 de 2019
Fecha de aceptación	Mayo 22 de 2019
Inclusión en Science Citation Index o bases relacionadas por la CNEAI (indíquese)	SCI
Factor de impacto	JCR 4.098 23/155 Q1

Coautor2 X Doctor □ No doctor . Indique nombre y apellidos	Vicente García Díaz
Coautor3 X Doctor □ No doctor . Indique nombre y apellidos	Carlos Enrique Montenegro
Coautor4 X Doctor □ No doctor . Indique nombre y apellidos	Rubén González Crespo
Coautor5 □ Doctor X No doctor . Indique nombre y apellidos	Claudio Camilo González Clavjio
Coautor6 □ Doctor □ No doctor . Indique nombre y apellidos	
Coautor7 □ Doctor □ No doctor . Indique nombre y apellidos	

Trabajo, Artículo 3

Titulo (o título abreviado)	Decision-making model at Higher Educational Institutions based on Machine Learning
Fecha de publicación	Pendiente de publicación
Fecha de aceptación	Mayo de 2019
Inclusión en Science Citation Index o bases relacionadas por la CNEAI (indíquese)	SCI
Factor de impacto	JCR 0,910 72/104 Q3

Coautor2 X Doctor □ No doctor . Indique nombre y apellidos	Vicente García Díaz
Coautor3 X Doctor □ No doctor . Indique nombre y apellidos	Carlos Enrique Montenegro
Coautor4 □ Doctor □ No doctor . Indique nombre y apellidos	
Coautor5 □ Doctor □ No doctor . Indique nombre y apellidos	
Coautor6 □ Doctor □ No doctor . Indique nombre y apellidos	
Coautor7 □ Doctor □ No doctor . Indique nombre y apellidos	

En caso de compendio de un número de artículos superior a seis, se incorporarán hojas suplementarias conforme a este modelo

RESUMEN

Las decisiones que se llevan a cabo a nivel estratégico en las Instituciones de Educación Superior (IES) involucran las políticas, las estrategias y las acciones que afectan a toda la comunidad académica. Así, la alta jerarquía administrativa enfrenta diferentes retos durante la toma decisiones, las cuales en su mayoría se basan en intuiciones o experiencias de administraciones pasadas.

La separación que existe entre los diferentes *stakeholders* y la carencia en el uso de algoritmos computacionales eficientes en las instituciones ha conllevado que el proceso de toma de decisiones: A) Sea lento, B) No tenga en cuenta toda la información necesaria, por lo que no existe una visión completa de los diversos escenarios, C) El impacto académico sea reducido, entre otros.

En esta investigación nos enfocamos en apoyar la toma de decisiones académicas a nivel estratégico, siendo las preocupaciones académicas de los administradores académicos respecto a la tasa de graduados la misión preminente que vamos a apoyar. Dicho aspecto es el motivo por el cual diseñamos un modelo de toma de decisiones en las IES. Para ello, comparamos cinco algoritmos supervisados de clasificación, de lo cual planteamos una metodología de pronósticos académicos impulsados por datos. Analizamos diversas métricas de evaluación de algoritmos y finalmente propusimos UDLearn, un modelo de aprendizaje de máquina que facilita la toma de decisiones académicas a la alta jerarquía administrativa de las IES.

PALABRAS CLAVE

Máquinas de aprendizaje, Soporte de máquinas vectoriales, Redes neuronales, Árboles de decisiones, Regresión logística, Árboles aleatorios, toma de decisiones, apoyo en la toma de decisiones, instituciones de educación superior, algoritmos de clasificación.

SUMMARY

Decisions made at the strategic level of Higher Educational Institutions (HEIs) affect policies, strategies, and actions that the institutions make as a whole. The high hierarchical management faces diverse challenges during the decision-making process, that sometimes are rely on intuition or past experiences.

The disengagement of the stakeholders and the lack of using efficient computational algorithms lead to: A) The decision process takes longer, B) The “whole picture” is not involved along with all data necessary and C) Small academic impact is produced by the decision, among others.

This research focuses on supporting decision-making at strategic level, being academic administrators concerns about graduation rates the preeminent mission to bolster. Therefore, we design a decision-making model at Higher Educational Institutions. We compared five supervised classification algorithms, from where we posed a methodology for data-driven academic prognosis. We analyze diverse algorithms evaluation metrics and finally, we propose UDLearn, a machine learning model to facilitate decision-making at the high hierarchy of the HEIs.

KEYWORDS

Machine learning, Support vector machines, Artificial neural networks, Decision tress, Logistic regression, Random forests, decision-making, decision support systems, higher educational institutions, classification algorithms.

Contenido

1. INTRODUCCIÓN

La misión de las Instituciones de Educación Superior (IES) es educar a profesionales altamente competitivos que apoyen el avance de la ciudad y/o el país en el que se desenvuelven. De esta manera, la educación juega un rol importante en la reputación de un país alrededor del mundo [1]. De ahí, que la tasa de graduados se ha convertido en un indicador mundialmente aceptado que refleja el éxito de los estudiantes y el de las universidades [2].

La toma de decisiones de los altos cargos administrativos de las IES tiene grandes implicaciones sobre toda la comunidad académica. Se impacta el éxito de los estudiantes cuando las partes interesadas (altos directivos y creadores de políticas, subordinados y estudiantes) están desacopladas durante el proceso decisional.

Las mencionadas decisiones académicas se basan en su mayoría en intuiciones o experiencias pasadas [3]. Adicionalmente, se carece de eficientes algoritmos computacionales que proporcionen un propicio entendimiento del "cuadro completo" donde se evalúe información de las diferentes partes interesadas. Esto, sin contar con los obstáculos respecto a las condiciones tecnológicas que los modelos educativos presenciales aun enfrentan, tales como la adquisición de datos. Por ejemplo, a nivel administrativo la información se suele encontrar alojada en diversos silos, haciendo que los formatos de los datos varíen significativamente [4]. Incluso algunos procedimientos transaccionales como el registro de la asistencia de los estudiantes aún se hacen de forma manual en libros de papel.

Las máquinas de aprendizaje usan algoritmos que hoy en día están en creciente tendencia para el análisis de información, mostrando una excepcional capacidad para reconocer patrones y predecir resultados de diversos conjuntos de datos sin importar su ámbito de procedencia. Mayormente las investigaciones realizadas con máquinas de aprendizaje, se han enfocado en algoritmos supervisados, dado su fortaleza en la producción de modelos incorporables en proceso de toma de decisiones [5]. Así mismo, nuestra investigación está basada en algoritmos supervisados ya que su pronóstico depende en un conjunto de datos y los eventos que han tenido lugar sobre estos datos en el pasado. De esta manera, se correlacionan un valor monitoreado a un valor objetivo [6], apoyando así la confiabilidad del modelo.

Se seleccionaron cinco clasificadores como lo son: Redes neuronales (RN), Soporte de máquinas vectoriales (SMV), Árboles de decisiones (AD), Bosques aleatorios (BA), y Regresión logística (RL). Estos algoritmos supervisados han sido los más estudiados y relevantes en trabajos similares y se escogieron dada su alta precisión y eficiencia como técnicas de clasificación en el ámbito académico [7][8]. Se consideraron simultáneamente estos algoritmos desde una perspectiva comparativa. Para evaluar su desempeño se utilizaron diversas métricas: F1score, Precisión, Memoria (Recall), y Área bajo la curva (AUC), dada su popularidad para clasificar el desempeño en clases supervisadas [9]. Una vez realizado el análisis y comparación de estos, se propuso, respecto a las variables y características evaluadas, un algoritmo de máquina de aprendizaje basado en la técnica de clasificación de Soporte de máquinas vectoriales.

Aunque existen varios trabajos similares que utilizan algoritmos de máquinas de aprendizaje para identificar estudiantes, nuestra investigación difiere de las existentes en: A) Los datos seleccionados en la caso de estudio provienen de una IES con un modelo educativo presencial, más de 10 años de ventana de observación y una decanatura completa de cinco programas académicos, B) Se incluyen en la arquitectura de los algoritmos diversas y numerosas características de los datos, lo que conlleva a alcanzar una precisión global mayor, C) Ni las partes interesadas, ni el objetivo de apoyar a los altos cargos administrativos de las IES han sido evaluados antes. El objetivo de nuestro trabajo es apoyar a los altos directivos de las IES durante la compleja labor de toma de decisiones, a través del desarrollo de un modelo que incluya algoritmos computacionales altamente confiables para este propósito.

Para determinar la problemática más preminente que inquieta a los directivos de las IES durante la toma de decisiones, se realizó una encuesta en la que participaron vice rectores, directivos y decanos de universidades tanto públicas como privadas de más de 40 universidades en todo Colombia. Se identificó que la mayor problemática para ellos corresponde a la tasa de graduación de estudiantes, lo cual se convirtió en el motor de nuestro trabajo. Es así como nos enfocamos en segmentar a los estudiantes a través de la clasificación de su desempeño académico, evaluando el tiempo en que se esperaba se graduasen los estudiantes. De esta manera se buscó una variable booleana (x,y) en la que la predicción se centra en determinar si un estudiante se graduará o no.

De la detallada información que se obtiene del modelo, se pueden llevar a cabo diferentes acciones como: priorizar los esfuerzos y las intervenciones a los estudiantes, aumentar la tasa de retención de estudiantes, crear estrategias que mitiguen la reprobación prematura de ciertas asignaturas, disminuir las tasas de deserción, crear planes de aprendizaje a futuro, incrementar la calidad respecto a los indicadores de la IES, entre otros.

2. OBJETIVOS

El objetivo principal de este trabajo es determinar qué modelo de aprendizaje de máquina (en base a la comparación de diferentes algoritmos) facilita la toma de decisiones académicas a la alta jerarquía administrativa de las Instituciones de Educación Superior.

El objetivo principal se divide en los siguientes objetivos específicos:

- **Determinar las decisiones académicas en las que la alta jerarquía administrativa de las IES requiere apoyo.** Se debe intervenir en la toma de decisiones a nivel estratégico, ya que las decisiones a este nivel generan mayor impacto en la comunidad académica. Para esto, se debe analizar y entender el gobierno institucional y sus necesidades a la hora de tomar decisiones. El desarrollo de una encuesta estructurada a decanos y directores de programas en facultades de ingeniería de todo Colombia, junto con el análisis estadístico de resultados, deslumbran sus preocupaciones respecto a las tasas de graduados, ya que esta métrica es crucial a la hora de evaluar el éxito de los estudiantes y de la institución.

- **Presentar una metodología impulsada por datos para pronósticos académicos**. La metodología general de la investigación se construye bajo los estándares de la minería de datos, como un proceso cíclico con un enfoque en el que la labor fluye de manera tal que en cada paso se pueda tomar acción sobre los datos analizados, se mejore el entendimiento de la visión general de la academia, se reduzcan tiempos y se mejore la confiablidad en las predicciones.

- **Proponer un marco de toma de decisiones académicas para las Instituciones de Educación Superior.** Se deben evaluar algoritmos de máquinas de aprendizaje que se adecuen a la problemática identificada, analizando los resultados bajo diversas métricas de evaluación. Teniendo en cuenta el algoritmo que presenta el mejor desempeño, se construye un modelo que prediga la tasa de graduados. Se presentan los parámetros y pasos a través de un marco de toma de decisiones que sirve como referencia para los directores y altos administrativos de las IES. Se debe facilitar la interacción persona-computadora, por lo cual los resultados se presentan de manera amigable a través de una herramienta Web que apoya a los directivos y altos administrativos de las IES en la toma de decisiones. Se presenta la arquitectura Web de dicho prototipo, así como los resultados que brinda.

3. PROPUESTA

Inicialmente, esbozamos la metodología propuesta para el pronóstico académico e indicamos los pasos requeridos para tomar acción sobre los datos. Se presenta el modelo propuesto para la toma de decisiones, el algoritmo de clasificación, así como todo el desarrollo teórico y tecnológico de la investigación.

3.1 INTRODUCCIÓN

El modelo propuesto se enfoca en apoyar la toma de decisiones académicas a través del pronóstico del número de graduados, lo cual va a permitir que los altos cargos administrativos de las IES tomen acción sobre los datos obtenidos, así como un mejor entendimiento cuando se tomen decisiones estratégicas respecto a la tasa de retención y deserción de estudiantes, planeación de recursos, diseño de currículos y administración de recursos y talentos, entre otros. Adicionalmente, este enfoque reduce tiempo de procesamiento, mejora la confiabilidad de las predicciones y promueve la formulación y entendimiento de patrones ocultos.

La metodología general que se utilizó en esta investigación, y que se representa en la Ilustración 1, se creó sobre el estándar del proceso cíclico de minería de datos y enfoque impulsado por datos, como una serie de pasos a lo largo de un flujo de trabajo [6].

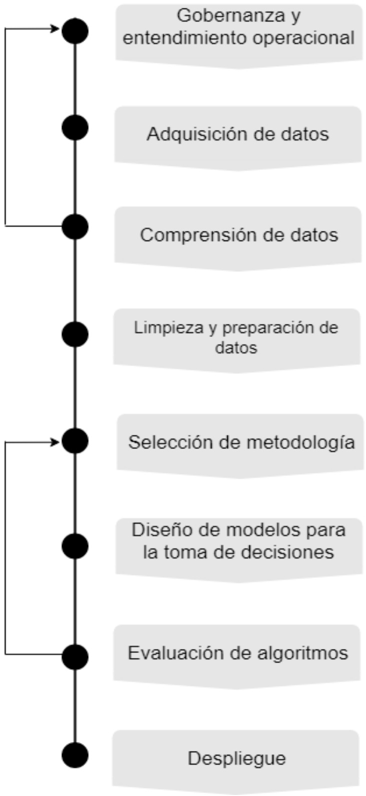

Gobernanza y entendimiento operacional
Adquisición de datos
Comprensión de datos
Limpieza y preparación de datos
Selección de metodología
Diseño de modelos para la toma de decisiones
Evaluación de algoritmos
Despliegue

Ilustración 1: Metodología impulsada por datos para pronósticos académicos

1. Gobernanza y entendimiento operacional: Se deben establecer los objetivos de la predicción académica identificando las necesidades de las partes interesadas y el posible algoritmo a ejecutar. Se debe analizar el ambiente en el que se desarrolla la IES y entender cómo afectan las políticas externas a sus decisiones. Es necesario examinar la actual situación académica, las estrategias, procesos, personas a cargo, tiempos y recursos.
2. Adquisición de datos: La fuente de datos suple las necesidades de cualquier propósito de análisis. Las IES con un modelo de educación presencial enfrentan varias dificultades en la adquisición de datos para la toma de decisiones. Los datos deben ser recolectados, integrados y almacenados en una base de datos.
3. Comprensión de datos: De acuerdo con el modelo educativo de la IES y sus procesos operacionales, la organización de sus datos, así como su formato y almacenamiento se realiza de diferentes maneras. Los datos deben ser revisados y descritos dentro de un contexto. El pronóstico depende de la cantidad y calidad de los datos.
4. Limpieza y preparación de datos: La selección, limpieza y formato de datos se debe realizar teniendo en miras a la selección de características importantes que se buscan en los datos. Se debe fijar la infraestructura y administración de almacenamiento de los datos.
5. Selección de metodología: El algoritmo de máquina de aprendizaje se selecciona en este paso. Su funcionalidad central es aprender de datos del pasado y generalizarlos al futuro. Si la variable de salida es categórica se debe utilizar un algoritmo de clasificación, de lo contrario se debe utilizar un algoritmo de regresión.
6. Diseño de modelos para la toma de decisiones: Se aplican los datos preparados previamente al algoritmo de aprendizaje seleccionado. Dependiendo del algoritmo seleccionado en el paso anterior, algunos pasos extra como la normalización de datos se deben llevar a cabo. En este paso se realiza el entrenamiento, evaluación y validación del algoritmo.
7. Evaluación de algoritmos: Se realiza una comparación de los diferentes algoritmos para escoger el más apropiado respecto a los objetivos del pronóstico. La validación del modelo puede ser estimada a través de diversas métricas de eficiencia.
8. Despliegue: Las técnicas de interacción persona-computadora deben considerarse con el fin de presentar la información a los directivos de las IES de manera amigable y fácil de entender. Los reportes de las predicciones deben estar disponibles en una herramienta online definida para cada perfil de usuario. El principal resultado de este paso es apoyar de manera confiable las decisiones, inducir a la formulación de hipótesis y proveer un mayor entendimiento de la situación de la IES.

En la Ilustración 2 se presenta la propuesta de un marco de toma de decisiones, esbozando uno de los objetivos centrales de este estudio. El enfoque, que fluye de abajo hacia arriba, representa la integración de los humanos con la inteligencia computacional manejada dentro de tres fases centrales.

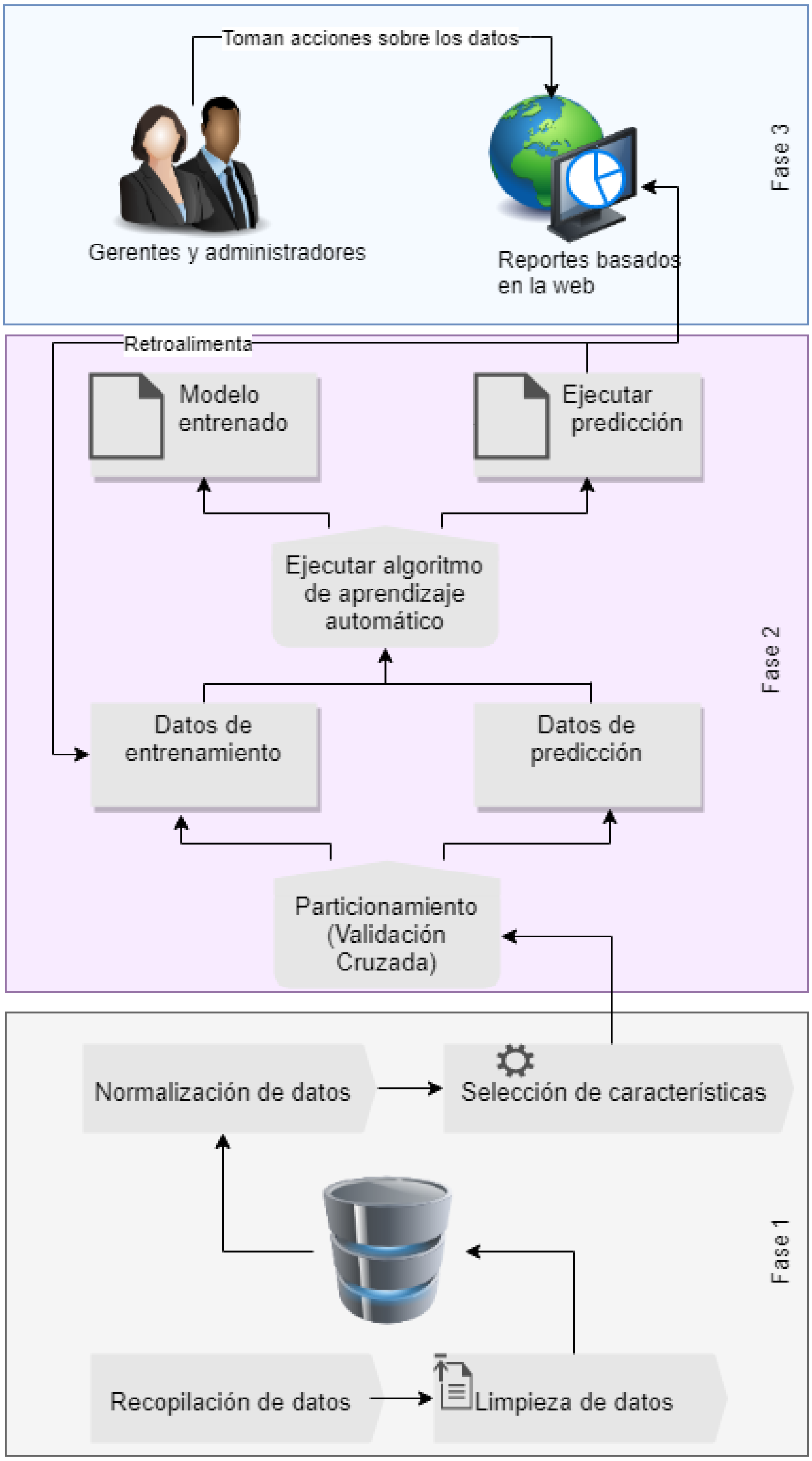

Toman acciones sobre los datos
Gerentes y administradores
Reportes basados en la web
Fase 3
Retroalimenta
Modelo entrenado
Ejecutar predicción
Ejecutar algoritmo de aprendizaje automático
Fase 2
Datos de entrenamiento
Datos de predicción
Particionamiento (Validación Cruzada)
Normalización de datos
Selección de características
Fase 1
Recopilación de datos
Limpieza de datos

Ilustración 2: Marco propuesto para la toma de decisiones académicas en las IES

La primera fase requiere la adquisición y limpieza de datos. Los modelos educativos a distancia y mixtos dependen de un sistema de información centralizado lo cual difiere del modelo educativo presencial, que enfrenta obstáculos a la hora de recolectar datos. La descripción y valores de las características utilizadas en el modelo se encuentran disponibles en la Tabla 1 de este documento.

En la segunda fase se construye el algoritmo de máquina de aprendizaje para predecir el número de estudiantes a graduarse. Tras realizar la comparación y el análisis de cinco robustos algoritmos de clasificación supervisada, se eligió el Soporte de máquinas vectoriales como la técnica de clasificación para la construcción del modelo, no solo por su superior resultado en la comparativa sino por su:

- Precisión superior y resultados compactos y comprensibles [10].
- Proporción de los mejores límites de decisión para problemas de clasificación [11][12].
- Precisión utilizando características académicas similares [8][13].
- Menor propensión al sobreajuste que otros modelos [14].

Los pasos principales para la construcción del algoritmo son los siguientes:

1. Normalizar los datos dándoles valores entre [0-1] utilizando la transformación: $p^n = (p - p^{mean})./p^{std}$ donde p^{mean} es el promedio de los vectores de entrada del conjunto de daos, y p^{std} es el vector que contiene la desviación estándar de cada elemento de los vectores de entrada.
2. Establecer la precisión deseada (ej. β >83%)
3. Establecer los parámetros iniciales del Soporte de Maquina vectorial
    a. Establecer el tipo de soporte de maquina vectorial (ej. C-SVC)
    b. Establecer el tipo de función del kernel (ej. Función básica radial)
    $K(x_m, y_m) = exp(-\gamma \|x_m, y_m\|^2)$ donde γ >0
    Parámetros del kernel:
    i. Establecer el gamma en la función del kernel (ej. γ =1.0)
    ii. Establecer C (ej. C =10.0)
4. Particionar los datos en dos subconjuntos: 70% para entrenamiento y 30% para predicción utilizando la técnica de muestreo estratificado para obtener un grupo homogéneo.
5. Particionar el subconjunto de entrenamiento del paso (4) en 5 pliegues de grupos que más adelante se ejecutarán como 5-validacion cruzada de pliegues.
6. Computar el algoritmo con los parámetros iniciales del paso (3) buscando por $K(x_m, y_m)$ donde (x_m, y_m) son variables booleanas de acuerdo con la formulación del problema original (ej. X = graduado, y = no graduado).
7. Almacenar la precisión obtenida en cada k-pliegue, así como los valores de γ y C que se ingresaron en cada ejecución.

8. Repetir el paso (6) hasta encontrar los mejores parámetros para γ y C
    a. Ajustar los parámetros en cada ejecución hasta que la exactitud sea satisfactoria para los investigadores.
9. Computar el algoritmo del paso (6) para entrenar el conjunto de entrenamiento en su totalidad, usando los valores obtenidos para γ y C.
10. Computar $K(x_m, y_m)$ en el paso (5) en el conjuto de predicción.
11. Si la exactitud deseada se alcanza, se aprueba y se guarda el ultimo resultado obtenido, si no existe dicho umbral, se reduce β .

Finalmente, en la tercera fase se entregan los resultados de la predicción en una plataforma Web a los directivos de las IES, de manera que sea fácil de acceder y entender.

3.2 TOMA DE DECISIONES EN LAS INSTITUCIONES DE EDUCACIÓN SUPERIOR

Las IES son un tipo particular de organización del sector terciario de la economía. Ellas tienen la responsabilidad principal de la gobernanza y la administración de sus finanzas, actividades y personal, conservando así su autonomía para decidir sobre sus operaciones organizativas. La forma en cómo se toman las decisiones, las prioridades institucionales, las estrategias, los objetivos y las asignaciones de recursos son típicamente funciones del gobierno institucional. Así, se observan tres tipos de gobierno [15].

1. Académico: Los miembros de la facultad tienen la autoridad para decidir en áreas de profesorado, investigación, academia, currículos y administración.
2. Burocrático: La universidad conserva capas jerárquicas con divisiones de trabajo caracterizadas por procedimientos, administraciones fijas y órdenes directas de líderes superiores.
3. Corporativo: Al observar la educación como un servicio, los estudiantes se convierten en los clientes centrales [2]. De esta manera las universidades siguen la práctica de las empresas para resaltar las necesidades del cliente y la competencia en el mercado.

En la práctica, la gobernanza universitaria es una variación y combinación de estas tres categorías, las cuales trabajan mancomunadamente en beneficio del éxito de los estudiantes. De ese modo, las IES llevan a cabo sus decisiones para asegurar dicho éxito. Es por esto por lo que, para clasificar las decisiones en las IES, consideramos una estructura piramidal jerárquica (Ilustración 3), que divide las decisiones de acuerdo con el trabajo y las responsabilidades de los colaboradores.

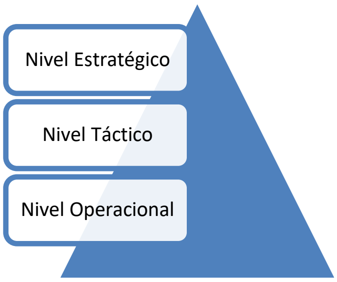


Ilustración 3: Niveles de estructura en las decisiones de las IES

1. Nivel estratégico: Este nivel define las políticas y estrategias de la organización integrando los objetivos y las acciones a realizar como un todo. Son el nivel más ambicioso de la planeación estratégica y sus cargos se representan por el consejo superior, el rector, directores y decanos.
2. Nivel táctico: Identifica y ejecuta al detalle los planes realizados a nivel estratégico. Se encargan de su implementación, ejecución y control.
3. Nivel operacional: El nivel base de la cadena se encarga de los procesos día a día y a través de su trabajo se sostiene toda la estructura. Tareas específicas y actividades transaccionales son desarrolladas para apoyar las operaciones de la institución.

Se observa que cada uno de los niveles de la pirámide tiene un proceso de toma de decisiones, que de arriba hacia abajo afecta significativamente a una mayor porción de la comunidad académica. Aunque algunas aplicaciones de software apoyan estas decisiones [16][17], los altos cargos generalmente trabajan con la información que proveen los niveles operacionales [18] [19]. Sin embargo, estos datos no son analizados y visualizados de forma fácil para apoyar la toma de decisiones a los altos directivos.

3.3 NECESIDADES DE LOS DIRECTIVOS DE LAS IES EN LA TOMA DE DECISIONES

Se llevó a cabo una encuesta para entender, de primera mano y de manera profunda, cuáles eran las inquietudes y problemáticas a las que se enfrentaban los directivos de las IES durante el proceso de toma de decisiones. La encuesta se llevó a cabo en diferentes pasos. Primero entrevistamos a tres directores de facultad de diversos modelos educacionales (presencial, virtual y mixto) para que nos dieran una primera visión de los grupos de decisiones académicas sobre las que ellos decidían (de ahora en adelante a estos grupos les llamamos categorías). Con el apoyo de la Red de Programas de Ingeniería de Sistemas y afines (REDIS) se hizo el envío de esta encuesta a vicerrectores, directores de departamento y decanos de todo Colombia. No obstante, antes de hacer el envío de estas encuestas, el diseño de las mismas contó con la colaboración y revisión de directivos de las IES en Bogotá [20].

En la base de datos del Ministerio de Educación en Colombia [21], encontramos que las universidades que enseñan ingeniería de computación como un programa profesional son 74, por lo que este se convirtió en el tamaño de la población ($\mathrm{N} = 74$) y calculamos un tamaño de la muestra de $\mathrm{n} = 42$ para obtener un nivel de confianza de $\mathrm{C} = 95\%$. Así, la encuesta fue contestada por 42 vice rectores, directivos, y decanos de facultades de ingeniera de universidades de todo el país. Usando el paquete SPSS [22], se estimó que el coeficiente Alpha Cronbach fue del $\alpha = 0{,}906$, resultando así un estudio altamente confiable.

De los resultados obtenidos, se encontró que sus inquietudes a nivel académico eran mayormente la tasa de graduados y el desempeño académico de los estudiantes. Otros hallazgos reflejan algunos de los factores que pueden afectar las tasas de graduación tales como los procesos administrativos por lo que los estudiantes deben pasar para obtener el título.

3.4 VALIDACIÓN EN UN CASO DE ESTUDIO

Para ilustrar la metodología en un caso de estudio real, se llevó a cabo un experimento usando datos reales provenientes de la Facultad de Ingeniería de la Universidad Distrital Francisco José de Caldas [23]. Esta es una universidad pública y presencial en Colombia, contando su Facultad de Ingeniería con más de 6000 estudiantes, 400 profesores, 5 programas de grado profesional y 10 programas de maestría y doctorado. El primer conjunto de datos incluía estudiantes de grado matriculados dentro de los años 2004-2014 (con un total de 12477 estudiantes). Una vez se realizó la limpieza de datos, se tomaron los datos de 6103 estudiantes con 55220 registros.

En la fase de recolección de datos, dado la política de privacidad de la información de la universidad, la información fue limitada, lo cual se sobrellevó derivando y calculando atributos extras provenientes de los datos dados. La Tabla 1 muestra las características de los datos que se utilizaron para construir el modelo.

Tabla 1: Características de los datos ingresadas al modelo

ID	Nombre	Descripción	Medida
Datos de entrada provisto por la administración de la universidad			
1	Estrato residencia	Estrato socio económico de acuerdo con la dirección de residencia del estudiante	{1,2,3,4,5,6}
2	Promedio de las materias cursadas	Promedio aritmético de las notas finales de las asignaturas cursadas	Nota entre 0,0 y 5,0
3	Semestre	Semestre actual en el que se encuentra matriculado	{1,2,3,4,5,6,7,8,9,10,11,12}

4	Cantidad de materias cursadas en un semestre	Materias matriculadas por semestre	Cantidad
5	Materias aprobadas por semestre	Materias satisfactoriamente aprobadas en un semestre	Cantidad
6	Materias reprobadas por semestre	Materias reprobadas en cada semestre matriculado	Cantidad
7	Materias validadas por semestre	Materias que fueron reprobadas, pero tras presentar un examen sumatorio se convirtieron en aprobadas	Cantidad
Datos de entrada calculados a partir de los datos dados y que fueron incluidos en el estudio			
8	Mediana de la nota	Valor de la nota separando la mitad más alta de las notas de la mitad más baja	Nota entre 0,0 y 5,0
9	Máxima nota	Máximo valor de la nota de todas las notas finales	Nota entre 0,0 y 5,0
10	Mínima nota	Mínimo valor de la nota de todas las notas finales	Nota entre 0,0 y 5,0
11	Rango de las notas	Diferencia entre la nota más alta y la nota más baja	Cantidad
12	Primer cuartil de notas	La nota media entre la nota más baja y la mediana del conjunto de datos	Cantidad
13	Segundo cuartil de notas	Nota mediana	Cantidad
14	Tercer cuartil de notas	La nota media entre la nota más alta y la mediana del conjunto de datos	Cantidad
15	Desviación estándar de las notas	Raíz cuadrada de la nota observada menos la nota promedio al cuadrado	Cantidad
16	Cantidad de materias cursadas	Sumatoria de todas las materias matriculadas durante el total de la carrera	Cantidad
17	Cantidad de	Sumatoria de todas las materias	Cantidad

	materias aprobadas	satisfactoriamente aprobadas durante toda la toda la carrera	
18	Cantidad de materias reprobadas	Sumatoria de todas las materias reprobadas durante toda la toda la carrera	Cantidad
19	Cantidad de materias validadas	Sumatoria de todas las materias que durante la carrera fueron aprobadas una vez se realizó el examen de validación	Cantidad

3.5 ALGORITMOS DE MÁQUINAS DE APRENDIZAJE

Se escogieron, analizaron y compararon los cinco algoritmos de clasificación con mayor relevancia y mejores resultados en el ámbito de la educación [8][7][24][25]. Con el fin de obtener una mejor predictibilidad, consideramos los más robustos como son: Redes neuronales, Soporte de máquinas vectoriales, Árboles de decisiones, Bosques aleatorios y Regresión logística. Posteriormente, una vez usadas las métricas de efectividad definidas en la Sección 4.1, se identificó y escogió el algoritmo que mejor predecía la cantidad de estudiantes por graduarse y no graduarse.

1. Redes neuronales:

Una red neuronal artificial es una abstracción simple de una red neuronal biológica. Está construida matemáticamente y diseñada para realizar tareas específicas [26]. Las redes neuronales son el método de referencia en el contexto de la educación porque son adecuadas para tareas comunes en este campo, como la clasificación, el reconocimiento de patrones y la predicción.

Trabajamos específicamente con una red neuronal supervisada, diseñada con dos neuronas de salida (graduadas o no graduadas) y tres capas. La función de activación y transferencia utilizada es un sigmoide tangente hiperbólico conocida también como tansig (tanh). Esta función generalmente se prefiere a la función de transferencia logística en las capas ocultas debido a que las entradas están normalizadas. Por otro lado, la función de transferencia utilizada en la capa de salida es softmax [27]. Esta es la función ofrecida por knime [28] (que es la plataforma analítica que se ha utilizado en el trabajo) en la capa de salida. La función de transferencia tiene esta forma:

$$a_i = f(n_i) = \exp(n_i) \Big/ \sum_{j=1}^{x} \exp(n_i)$$

La función softmax mostrada arriba, donde n es un vector de entradas y la capa de salida, y_j donde $j = 1,2,3, \dots s$. Las salidas de la función de transferencia de softmax se pueden interpretar como las probabilidades asociadas con cada clase. Cada salida estará entre el rango [0-1] y la suma de las salidas será igual a 1 [27]. La arquitectura de la red desarrollada en la investigación se muestra en la Ilustración 4.

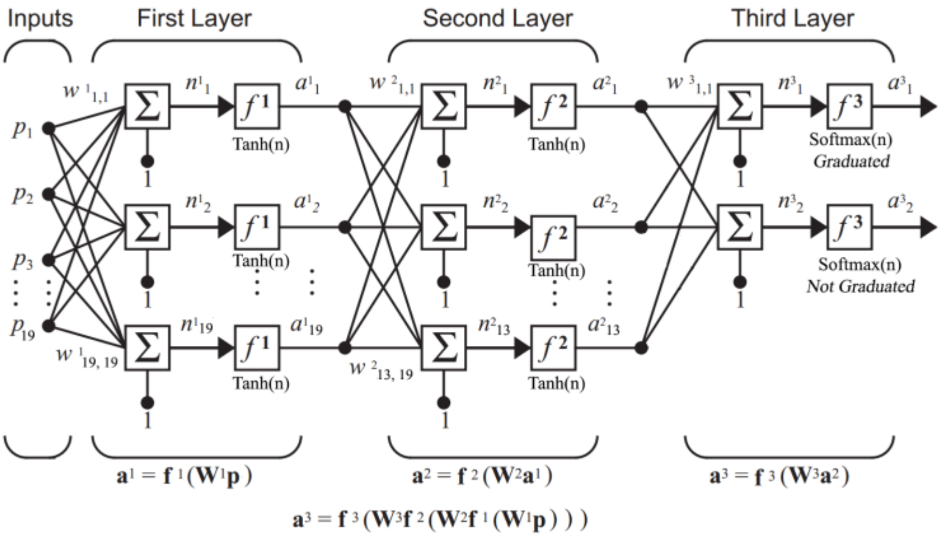


Ilustración 4: Arquitectura de la Red neuronal

La estructura de nuestro modelo de Red neuronal consta de tres capas: la primera capa, que es también la capa de entrada, la segunda capa o capa oculta y la tercera capa o capa de salida. Cada capa tiene su propia matriz de peso W. En nuestro caso, este valor viene dado inicialmente por knime. Para distinguir entre estas capas, agregamos el número de la capa como superíndice. Por lo tanto, la matriz de ponderación para la primera capa se escribe como W^1, la segunda capa como W^2 y la tercera capa como W^3. Así, W^1 (19,19) es el elemento para la neurona 19, característica 19. Además, cada capa tiene un vector de entrada neta n y un vector de salida a.

La primera capa tiene 19 entradas (características) representadas en un vector de entrada p. Cada uno de estos elementos está conectado a cada neurona, por lo que la primera capa, que también es la capa de entrada, tiene 19 neuronas. Como la literatura de revisión normalmente recomienda [29], las neuronas de capa oculta son 2/3 del tamaño de la capa de entrada. Así, nuestro caso corresponde a 13 neuronas para la segunda capa. Finalmente, el vector de salida se toma como una tercera capa con 2 neuronas, correspondiente a la predicción buscada (graduado o no graduado).

2. Soporte de máquinas vectoriales:

Es una técnica de clasificación ampliamente utilizada [30] que, dado un conjunto de objetos pertenecientes a una de dos categorías, construye un hiperplano en un espacio dimensional mayor, separando esas categorías. Las funciones tradicionales del kernel incluyen series lineales, polinómicas, de Gauss, de sigmoid y de Fourier. La elección del kernel altera bruscamente la naturaleza del límite de decisión. Cabe destacar que el kernel de Gauss (formulación radial) sobresale debido a su simplicidad, alta eficiencia, fácil acceso y menos cómputo [12], por lo que fue el kernel seleccionado.

La función de base radial es un algoritmo de aprendizaje kernel usado comúnmente en Soporte de máquinas vectoriales. No se define explícitamente su integración por lo que opera en un espacio dimensional infinito. Tiene la forma:

$$K(x_m, y_m) = exp(- \gamma \| x_m, y_m \|^2)$$

Donde $\gamma > 0$ es el parámetro que controla el radio de la función de base, que sirve para controlar la dispersión del kernel en el espacio de entrada. La función del kernel $\mathrm{K(x_m,y_m)}$ pretende medir la "similitud" entre $\mathrm{x_m}$ e $\mathrm{y_m}$ (cuanto más grande, más similar). En términos de la distancia euclidiana cuadrada, asumimos qué tan cerca están esos puntos entre sí. El proceso de clasificación binaria (es decir, estudiantes graduados y no graduados) implica adaptar un hiperplano (límite de decisión) a solo dos clases separables. Así, el hiperplano se convierte en una línea recta que separa dos clases [31]. La extensión de knime para Soporte de máquinas vectoriales, así como la validación cruzada permitió encontrar los mejores valores para γ (ej. $\gamma = 0{,}1$) y C (ej. $\mathrm{C} = 10$). SMV es la técnica de clasificación elegida debido a su exactitud superior, a los resultados compactos [10] y a que ofrece un mejor límite de decisión en cualquier problema de clasificación [11]. Su gran precisión se obtuvo utilizando características académicas similares [8], siendo, además, menos propensos al sobreajuste que otros modelos.

3. Arboles de decisiones

La estructura jerárquica de árbol se asemeja a una forma humana de toma de decisiones, proporcionando información extendida sobre la secuencia para clasificarla individualmente en una clase, descubriendo las reglas de una manera más comprensible [32].

Aunque hay muchos algoritmos de árbol de decisión específicos (ej. ID3, C4.4, C5.0, CART y CHAID), trabajamos con el más popular desarrollado por Quinlan [33]. C5.0 es significativamente más rápido y eficiente que sus predecesores C4.5 e ID3. C5.0 admite un potenciador (boosting), lo que les da a los árboles una mayor exactitud. Además, permite la ponderación de diferentes atributos y tipos de clasificación errónea y separa los datos automáticamente para ayudar a reducir el ruido. El índice de Gini se usa como una medida de división más apropiada para cada nodo [34]. La función de división tiene esta forma:

$$I_g\,(p) = 1 - \sum_{i=1}^{J} p_i^2$$

Donde J clases, suponen que $\mathrm{I} \in \{1, 2, 3, \ldots \mathrm{J}\}$ y $\mathrm{p_i}$ es la fracción de elementos etiquetados con la clase "i" en el conjunto. El conjunto de datos se divide en dos subconjuntos. La división se puede hacer en el valor más alto de la partición más baja, en el valor más bajo de la partición más baja, o en el promedio de las dos. Elegimos estos últimos puntos de división para forzar al modelo solo divisiones binarias en columnas nominales. Por otro lado, el método de poda utilizado es la Longitud de descripción mínima, que reduce la tasa de errores de clasificación, proporciona alta exactitud y rápidos tiempos de ejecución.

4. Bosques aleatorios:

Este clasificador combina el desempeño de numerosos árboles de decisiones para predecir el valor de una variable [35]. Cada árbol en el bosque da una clasificación y "votos" para esa clase. El bosque elige la clasificación que tenga más votos (sobre todos los árboles en el bosque). El predictor de regresión de BA tiene la forma:

$$f_{rf}^{K}(x) = \frac{1}{K} \sum_{K=1}^{k} T(x)$$

Cuando el BA recibe $\mathrm{y(x)}$, que es el vector de entrada, compuesto de los valores de diferentes características evidénciales analizadas para un área de entrenamiento determinada, BA crea un número K de árboles de regresión y promedia los resultados. Después se cultivan, tales árboles $\mathrm{T(x)1^K}$, los cuales van creciendo. Sobre la arquitectura utilizada, el atributo de división elegido fue Gain Radio [36], ya que no representó una diferencia significativa entre las otras opciones proporcionadas por knime (Information Gain, Information Gain Radio y Gini Index). Además, el número de modelos utilizados fue de cien, principalmente debido al tamaño del conjunto de datos y los resultados de las pruebas. El número de árboles de decisión para aprender del número de modelos establecidos también ha sido de cien. Además, se requiere el uso de semillas aleatorias estáticas para comenzar la predicción, y se utilizó el generado automáticamente por el software (es decir, 1508210392822).

5. Regresión logística:

Se utiliza para estimar valores discretos (por ejemplo, valores binarios) basados en un conjunto dado de variables independientes. Se conoce como regresión logit porque predice la probabilidad de una ocurrencia, de cualquier evento, al ajustar los datos en una función logit [37]. La regresión logística da límites de clase lineales. Debido al hecho de que utiliza una curva en forma de 'S' en lugar de una línea recta, es un ajuste natural para dividir los datos en grupos. La columna objetivo fue (graduado y no graduado) con verdadero como categoría de referencia. El gradiente promedio estocástico se usó como solucionador, principalmente porque minimiza el negativo de la función de probabilidad de registro y admite la regularización. Se basa en la idea del método de descenso de gradiente, lo que significa que en cada interacción el algoritmo se mueve en una dirección descendente del negativo de la función Log Probabilidad con un tamaño de paso Δ, que se denomina tasa de aprendizaje. La estrategia de velocidad de aprendizaje se fijó en un tamaño de paso de 0,2. Finalmente, las condiciones de terminación se establecieron en un número máximo de épocas (es decir, 200) y Epsilon (es decir, $\varepsilon = 0{,}001$).

3.6 Arquitectura de software de la herramienta web propuesta

Los resultados de la predicción son difíciles de acceder para los directores de las IES dado que deben tener los conocimientos técnicos para interpretar dichos resultados en los programas o plataformas donde se lleva a cabo el análisis. Además, los resultados presentados por estas plataformas se dan en tablas de datos difíciles de analizar y brindan la posibilidad de mostrar únicamente algunos gráficos como los diagramas de tortas, que no son fáciles de visualizar, comprender y gestionar.

Por lo tanto, vamos más allá del modelo en sí, presentando un conjunto de herramientas basadas en la Web para directivos académicos que permite predecir el número de estudiantes graduados y no graduados; exponiendo los resultados de una manera amigable y fácil de entender en dicha interfaz. El principal desafío es permitir la carga de nuevos datos y predicciones y hacerlos transparentes para los usuarios.

Aunque el algoritmo de aprendizaje automático todavía se ejecuta en knime, en nuestro prototipo proponemos una arquitectura de tres capas para integrar la infraestructura. Es importante resaltar que esta herramienta Web ya ha sido usada por los directivos de la universidad caso de estudio, quienes han expresado su satisfacción no solo con la utilidad y fácil uso de la herramienta sino con la accesibilidad y amigabilidad de la interfaz. Según la descripción del director de la facultad de Ingeniería, la aplicación “es apropiada para la retención estudiantil y estimación de graduados” debido a “la forma fácil en la que se entiende la información que no había sido desvelada antes”.

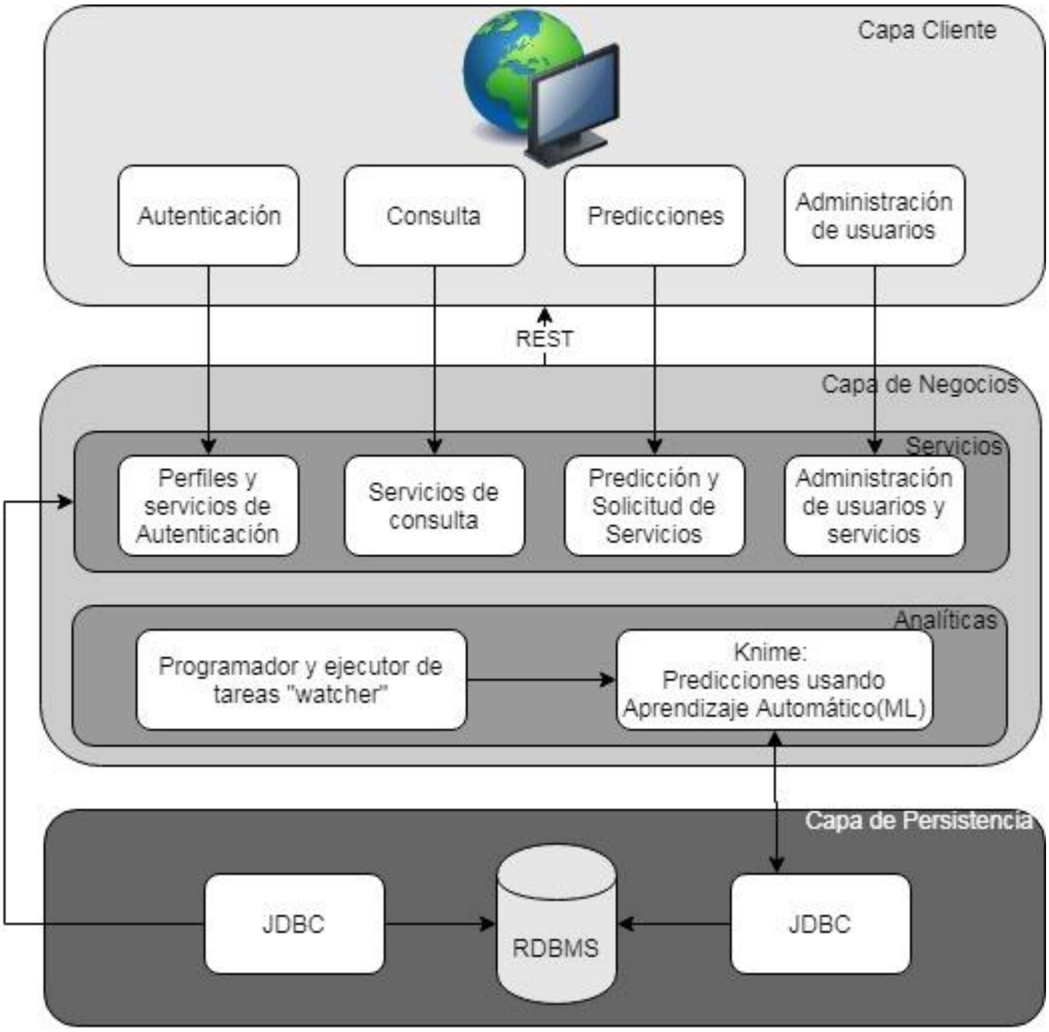

Capa Cliente
Autenticación
Consulta
Predicciones
Administración de usuarios
REST
Capa de Negocios
Servicios
Perfiles y servicios de Autenticación
Servicios de consulta
Predicción y Solicitud de Servicios
Administración de usuarios y servicios
Analíticas
Programador y ejecutor de tareas "watcher"
Knime: Predicciones usando Aprendizaje Automático(ML)
Capa de Persistencia
JDBC
RDBMS
JDBC

Ilustración 5: Arquitectura general del prototipo propuesto [38]

En este prototipo integramos la infraestructura de la aplicación en una arquitectura de tres capas:

1. Capa de cliente: Después de un registro seguro, el kit de herramientas Web ofrece cuatro módulos funcionales: A) autenticación (administrar la configuración del perfil y la actualización), B) consultar (permitir la visualización de las solicitudes realizadas, como tablas, filtros y gráficos representados desde la capa empresarial), C) predicciones (administrar la solicitud de predicción y obtener los datos) y D) gestión de usuarios (permitir crear, actualizar y eliminar usuarios). Los módulos acceden a la capa empresarial a través del esquema REST basado en el protocolo HTTP, que forma parte de la capa empresarial. El desarrollo se realiza en Java utilizando las bibliotecas React y Flux [39] para crear la interfaz de usuario. La biblioteca d3 [40] se utiliza para visualizar gráficos.
2. Capa de negocios: Comprende la lógica de negocios y está representada en dos subcapas: la capa de servicios y la capa de análisis.

    - Subcapa de servicio: Se divide en módulos que exponen los cuatro servicios de acuerdo con su funcionalidad. La asignación de objeto relacional Hibernate [41] aplica la interfaz de persistencia de Java JPA [42]. Liquibase [43] se utiliza como un sistema migratorio para completar la información base en la capa de persistencia.
    - Subcapa de análisis: la aplicación de servidor llamada "observador" usa la extensión Spring-Schedule para ejecutar una tarea de instancia periódicamente desde knime [28]. Este último contiene los flujos de trabajo para procesar las funcionalidades de entrenamiento y predicción. La comunicación entre estos dos componentes se realiza con líneas de comando que llaman en la instancia de knime y especifican el flujo de trabajo correspondiente que se ejecutará.
3. Capa de persistencia: Utiliza MySql [44] como el Sistema de Gestión de Base de Datos Relacional. Las fuentes de datos suministran la información requerida en el kit de herramientas. Los usuarios pueden subir nuevos datos en formato CSV. Se proporciona una plantilla para garantizar que los datos contengan todas las funciones necesarias para ejecutar la predicción.

4. RESULTADOS

En este capítulo se presenta la comparación del rendimiento de los algoritmos, así como la herramienta Web que facilita la visualización de resultados.

4.1 COMPARACION DE ALGORITMOS DE CLASIFICACION

Los cinco algoritmos fueron analizados bajo las mismas métricas. El resumen de estos resultados se encuentra en la Tabla 2. Nuestro criterio de decisión fue F-Score dado que provee la armónica media entre la precisión y la recuperación (recall) como se muestra en la siguiente formula:

$$F_1 score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Tabla 2: Resultados de las métricas evaluadas en los algoritmos

	Soporte de máquinas Vectoriales SMV	**Redes Neuronales RN**	**Árboles de Decisiones AD**	**Bosques Aleatorios BA**	**Regresión Logística RL**
F-Score	0,8951569	0,8866559	0,889378	0,8910678	0,8894826
Precisión	0,8599858	0,8598123	0,866170	0,8644938	0,8704747
Recuperación (Recall)	0,9333276	0,9152296	0,913863	0,9193273	0,9093392
Exactitud	0,8454062	0,8345406	0,839249	0,8410600	0,8402148
Error	0,1545937	0,1654593	0,160750	0,1589399	0,1597851
AUC	0,8972	0,8986	0,883	0,8994	0,9028
Cohen Kappa	0,601	0,596	0,596	0,598	0,602

Al comparar SMV con RN, AD, BA, y RL, encontramos que SMV proporciona resultados ligeramente mejores con respecto al puntaje F (los valores varían entre 0,00577 y 0,00408). Cada clasificador se probó utilizando el mismo conjunto de datos y características de datos utilizando la plataforma analítica abierta knime. SMV se destaca como el mejor en la mayoría de las métricas. Muestra su precisión superior alcanzando el mejor F-Score (89,51%).

De manera factible, la razón para que SMV obtiene el mejor rendimiento es porque clasifica de manera eficiente los datos separables no lineales cuando se usa la función de kernel apropiada, es altamente tolerante a los sobreajustes y es muy preciso (es decir, 89,72%). De manera similar ocurre con BA, que se sabe que es un mejor clasificador cuando se presentan múltiples variables categóricas [35].

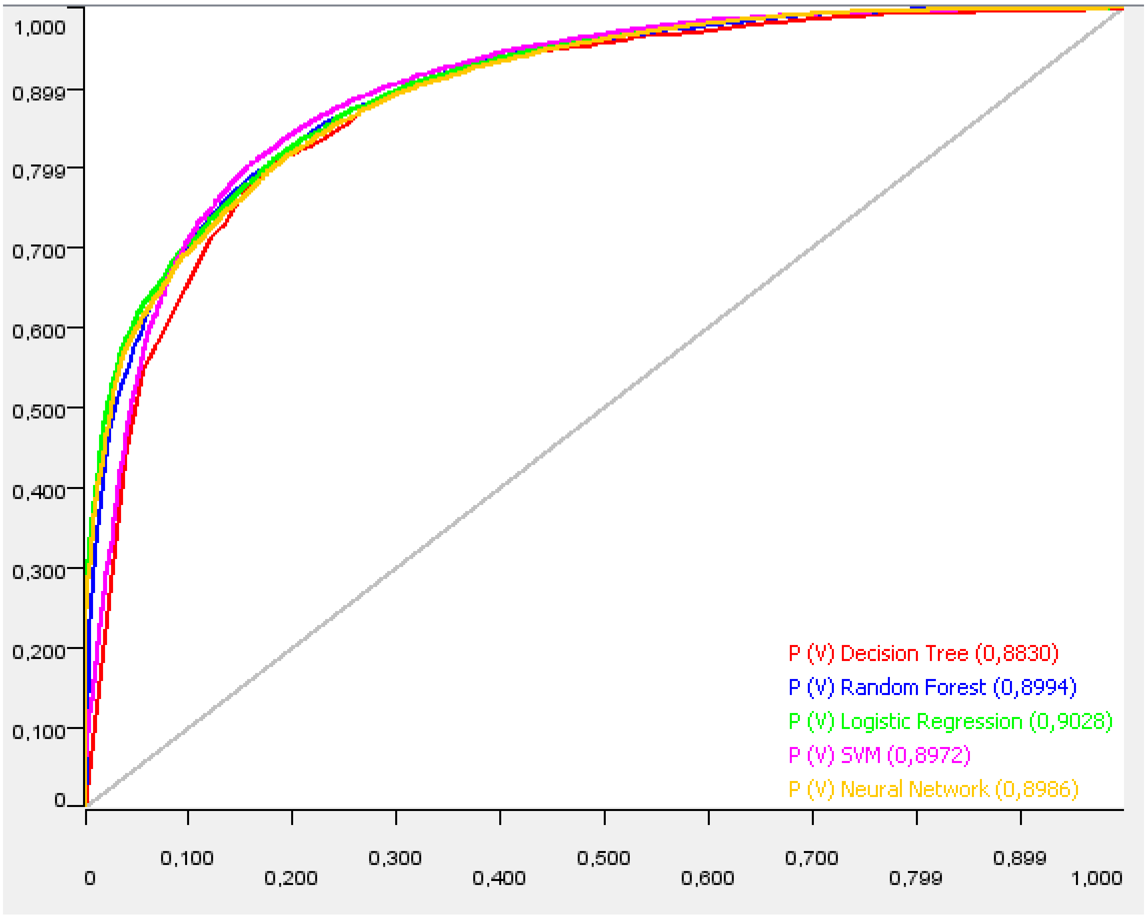


Ilustración 6: Áreas Bajo la Curva (AUC) de los algoritmos analizados.

En términos de Área bajo la curva (AUC), que representa el rendimiento esperado, el clasificador RL, obtiene el mejor resultado (es decir, 0,9028) con 0,0056 puntos sobre SMV y el valor kappa (es decir, 0,602). Por el contrario, la Red neuronal artificial tenía las peores métricas globales con el valor kappa más bajo (es decir, 0,596) y la exactitud (es decir, 83,45%). Dado que nuestros dos resultados binarios (graduados y no graduados) son igualmente importantes para nuestro objetivo modelo, el F-Score es la métrica adecuada para buscar un equilibrio entre Precisión y Recuperación (recall). Por lo tanto, y para los resultados sobresalientes en las otras métricas evaluadas, el Soporte de máquinas vectoriales es el algoritmo de aprendizaje automático utilizado en nuestro modelo propuesto.

Sin embargo, es importante observar que antes de la comparación de los algoritmos, y dada la reducida cantidad de atributos proporcionados por la universidad que sirvió como caso de estudio, donde se abordó un modelo educativo presencial, probamos si los datos proporcionados eran suficientes para representar la situación académica de los estudiantes ejecutando los algoritmos mencionados anteriormente con una cantidad diferente de características. Los resultados son expuestos en la Ilustración 7.

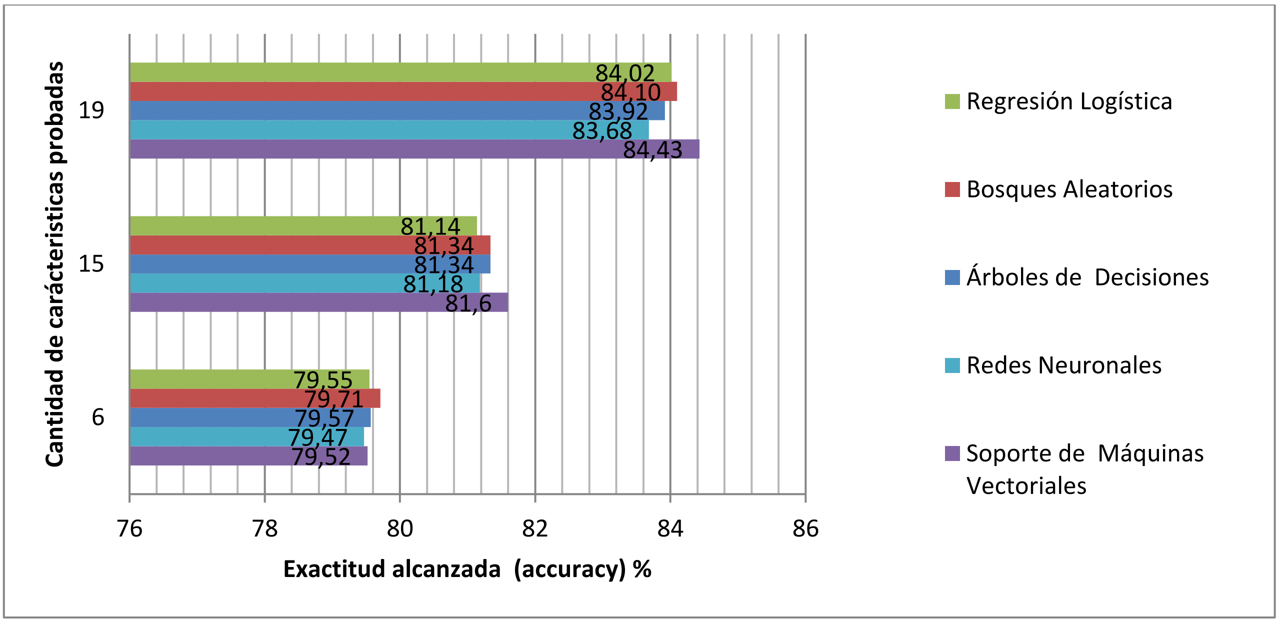
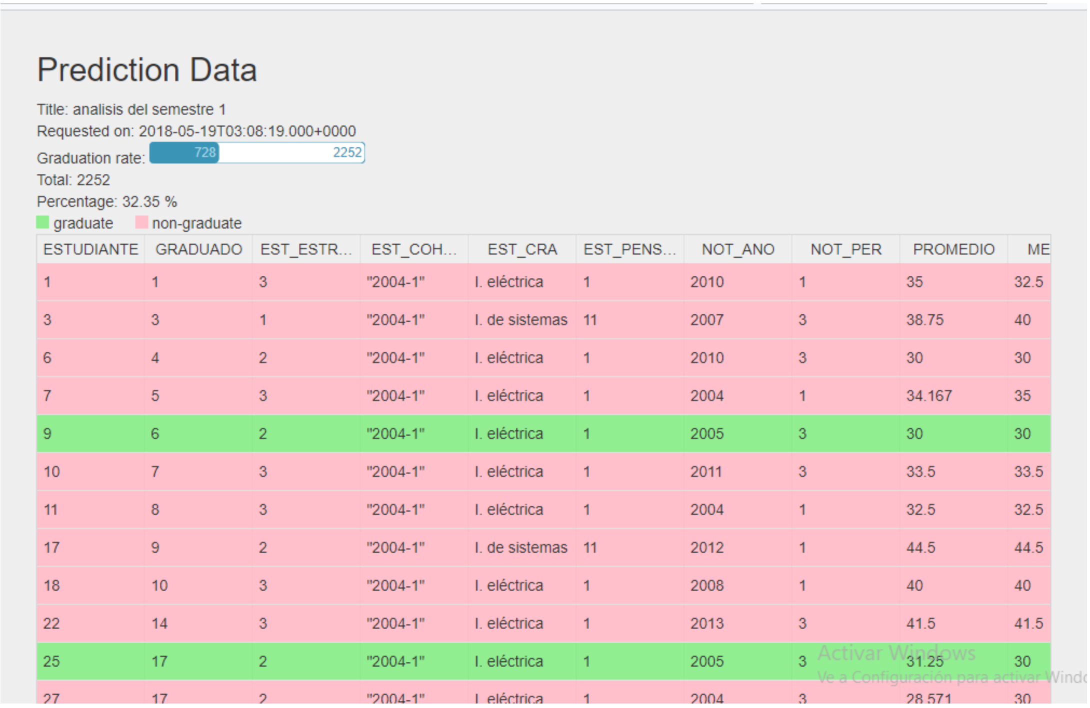


Ilustración 7: Exactitud alcanzada por los algoritmos con diferente cantidad de características de datos

El eje *x* en la Ilustración 7 indica el porcentaje de exactitud (accuracy) alcanzada por los algoritmos probados, mientras que el eje *y* indica los tres escenarios ejecutados donde varía la cantidad de características de los datos ingresados. A partir de los resultados de exactitud de este experimento, observamos que era relevante agregar más características al conjunto de datos al calcularlas. La exactitud aumenta a medida que se integran más características en el modelo. La mejora de la exactitud es de aproximadamente el 5% cuando se usan 19 características a cuando solo se usan 6 características. Cuando esto se refiere a los estudiantes clasificados correctamente significa que más de 300 estudiantes pasaron a estar clasificados correctamente, lo que se convierten en un hallazgo significativo.

4.2 VISUALIZACION DE LOS RESULTADOS

Con el objetivo de proporcionar una plataforma Web amigable para acceder a las predicciones realizadas por el modelo, apoyar las necesidades específicas de la Institución de Educación Superior que sirvió como estudio de caso y, superar los altos costos del servidor de knime, desarrollamos un prototipo de visualización de resultados. Algunos de los resultados se muestran en las Ilustraciones 8 y 9.



Prediction Data

Title: analisis del semestre 1

Requested on: 2018-05-19T03:08:19.000+0000

Graduation rate: 728 2252

Total: 2252

Percentage: 32.35 %

🟩 graduate 🟥 non-graduate

ESTUDIANTE	GRADUADO	EST_ESTR...	EST_COH...	EST_CRA	EST_PENS...	NOT_ANO	NOT_PER	PROMEDIO	ME
1	1	3	"2004-1"	I. eléctrica	1	2010	1	35	32.5
3	3	1	"2004-1"	I. de sistemas	11	2007	3	38.75	40
6	4	2	"2004-1"	I. eléctrica	1	2010	3	30	30
7	5	3	"2004-1"	I. eléctrica	1	2004	1	34.167	35
9	6	2	"2004-1"	I. eléctrica	1	2005	3	30	30
10	7	3	"2004-1"	I. eléctrica	1	2011	3	33.5	33.5
11	8	3	"2004-1"	I. eléctrica	1	2004	1	32.5	32.5
17	9	2	"2004-1"	I. de sistemas	11	2012	1	44.5	44.5
18	10	3	"2004-1"	I. eléctrica	1	2008	1	40	40
22	14	3	"2004-1"	I. eléctrica	1	2013	3	41.5	41.5
25	17	2	"2004-1"	I. eléctrica	1	2005	3	31.25	30
27	17	2	"2004-1"	I. eléctrica	1	2004	3	28.571	30

Ilustración 8: Consulta académica dentro del prototipo

El color verde representa a los estudiantes graduados, mientras que el color rojo representa a los estudiantes no graduados. La tasa de graduación se calcula en un 47,9% con más de 2940 estudiantes graduados. Consideramos esto como un resultado de gran potencial, descrito por el decano de la Universidad Distrital como: "adecuado para las decisiones de retención" porque es "una forma sencilla de entender la información que se utilizó", “no revelado antes”.

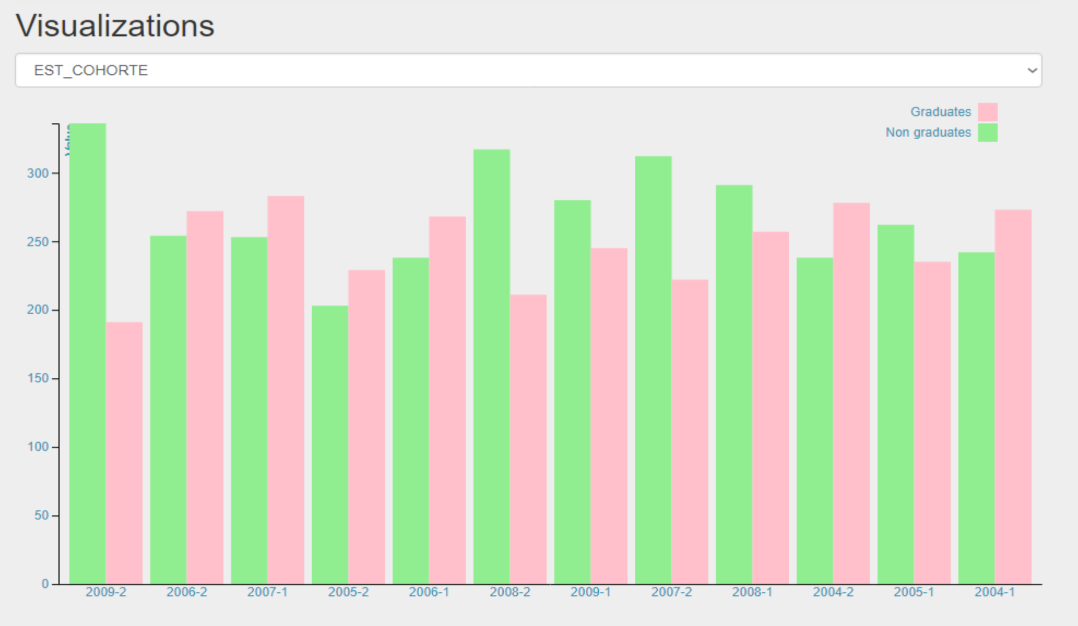


Ilustración 9: Estudiantes graduados vs estudiantes no graduados por cohorte

Las barras en la Ilustración 9 representan los estudiantes graduados (barras verdes) y los estudiantes no graduados (barras rojas) en cada cohorte. El eje *x* indica la cohorte, mientras que el eje *y* indica el número de estudiantes. Si bien el color verde prevalece en la mayoría de las cohortes observadas, la ligera diferencia con los estudiantes no graduados debería impulsar a los directivos de las IES para formular hipótesis y tomar medidas sobre estos preocupantes datos. Desde este punto de vista, el prototipo del kit de herramientas muestra los resultados de nuestro modelo propuesto y apoya a los gerentes y directores a través de información detallada y altamente confiable. Así, surgen nuevas oportunidades para lograr decisiones más favorables.

5. CONCLUSIONES

Existe una preocupación latente por parte de los directivos de las Instituciones de Educación Superior a la hora de tomar decisiones académicas, adicionalmente por la falta de toda la información necesaria proveniente de todas las partes interesadas. Este complejo proceso se ha confiado a decisiones del pasado e incluso a intuiciones.

En el campo educativo, el aprendizaje de máquina aún está emergiendo, mostrando gran efectividad para análisis y predicción de datos. Haciendo uso de diferentes algoritmos, los directivos de las IES obtienen una mayor comprensión de las diferentes variables involucradas al tomar una decisión, no solo en el área académica, sino en la planificación de recursos, la gestión de profesores, el diseño curricular y otros factores relacionados.

La elección del algoritmo de aprendizaje de máquina depende del problema a resolver. Por lo tanto, la metodología propuesta es un resultado esencial de la investigación, ya que especifica sugerencias importantes al intentar el pronóstico académico, desde la gobernanza y la comprensión operativa, hasta como desplegar la información obtenida en una herramienta Web. Además, la investigación proporciona la comparación de cinco algoritmos de aprendizaje automático robustos y su capacidad para reconocer una segmentación equilibrada de estudiantes graduados y no graduados.

Se reveló que Soporte de máquinas vectoriales tiene el mejor reconocimiento F-Score (es decir, 89,51%) y exactitud (es decir, 84,54%). Es por esto por lo que hemos utilizado Soporte de máquinas vectoriales como el algoritmo en el que se basa nuestro modelo de decisión para realizar predicciones académicas. Sin embargo, a pesar de su alta precisión y los resultados destacados de las métricas, este clasificador puede ser ineficiente y lento debido a sus dificultades de cálculo, así como a la complejidad del modelo con un conjunto de entrenamiento de escala más masiva [30]. La complejidad temporal en SMV oscila entre $O(n^2)$ y $O(n^3)$ donde n es el tamaño del conjunto de entrenamiento [45], pero, hasta donde conocemos, ninguno de los estudios similares ha utilizado un conjunto de datos más extenso que el nuestro. Sin embargo, creemos que, si se analiza un conjunto de datos más grande, el procesamiento del tiempo puede convertirse en una desventaja del modelo sin afectar su alta confiabilidad.

En el presente estudio, hemos superado la falta de atributos de datos en el caso de estudio. Después de varios experimentos, encontramos que los algoritmos alcanzaron una mejora de precisión significativa cuando se agregaron más atributos al calcularlos a partir de los datos sin procesar.

Durante los experimentos, utilizamos knime, una plataforma analítica que se adapta a nuestras necesidades, pero también existen algunas plataformas analíticas diferentes. Alternativamente a knime, plataformas como TensorFlow [46], RapidMiner [47], Alteryx [48], entre otras, permiten, a través de nodos o bibliotecas, acceder a otras herramientas equivalentes a las utilizadas en este trabajo. Dado que tenemos que enfrentar un alto costo para el servidor knime, y también cumplir con las necesidades específicas de la universidad del caso de estudio, desarrollamos nuestra aplicación Web que permite ejecutar predicciones y visualizar resultados.

La realización del modelo propuesto ofrece resultados de predicciones confiables e información reveladora sobre la situación académica de las IES. Además, los resultados logrados a lo largo de la ejecución del modelo y la retroalimentación positiva recibida por personal administrador de la universidad del caso de estudio, nos sugieren que es posible apoyar la toma de decisiones académicas a través de nuestro modelo de manera eficiente.

Esta investigación representa un modelo innovador que apoya a los altos administradores jerárquicos de IES durante la toma de decisiones académicas con respecto a los índices de graduación. Nos preocupamos por las partes interesadas que no se abordaron y que tienen un impacto significativo en toda la comunidad institucional cuando deciden. Además, incentivamos la formulación de hipótesis y la decisión con respecto a las tasas de retención, las políticas de exclusión de los estudiantes, las tasas de deserción escolar y el fortalecimiento de los programas a partir del modelo de toma de decisiones propuesto.

6. TRABAJO FUTURO

En un futuro próximo, se diseñará una propuesta de modelo híbrido entre el Soporte de máquinas vectoriales y el Bosque aleatorio para la predicción de tasa de graduados. Esto, debido a que los algoritmos anteriormente mencionados obtuvieron un desempeño superior en el problema de clasificación de estudiantes.

Una de las principales limitaciones de este estudio fue la adquisición de datos. Por lo tanto, ejecutaremos el mismo modelo propuesto con atributos de datos más grandes que incluyen datos socioeconómicos y demográficos de los estudiantes. Por ejemplo, el género de los estudiantes, el costo de la matrícula, el promedio de notas de la escuela secundaria, los ingresos mensuales, las personas a cargo y la ciudad de donde vinieron, por mencionar algunos atributos que valdría la pena analizar. Cuando se tengan más parámetros de entrada, el uso de un parámetro de optimización como el degradado descendente [49] podría ayudar a optimizar nuestro problema.

La herramienta Web presentada fue una versión preliminar de prueba de visualización de datos. Sin embargo, se requiere un arduo trabajo a futuro en el mejoramiento de esta, no solo para la visualización de información más amigable que cumpla los estándares de interacción persona-computadora, sino para responder a las necesidades reales de carga y tiempos de respuesta, entre otros.

7. INFORME DE FACTOR DE IMPACTO DE LAS PUBLICACIONES

A continuación, se presenta el informe de factor de impacto de los tres artículos utilizados para presentar el trabajo de tesis doctoral en la modalidad de compendio de publicaciones. La copia completa de los artículos figura en el anexo del documento.

Título: Supporting academic decision making at higher educational institutions using Machine Learning–based algorithms
Autores: Yuri Nieto, Vicente García-Díaz, Carlos Montenegro, Rubén González Crespo.
Revista: Journal of Soft Computing
Factor de impacto JCR: Computer Science, Artificial Intelligence, Computational intelligence and complexity, Control Engineering, Mathematics.
JCR 2.784 40/106 (segundo cuartil).
Volumen y fecha de publicación: Volumen 23, Número 12, año 2019.
Páginas: 4145-4153
DOI: 10,1007/s00500-018-3064-6
Aportación: Al ser la primera aportación del trabajo de investigación, este articulo muestra cómo se llevó a cabo el estudio para categorizar las decisiones académicas en las que los directivos de las universidades necesitan apoyo durante la toma de decisiones. De esta manera se identificó y definió que la decisión académica que se apoyaría con este trabajo sería la tasa de graduación estudiantil. Se establece la metodología de trabajo y se toma una universidad como caso de estudio para llevar a cabo la experimentación con datos reales. Se presenta la arquitectura y ejecución de dos algoritmos de máquinas de aprendizaje, así como el análisis de su comparación usando la matriz de confusión y el área bajo la curva. Se concluye que las Redes neuronales y los Soportes de máquinas vectoriales son algoritmos eficientes para realizar las predicciones de estudiantes a graduarse y que este último tiene un rendimiento superior.

Título: Usage of machine learning for strategic decision making at higher educational institutions
Autores: Yuri Nieto, Vicente García-Díaz, Carlos Montenegro, Claudio Camilo González, Rubén González-Crespo.
Revista: IEEE ACCESS
Factor de impacto JCR: Computer Science, Information Systems, Engineering, Electrical and Electronic science, Telecommunications science
JCR 4.098 23/155 (primer cuartil)
Volumen y fecha de publicación: Volumen 7, año 2019
Páginas: 75007-75017
DOI: 10,11109/access.2019.2919343
Aportación: Este artículo se enfoca en categorizar las decisiones académicas de las Instituciones de Educación Superior (IES) de acuerdo con el tipo de gobierno de las instituciones y su estructura en un nivel

jerárquico. Se revisan los trabajos similares respecto al uso de las máquinas de aprendizaje en la educación y se corrobora que las máquinas de aprendizaje no han sido utilizadas para apoyar específicamente a los directivos en la toma de decisiones. Se presenta la arquitectura y experimentación de tres algoritmos de máquinas de aprendizaje adicionales al artículo anterior, como lo son: Árboles de decisiones, Bosques aleatorios y Regresión logística. Se comparan los resultados de dichos algoritmos y se observa que los mejores resultados los obtienen los Bosques aleatorios con una eficiencia del 84.11%. Finalmente se concluye la viabilidad de los resultados proporcionados por estos algoritmos para la identificación temprana de estudiantes tendientes a no graduarse, lo cual permite al gobierno estratégico de las IES planear actividades correspondientes a tasas de deserción, políticas estudiantiles, tasas de retención, reforzamiento de programas entre otros.

Título: Decision-making model at Higher Educational Institutions based on Machine Learning
Autores: Yuri Nieto, Vicente García-Díaz, Carlos Montenegro
Revista: Journal of Universal Computer Science
Factor de impacto JCR: Computer Science, Theory and Methods, Software engineering
JCR 0,910 72/104 (tercer cuartil)
Volumen y fecha de publicación: Volumen 25, año 2019
DOI: Aceptado, pendiente por ser asignado.
Aportación: Nuestra última publicación resume la metodología del trabajo realizado durante estos tres años en un modelo gráfico basado en los estándares de la minería de datos y los pronósticos, impulsado por datos, en el cual se explica paso a paso los procesos realizados desde el entendimiento del gobierno y operaciones de las universidades, hasta la evaluación del algoritmo y su despliegue. Se detalla la construcción del modelo de máquina de aprendizaje que predice la tasa de graduados. Dicho modelo, se obtuvo gracias a la evaluación de los cinco algoritmos probados en los artículos anteriores. Se presentan los parámetros y pasos a través de un modelo de toma de decisiones que facilita un marco de referencia para los directores y altos administrativos de las IES. Por último, se presenta la arquitectura y la funcionalidad de una herramienta Web que muestra los resultados de manera amigable para apoyar a los directivos y altos administrativos de las IES en la toma de decisiones.

8. REFERENCIAS

[1] T. N. C. Vo and H. P. Nguyen, "A Knowledge-Driven Educational Decision Support System," in *2012 IEEE RIVF International Conference on Computing & Communication Technologies, Research, Innovation, and Vision for the Future*, 2012, pp. 1–6.

[2] M. Guilbault, "Students as customers in higher education : The (controversial) debate needs to end," *J. Retail. Consum. Serv.*, pp. 8–11, 2017.

[3] K. Stefanova and D. Kabakchieva, "Educational data mining perspectives within university big data environment," in *2017 International Conference on Engineering, Technology and Innovation: Engineering, Technology and Innovation Management Beyond 2020: New Challenges, New Approaches, ICE/ITMC 2017*, 2018, pp. 264–270.

[4] Y. Nieto and C. Montenegro, "System Architecture Based on Learning Analytcis to Educational Decision Makers Toolkit," *Adv. Comput. Sci. Enginnering*, vol. 13, no. 2, pp. 89–105, 2014.

[5] Hakim Lounis and Tamer Fares, "Using efficient Machine-Learning Models to Assess Two Important Quality Factors: Maintainability and Reusability," pp. 170–177, 2011.

[6] A. Diez-Olivan, J. Del Ser, D. Galar, and B. Sierra, "Data fusion and machine learning for industrial prognosis: Trends and perspectives towards Industry 4.0," *Inf. Fusion*, vol. 50, pp. 92–111, 2019.

[7] A. Hoffait and M. Schyns, "Early detection of university students with potential difficulties," *Decis. Support Syst.*, 2017.

[8] E. B. Costa, B. Fonseca, M. A. Santana, F. F. de Araújo, and J. Rego, "Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses," *Comput. Human Behav.*, vol. 73, pp. 247–256, 2017.

[9] Z. Yongshan, W. Jia, Z. Chuan, and C. a Zhihua, "Instance cloned extreme learning machine," *Pattern Recognit.*, vol. 68, pp. 52–65, 2017.

[10] C. Stoean and R. Stoean, "Post-evolution of variable-length class prototypes to unlock decision majing within support vector machines," *Appl. Soft Comput.*, no. 25, pp. 159–173, 2014.

[11] Y. Kaneda, Y. Pei, Q. Zhao, and Y. Liu, "Study on the effect of learning parameters on decision boundary making algorithm," in *EEE International Conference on Systems, Man and Cybernetics*, 2014, pp. 705–710.

[12] X. Wang, F. Huang, and Y. Cheng, "Computational performance optimization of support vector machine based on support vectors," *Neurocomputing*, vol. 211, pp. 66–71, 2016.

[13] N. Ghatasheh, "Knowledge Level Assessment in e-Learning Systems Using Machine Learning and User Activity Analysis," vol. 6, no. 4, 2015.

[14] G. Czibula, I. Gergely, and R. Gaceanu, "A support vector machine model for intelligent selection fo data representations," *Appl. Soft Comput.*, vol. 18, pp. 70–81, 2014.

[15] S. Chan and C. Yang, "Governance styles in Taiwanese universities : Features and e ff ects," *Int. J. Educ. Dev.*, pp. 1–7, 2017.

[16] F. A. Goni, A. G. Chofreh, M. Mukhtar, S. Sahran, S. A. Shukor, and J. J. Klemeš, "Strategic alignment between sustainability and information systems: A case analysis in Malaysian Public Higher Education Institutions," *J. Clean. Prod.*, vol. 168, pp. 263–270, 2017.

[17] E. Indrayani, "Management of Academic Information System (AIS) at Higher Education in the City of Bandung," *Procedia - Soc. Behav. Sci.*, vol. 103, pp. 628–636, 2013.

[18] K. V. Pincus, D. E. Stout, J. E. Sorensen, K. D. Stocks, and R. A. Lawson, "Forces for change in higher education and implications for the accounting academy," *J. Account. Educ.*, vol. 40, pp. 1–18, 2016.

[19] U. M. Azeiteiro, P. Bacelar-Nicolau, F. J. P. Caetano, and S. Caeiro, "Education for sustainable development through e-learning in higher education: Experiences from Portugal," *J. Clean. Prod.*, vol. 106, pp. 308–319, 2015.

[20] Y. Nieto, V. García-Díaz, C. Montenegro, and R. G. Crespo, "Supporting academic decision making at higher educational institutions using machine learning-based algorithms," *Soft Comput.*, pp. 4145–4153, 2018.

[21] "Ministerio de Educación Nacional de Colombia." [Online]. Available: https://www.mineducacion.gov.co/portal/. [Accessed: 18-Sep-2016].

[22] IBM, "SPSS Statistics." [Online]. Available: https://www.ibm.com/co-es/products/spss-statistics/pricing. [Accessed: 20-Aug-2019].

[23] "UDIN Facultad de Ingeniería Universidad Distrital Francisco José de Caldas." [Online]. Available: https://ingenieria.udistrital.edu.co/. [Accessed: 18-Sep-2016].

[24] A. J. Stimpson, M. L. Cummings, and S. Member, "Assessing Intervention Timing in Computer - Based Education Using Machine Learning Algorithms," *IEEE Access*, vol. 2, pp. 78–87, 2014.

[25] M. Tan and P. Shao, "Prediction of student dropout in E-learning program through the use of machine learning method," *Int. J. Emerg. Technol. Learn.*, vol. 10, no. 1, pp. 11–17, 2015.

[26] H. Ramchoun, M. Amine, J. Idrissi, Y. Ghanou, and M. Ettaouil, "Multilayer Perceptron: Architecture Optimization and Training," *Int. J. Interact. Multimed. Artif. Intell.*, vol. 4, no. 1, p. 26, 2016.

[27] M. T. Hagan, H. B. Demuth, and M. H. Beale, *Neural Network Design*, 2nd ed. 2014.

[28] M. Berthold, N. Cebron, and D. Fabian, "KNIME: The Konstanz Information Miner," in *Studies in Classification, Data Analysis, and Knowledge Organization*, Spinger, 2007.

[29] S. Karsoliya, "Approximating Number of Hidden layer neurons in Multiple Hidden Layer BPNN Architecture," *Int. J. Eng. Trends Technol.*, vol. 3, no. 6, pp. 714–717, 2012.

[30] C. Liu, W. Wang, M. Wang, F. Lv, and M. Konan, "An efficient instance selection algorithm to reconstruct training set for support vector machine," *Knowledge-Based Syst.*, vol. 116, pp. 58–73, 2017.

[31] S. S. Athani, S. A. Kodli, M. N. Banavasi, and P. G. S. Hiremath, "Student performance predictor using multiclass support vector classification algorithm," in *IEEE International Conference on Signal Processing and Communication, ICSPC 2017*, 2018, no. July, pp. 341–346.

[32] S. Helal *et al.*, "Predicting academic performance by considering student heterogeneity," *Knowledge-Based Syst.*, vol. 161, pp. 134–146, 2018.

[33] D. Delen, H. Zaim, and C. Kusey, "A comparative analysis of machine learning systems for measuring the impact of knowledge management practices," *Decis. Support Syst.*, vol. 54, pp. 1150–1160, 2013.

[34] S. Huaining and H. Xuegang, "Attribute selection for decision tree learning with class constraint," *Chemom. Intell. Lab. Syst.*, vol. 162, pp. 123–129, 2017.

[35] V. Rodriguez, M. Sanchez, and M. Chica, "Machine learning predictive models for mineral prospectivity : An evaluation of neural networks , random forest , regression trees and support vector machines," *Ore Geol. Rev.*, vol. 71, pp. 804–818, 2015.

[36] Y. Nieto, V. Garcia-Diaz, C. Montenegro, C. Gonzalez, and R. Gonzalez, "Usage of Machine Learning for Strategic Decision Making at Higher Educational Institutions," *IEEE Access*, vol. 7, pp. 75007–75017, 2009.

[37] S. Singh, R. Taylor, M. Rahman, and B. Pradha, "Developing robust arsenic awareness prediciton models using machine learning algorithms," *J. Enviromental Manag.*, vol. 211, pp. 125–137, 2018.

[38] Y. Nieto, V. García, C. Montenegro, R. Gonzalez, and C. Gonzalez, "Decision-making model at Higher Educational Institutions based on Machine Learning," *J. Univers. Comput. Sci.*, vol. Aceptado p, 2019.

[39] "React – Una biblioteca de JavaScript para construir interfaces de usuario." [Online]. Available: https://es.reactjs.org/. [Accessed: 22-Aug-2019].

[40] "D3.js - Data-Driven Documents." [Online]. Available: https://d3js.org/. [Accessed: 22-Aug-2019].

[41] "¿Qué es Java Hibernate? – Educacion IT." [Online]. Available: https://blog.educacionit.com/2013/02/07/que-es-java-hibernate/. [Accessed: 23-Aug-2019].

[42] "JPA (Java Persistence API)." [Online]. Available: https://www.ibm.com/support/knowledgecenter/es/SSD28V_liberty/com.ibm.websphere.wlp.core.doc/ae/cwlp_jpa.html. [Accessed: 23-Aug-2019].

[43] "Liquibase, Database Refactoring." [Online]. Available: https://www.liquibase.org/. [Accessed: 23-Aug-2019].

[44] "MySQL." [Online]. Available: https://www.mysql.com/. [Accessed: 23-Aug-2019].

[45] A. Stecto, F. Dinmohammadi, X. Zhao, and V. Robu, "Machine learning methods for wind turbine condition monitoring: A review," *Renew. Energy*, vol. 133, pp. 620–635, 2019.

[46] "TensorFlow." [Online]. Available: https://www.tensorflow.org/. [Accessed: 23-Aug-2019].

[47] "RapidMiner©." [Online]. Available: https://rapidminer.com/. [Accessed: 23-Aug-2019].

[48] "Alteryx: Self-Service Data Analytics Platform." [Online]. Available: https://www.alteryx.com/. [Accessed: 23-Aug-2019].

[49] Y. Jeon, Y. Park, and S. Lee, "Machine Learning Optimization of Parameters for Noise Estimation," *J. Univers. Comput. Sci.*, vol. 24, no. 9, pp. 1271–1281, 2018.

9. BIBLIOGRAFÍA DE LA INVESTIGACIÓN

- A. Clark, “IT Governance: Determining who decides,” Educase Center for Applied Reserach, vol. 2005, no. 24. Educase, Boulder, pp. 1–13, 2007.
- A. Diez-Olivan, J. Del Ser, D. Galar, and B. Sierra, “Data fusion and machine learning for industrial prognosis: Trends and perspectives towards Industry 4.0,” Information Fusion, Vol. 50 (2019), pp. 92–111. https://doi.org/https://doi.org/10.1016/j.inffus.2018.10.005,
- A. Hoffait and M. Schyns, “Early detection of university students with potential difficulties,” Decis. Support Syst., 2017.
- A. J. Stimpson, M. L. Cummings, and S. Member, “Assessing Intervention Timing in Computer - Based Education Using Machine Learning Algorithms,” IEEE Access, vol. 2, pp. 78–87, 2014.
- A. La, Ľ. Bajzíková, and I. Dedze, “Barriers and drivers of innovation in higher education: Case study-based evidence across ten European universities,” Int. J. Educ. Dev., vol. 55, no. May, pp. 69–79, 2017.
- A. L. Dyckhoff, D. Zielke, M. Bültmann, and M. A. Chatti, “Design and Implementation of a Learning Analytics Toolkit for Teachers,” Educ. Technol. Soc., vol. 15, pp. 58–76, 2012.
- A. L. Dyckhoff, V. Lukarov, A. Muslim, M. A. Chatti, and U. Schroeder, “Supporting Action Research with Learning Analytics,” in Learning Analytics and Knowledge 2013, 2013, pp. 220–229.
- A. Muklason, A. Parkes, and E. Ozcan, “Fairness in examination timetabling: Student preferences and extended formulations,” Appl. Soft Comput., vol. 55, pp. 302–318, 2017.
- A. Oztekin, D. Delen, and A. Turkylmaz, “A Machine learning-based usability evaluation method for eLearning systems,” Decis. Support Syst., vol. 56, pp. 66–73, 2013.
- A. R. T. Góes, M. T. Arns Steiner, and P. J. Steiner Neto, “Education Quality Measured by the Classification of School Performance Using Quality Labels,” Appl. Mech. Mater., vol. 670, pp. 1675–1683, 2014.
- A.-P. Pavel, A. Fruth, and M.-N. Neacsu, “ICT and E-Learning – Catalysts for Innovation and Quality in Higher Education,” Procedia Econ. Financ., vol. 23, no. October 2014, pp. 704–711, 2015.
- A. Sandoval, C. Gonzalez, R. Alarcon, K. Pichara, and M. Montenegro, “Centralized student performance prediction in large courses based on low-cost variables in an institutional context,” Internet High. Educ., vol. 37, pp. 76–89, 2018.
- A. Shanthini, G. Vinodhini, and R. Chandrasekaran, “Predicting students’ academic performance in the university using meta decision tree classifiers,” J. Comput. Sci., vol. 14, no. 5, pp. 654–662, 2018.
- A. Stecto, F. Dinmohammadi, X. Zhao, and V. Robu, “Machine learning methods for wind turbine condition monitoring: A review,” Renew. Energy, vol. 133, pp. 620–635, 2019.
- A. Trabesli, Z. Elouedi, and E. Lefevre, “Decision tree classifiers for evidential attribute values and class labels,” Fuzzy Sets Syst., vol. 366, pp. 46–62, 2019.
- A. Valentín, P. M. Mateos, M. M. González-Tablas, L. Pérez, E. López, and I. García, “Motivation and learning strategies in the use of ICTs among university students,” Comput. Educ., vol. 61, pp. 52–58, 2013.
- A. Y. Noaman and F. F. Ahmed, “ERP Systems Functionalities in Higher Education,” in International

Conferences on Communication, Management and Information Technology, 2015, vol. 65, pp. 385–395.

- C. Dewberry, Statistical methods for organizational research. New York: Routledge, 2004.
- C. Gonzalez, E. Elhariri, N. El-Bendary, and A. Fernandez, “Machine Learning Based Classification Approach for predicting Students Performance in Blended Learning,” in Advances in Intelligent Systems and Computing, 2016, vol. 407, pp. 47–56.
- Chih-Wei Hsu, Chih-Chung Chang and C.-J. L. “A Practical Guide to Support Vector Classification,” BJU Int., vol. 101, no. 1, pp. 1396–400, 2016.
- C. Hopp and G. A. Hoover, “How prevalent is academic misconduct in management research?” J. Bus. Res., vol. 80, no. July, pp. 73–81, 2017.
- C. Liu, W. Wang, M. Wang, F. Lv, and M. Konan, “An efficient instance selection algorithm to reconstruct training set for support vector machine,” Knowledge-Based Syst., vol. 116, pp. 58–73, 2017.
- C. M. Bishop, “Neural networks for pattern recognition,” J. Am. Stat. Assoc., vol. 92, p. 482, 1995.
- C. Stoean and R. Stoean, “Post-evolution of variable-length class prototypes to unlock decision making within support vector machines,” Appl. Soft Comput., no. 25, pp. 159–173, 2014.
- C. C. White, “A survey on the integration of decision analysis and expert systems for decision support,” IEEE Trans. Syst. Man. Cybern., vol. 20, no. 2, pp. 358–364, 1990,
- D. Delen, H. Zaim, and C. Kusey, “A comparative analysis of machine learning systems for measuring the impact of knowledge management practices,” Decis. Support Syst., vol. 54, pp. 1150–1160, 2013.
- D. Yu, J. Hu, Q. Li, Z. Tang, J. Yang, and H. Shen, “Constructing Query-Driven Dynamic Machine Learning Model With Application to Protein-Ligand Binding Sites Prediction,” vol. 14, no. 1, pp. 45–58, 2015.
- E. Aguiar, H. Lakkaraju, N. Bhanpuri, D. Miller, B. Yuhas, and K. L. Addison, “Who, when, and Why: A Machine Learning Approach to Prioritizing Students at Risk of Not Graduating High School on Time,” in Proceedings of the Fifth International Conference on Learning Analytics And Knowledge - LAK ’15, 2015, pp. 93–102.
- E. Alptekin and K. Ertugrul, “An integrated decision framework for evaluating and selecting e-learning products,” Appl. Soft Comput., vol. 11, p. 2990-2998, 2010,
- E. B. Costa, B. Fonseca, M. A. Santana, F. F. de Araújo, and J. Rego, “Evaluating the effectiveness of educational data mining techniques for early prediction of students’ academic failure in introductory programming courses,” Comput. Human Behav., vol. 73, pp. 247–256, 2017.
- E. Faham, A. Rezvanfar, S. H. Movahed Mohammadi, and M. Rajabi Nohooji, “Using system dynamics to develop education for sustainable development in higher education with the emphasis on the sustainability competencies of students,” Technol. Forecast. Soc. Change, vol. 123, pp. 307–326, 2016.
- F. A. Goni, A. G. Chofreh, M. Mukhtar, S. Sahran, S. A. Shukor, and J. J. Klemeš, “Strategic alignment between sustainability and information systems: A case analysis in Malaysian Public Higher Education Institutions,” J. Clean. Prod., vol. 168, pp. 263–270, 2017.
- F. Moreira, M. J. Ferreira, C. P. Santos, and N. Durao, “Evolution and use of mobile devices in higher education: A case study in Portuguese Higher Education Institutions between 2009/2010 and 2014/2015,” Telemat. Informatics, vol. 34, pp. 838–852, 2016.
- F. Saeed and A. Dixit, “A decision support system approach for accreditation & quality assurance

council at higher education institutions in Yemen," in IEEE 3rd International Conference on MOOCs, Innovation and Technology in Education (MITE), 2015, pp. 163–168.

- F. V. Elena, A. M. Manuel, and G. G. Carina S., "Which Use Give Teachers at La Laguna University to ICTs?," Procedia - Soc. Behav. Sci., vol. 93, pp. 1646–1651, 2013.
- G. Czibula, I. Gergely, and R. Gaceanu, "A support vector machine model for intelligent selection fo data representations," Appl. Soft Comput., vol. 18, pp. 70–81, 2014
- G. Hackeling, Mastering Machine Learning with scikit-learn. 2014
- G. R. Jones, Organizational Theory, Design, and Change, 7th ed. Prentice Hall, 2011.
- Hakim Lounis and Tamer Fares, "Using efficient Machine-Learning Models to Assess Two Important Quality Factors: Maintainability and Reusability," pp. 170–177, 2011.
- H. M. Vo, C. Zhu, and N. A. Diep, "The effect of blended learning on student performance at course-level in higher education: A meta-analysis," Stud. Educ. Eval., vol. 53, pp. 17–28, 2017
- H. Lakkaraju et al., "A Machine Learning Framework to Identify Students at Risk of Adverse Academic Outcomes," in International Conference on Knowledge Discovery and Data Mining KDD 2015, 2015, pp. 1909–1918.
- H. Ramchoun, M. Amine, J. Idrissi, Y. Ghanou, and M. Ettaouil, "Multilayer Perceptron: Architecture Optimization and Training," Int. J. Interact. Multimed. Artif. Intell., vol. 4, no. 1, p. 26, 2016.
- I. González-González and A. I. Jiménez-Zarco, "Using learning methodologies and resources in the development of critical thinking competency: An exploratory study in a virtual learning environment," Comput. Human Behav., vol. 51, pp. 1359–1366, 2015.
- I. H. Witten, E. Frank, and M. a Hall, Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed. San Francisco: Morgan Kauffman Publishers, 2011.
- I. M. Taucean and M. Tamasila, "Research Challenges for eLearning Support in Engineering and Management Training," Procedia - Soc. Behav. Sci., vol. 124, pp. 210–218, 2014.
- I. Livereis, T. Mikropoulos, and P. Pintelas, "A Decision Support System for Predicting Student Performance," Themes Sci. Technol. Educ., vol. 9, no. 1, pp. 43–57, 2016.
- I. S. Bianchi and R. D. Sousa, "IT Governance Mechanisms in Higher Education," Procedia Comput. Sci., vol. 100, pp. 941–946, 2016.
- I. Smeureanu and N. Isaila, "Information technology, support for innovation in education sciences," Procedia - Soc. Behav. Sci., vol. 15, pp. 751–755, 2011.
- J. Heaton, Introduction to Neural Networks with Java, 2nd ed., vol. 99. St. Louis: Heaton Research, Inc, 2008.
- J. Hu, H. Liu, Y. Chen, and J. Qin, "Strategic planning and the stratification of Chinese higher education institutions," Int. J. Educ. Dev., no. 2016, pp. 1–8, 2017.
- K. H. Wang et al., "Using community-based participatory research and organizational diagnosis to characterize relationships between community leaders and academic researchers," Prev. Med. Reports, vol. 7, pp. 180–186, 2017.
- K. Ihalagedara, R. Kithuldeniya, S. Weerasekara, and S. Deegalla, "Feasibility of Using Machine Learning to Access Control in Squid Proxy Server," pp. 491–494, 2015.
- K. J. Gerritsen-van Leeuwenkamp, D. Joosten-ten Brinke, and L. Kester, "Assessment quality in tertiary

education: An integrative literature review," Stud. Educ. Eval., vol. 55, no. July, pp. 94–116, 2017.

- K. Stefanova and D. Kabakchieva, "Educational data mining perspectives within university big data environment," in 2017 International Conference on Engineering, Technology and Innovation: Engineering, Technology and Innovation Management Beyond 2020: New Challenges, New Approaches, ICE/ITMC 2017, 2018, pp. 264–270,2018.
- K. V. Pincus, D. E. Stout, J. E. Sorensen, K. D. Stocks, and R. A. Lawson, "Forces for change in higher education and implications for the accounting academy," J. Account. Educ., vol. 40, pp. 1–18, 2016.
- L. Nanni, C. Fantozzi, and N. Lazzarani, "Coupling different methods for overcoming the class imbalance problem," Neurocomputing, vol. 158, pp. 48–61, 2015.
- M. Abdahllah, "A decision support model for long-term course planning," Decis. Support Syst., vol. 74, pp. 33–45, 2015.
- M. A. Chatti, A. L. Dyckhoff, U. Schroeder, and H. Thüs, "A reference model for learning analytics," Int. J. Technol. Enhanc. Learn., vol. 4, no. 5/6, pp. 1–22, 2012.
- M. Berthold, N. Cebron, and D. Fabian, "KNIME: The Konstanz Information Miner," in Studies in Classification, Data Analysis, and Knowledge Organization, Spinger, 2007.
- M. Darzi, S. Niaki, and M. Khedmati, "Binary classification of imbalanced datasets: The case of CoIL challenge 2000," Expert Syst. Appl., vol. 128, pp. 169–186, 2019.
- M. E. Zorrilla, D. García, and E. Álvarez, "A decision support system to improve e-learning environments," in Proc. of the EDBT/ICDT Workshops, 2010, pp. 1–8.
- Microfost Azure, "How to choose machine learning algorithms," docs.microsoft.com, 2017. [Online]. Available: https://docs.microsoft.com/es-es/azure/machine-learning/studio/algorithm-choice. [Accessed: 17-Jan-2018].
- M. Goga, S. Kuyoro, and N. Goga, "A Recommender for Improving the Student Academic Performance," Procedia - Soc. Behav. Sci., vol. 180, no. November 2014, pp. 1481–1488, 2015.
- M. Fischetti, "Fast training of Support Vector Machines with Gaussian Kernel," Discret. Optim., vol. 22, pp. 183–194, 2016.
- M. Guilbault, "Students as customers in higher education: The (controversial) debate needs to end," J. Retail. Consum. Serv., pp. 8–11, 2017.
- M. Tan and P. Shao, "Prediction of student dropout in E-learning program through the use of machine learning method," Int. J. Emerg. Technol. Learn., vol. 10, no. 1, pp. 11–17, 2015.
- N. Cohen and J. Gattuso, CCTV Operational Requirements Manual, no. 28. 2009.
- M. T. Hagan, H. B. Demuth, and M. H. Beale, Neural Network Design, 2nd ed. 2014.
- N. Ghatasheh, "Knowledge Level Assessment in e-Learning Systems Using Machine Learning and User Activity Analysis," vol. 6, no. 4, 2015.
- N. Werghi and F. K. Kamoun, "A decision-tree-based system for student academic advising and planning in information systems programmes," International Journal of Business Information Systems, vol. 5. p. 1, 2010,
- P. Wittek, Quantum Machine Learning, 1st ed. Elsevier, 2014.
- R. Rodriguez and G. Rubio, "Teaching quality in academic research," Int. Rev. Econ. Educ., vol. 129, pp. 10–27, 2016.

- R. S. J. D. Baker, “Data Mining for Education,” International Encyclopedia of Education. Elsevier, 2010,
- R. Skyrius, G. Kazakevičienė, and V. Bujauskas, “From Management Information Systems to Business Intelligence: The Development of Management Information Needs,” Int. J. Interact. Multimed. Artif. Intell., vol. 2, no. 3, p. 31, 2013.
- S. Boyer, B. U. Gelman, B. Schreck, and K. Veeramachaneni, “Data Science Foundry for MOOCs,” pp. in Proceedings of the 2015 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2015, , pp. 1–10. 2015.
- S. Chan and C. Yang, “Governance styles in Taiwanese universities: Features and effects,” Int. J. Educ. Dev., pp. 1–7, 2017.
- S. Dawson and E. Heathcote, “SNAPP: Realising the affordances of real-time SNA within networked learning environments,” in International Conference on Networked learning, 2010, pp. 125–133.
- S. Granja, “Colombia mejora en acceso a la educación superior, pero falta calidad,” El Tiempo, Bogotá, pp. 1–5, 05-Jun-2017.
- S. Helal et al., “Predicting academic performance by considering student heterogeneity,” Knowledge-Based Syst., vol. 161, pp. 134–146, 2018.
- S. Huaining and H. Xuegang, “Attribute selection for decision tree learning with class constraint,” Chemom. Intell. Lab. Syst., vol. 162, pp. 123–129, 2017.
- S. Karsoliya, “Approximating Number of Hidden Layer Neurons in Multiple Hidden Layer BPNN Architecture,” Int. J. Eng. Trends Technol., vol. 3, no. 6, pp. 714–717, 2012.
- S. K. Thangavel, P. D. Bkaratki, and A. Sankar, “Student placement analyzer: A recommendation system using machine learning,” in 2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS), 2017, pp. 1–5.
- S. M. Shuhidan, N. Mastuki, and W. M. N. W. M. Nori, “Accounting Information System and Decision Useful Information Fit Towards Cost Conscious Strategy in Malaysian Higher Education Institutions,” Procedia Econ. Financ., vol. 31, no. 15, pp. 885–895, 2015.
- S. S. Athani, S. A. Kodli, M. N. Banavasi, and P. G. S. Hiremath, “Student performance predictor using multiclass support vector classification algorithm,” in IEEE International Conference on Signal Processing and Communication, ICSPC 2017, 2018, no. July, pp. 341–346.
- S. Singh, R. Taylor, M. Rahman, and B. Pradha, “Developing robust arsenic awareness prediction models using machine learning algorithms,” J. Enviromental Manag., vol. 211, pp. 125–137, 2018.
- T. Anastasios, S. Cleo, P. Effie, T. Olivier, and M. George, “Institutional Research Management using an Integrated Information System,” Procedia - Soc. Behav. Sci., vol. 73, no. 0, pp. 518–525, 2013.
- T. N. C. Vo and H. P. Nguyen, “A Knowledge-Driven Educational Decision Support System,” in 2012 IEEE RIVF International Conference on Computing & Communication Technologies, Research, Innovation, and Vision for the Future, 2012, pp. 1–6.
- U. M. Azeiteiro, P. Bacelar-Nicolau, F. J. P. Caetano, and S. Caeiro, “Education for sustainable development through e-learning in higher education: Experiences from Portugal,” J. Clean. Prod., vol. 106, pp. 308–319, 2015.
- V. Miguéis, A. Freitas, P. J. V Garcia, and A. Silva, “Early segmentation of students according to their academic performance: A predictive modelling approach,” Decis. Support Syst., vol. 115, pp. 36–51,

2018.

- V. Rodriguez, M. Sanchez, and M. Chica, “Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines,” Ore Geol. Rev., vol. 71, pp. 804–818, 2015.
- X. Deng, Q. Liu, Y. Deng, and S. Mahadevan, “An improved method to construct basic probability assignment based on the confusion matrix for classification problem,” Inf. Sci. (Ny)., vol. 340–341, pp. 250–261, 2016.
- X. Huang, A. Maier, J. Hornegger, and J. Suykens, “Indefinite Kernels in least squares support vector machines and principal component analysis,” Appl. Comput. Harmon. Anal., vol. 43, pp. 162–172, 2017.
- Xiao-YanLiu, “Private colleges teachers evaluation system based on support vector machine (SVM),” in International Conference on Applied Science and Engineering Innovation ASEI 2015, 2015, no. Asei, pp. 1918–1921.
- X. Wang, F. Huang, and Y. Cheng, “Computational performance optimization of support vector machine based on support vectors,” Neurocomputing, vol. 211, pp. 66–71, 2016.
- Y. Barlas. and V. Dicker., “A Dynamic Simulation Game for Strategic University Management (UNIGAME),” 2000,
- Y. Kaneda, Y. Pei, Q. Zhao, and Y. Liu, “Study on the effect of learning parameters on decision boundary making algorithm,” in EEE International Conference on Systems, Man and Cybernetics, 2014, pp. 705–710,
- W. S. Murray and L. A. Le Blanc, “A Decision Support System for Academic Advising,” 1995.
- Z. Yongshan, W. Jia, Z. Chuan, and C. a Zhihua, “Instance cloned extreme learning machine,” Pattern Recognit., vol. 68, pp. 52–65, 2017.

10. ANEXO DE PUBLICACIONES

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 22, 2019, accepted May 22, 2019, date of publication May 27, 2019, date of current version June 20, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2919343

Usage of Machine Learning for Strategic Decision Making at Higher Educational Institutions

YURI NIETO[1], VICENTE GACÍA-DÍAZ[1], CARLOS MONTENEGRO[2], CLAUDIO CAMILO GONZÁLEZ[3], AND RUBÉN GONZÁLEZ CRESPO[4]

[1]Department of Computer Science, University of Oviedo, 33007 Oviedo, Spain
[2]Engineering Department, Francisco Jose de Caldas District University, Bogotá 1100100, Colombia
[3]Engineering Department, National Open and Distance University UNAD, Bogotá 1100100, Colombia
[4]School of Engineering, Universidad Internacional de La Rioja (UNIR), 26002 Madrid, Spain

Corresponding author: Rubén González Crespo (ruben.gonzalez@unir.net)

This work was supported by the Corporation Unified National Higher Education (CUN).

ABSTRACT Decisions made at the strategic level of Higher Educational Institutions (HEIs) affect policies, strategies, and actions that the institutions make as a whole. Decision's structures at HEIs are depicted in this paper and their effectiveness in supporting the institutions' governance. The disengagement of the stakeholders and the lack of using efficient computational algorithms lead to 1) the decision process takes longer; 2) the ''whole picture'' is not involved along with all data necessary; and 3) small academic impact is produced by the decision, among others. Machine learning is an emerging field of artificial intelligence that using various algorithms analyzes information and provides a richer understanding of the data contained in a specific context. Based on the author's previous works, we focus on supporting decision-making at a strategic level, being deans' concerns the preeminent mission to bolster. In this paper, three supervised classification algorithms are deployed to predict graduation rates from real data about undergraduate engineering students in South America. The analysis of receiver operating characteristic (ROC) curve and accuracy are executed as measures of effectiveness to compare and evaluate decision tree, logistic regression, and random forest, where this last one demonstrates the best outcomes.

INDEX TERMS Decision trees, random forest, logistic regressions, machine learning, strategic decisions, Higher Educational Institutions.

I. INTRODUCTION

Notoriously a ''disengagement'' occurs regarding Higher Educational Institutions and education policymakers, students, managers and their subordinates [1]. Many barriers including technological conditions, rigid governance structure or vulnerability to continuous changes in government rules, may impede the support needed by managers and university's directors when making a decision.

In regard to technological conditions, face-to-face model education still addresses significant obstacles. For example, their administrative, as well as their academic information, is stored in various silos making formats employed vary significantly [2]. Additionally, as we observe in our case study university, some of the transactional processes such as the record of students' attendance or registered graduate documents, to mention a few examples, are still done manually in paper notebooks.

Nevertheless, in distance and blended educational models, the arena offers more resources. The amount of systematized information is natural because computers are highly suitable and practical for this work [3].

Stored data is not enough when directors and managers are deciding. Educational data, whether it is systematically or manually stored, should be analyzed to provide a proper presentation of valuable information to support these complex processes.

Therefore, the usage of efficient computational algorithms is vital to enhance this process. Through the latest years, Machine Learning has shown its outstanding capacity for pattern recognition and predicting outcomes for diverse datasets despite the field. Most of the work done in Machine Learning has focused on supervised algorithms. Their main strength is that they produce models that we can incorporate in the decision-making process [4]. In order to choose the most suitable learning algorithm, a clear objective is required, and an analysis of previous data must be performing. Thus, the feasibility of using a supervised algorithm

The associate editor coordinating the review of this manuscript and approving it for publication was Hazrat Ali.

over a not-supervised algorithm can be determined. Afterwards, depending on the viability of each, a choice between classification or regression algorithms needs to be made. Even though many studies have used ML algorithms to identify students; our research differs from the existing ones. (I) the data comes from a face-to-face educational model. (II) more diverse and numerous features from data collection are included on the algorithms architecture leading to achieve a higher overall accuracy that we analyze. (III) neither the stakeholders nor the objective goals have been evaluated before. (IV) we investigate deans and directors concerns when making academic decisions [5] as our work driver.

This article seeks to classify the decision's structure at HEIs and the influence of the institutional governance among them. This section is developed with the aim of depicting the impact and responsibility of strategic decisions not just in the academic context but the complete environment where the Higher Educational Institution is located.

An extensive literature review looks for the classification of the uses of Information and Communication Technologies (ICT) at HEIs and ongoing applications that used ML in the education field. The primary goal focuses on the comparison of three supervised Machine Learning (ML) algorithms that, used as predictors, would enhance decision at the strategic level. Specifically, we applied Decision Trees, Random Forests and Logistic Regression to predict graduation rates using real data from a face-to-face model education university in South America. Analysis of the ROC curve and Accuracy are executed as measures of effectiveness to compare and evaluate the three algorithms.

The rest of the paper is laid out as follows: In Section 2, a review of the current literature is discussed. In Section 3, we introduce the classification of decisions at Higher Educational Institutions. In Section 4, details of the method to compare Decision Tree, Random Forest and Logistic Regression in a real case study are illustrated. Results are presented in Section 5. Ultimately conclusions and discussions are reported in Section 6.

II. LITERATURE REVIEW

Information and Communication Technologies (ICT) have transformed the academic field, not just what we teach (curriculum) or how we teach it (pedagogy), but how the institutions respond and manage these changes [6]. Researchers set out investigating the ICT impact on education in the past, in particular in the e-learning arena.

The first subsection seeks to isolate the different uses of ICT at HEIs with the aim of highlight the small research held for the academic decision-making process. Hence, the focus was on classifying the ICT used in the educational field into the following categories; e-learning, academic research, quality measurement and decision- making process.

Afterward, we provide an overview of selected works that uses Machine Learning at Higher Educational Institutions for solving academic problems. We set their stakeholder, goals, and algorithms used. Moreover, some of the ongoing applications are highlighted to establish the reliability of using these algorithms in a face to face educational model with the dataset obtained.

A. ICT USES AT HIGHER EDUCATIONAL INSTITUTIONS

In the current dynamic environmental economy sectors, in particular, the tertiary sector (service sector) has to keep track with ICT and align these technologies to satisfy stakeholder's needs and expectations. In these contexts, universities have to adapt the services they provide to develop, improve, and enhance the quality of the provided services [7].

Generally speaking, the existing studies and application of ICT at Higher Educational Institutions have primarily focused on four major streams: e-learning, academic research, a quality measurement, and decision-making process.

1) E-LEARNING

Transforming the conventional face-to-face education model through technological platforms have set up blended-learning and distance learning models [8], [9]. Having virtual classrooms has changed communication and interaction between teachers and students, education resources, and others. Although e-learning has developed new educational models, researchers have become aware of the need to sustain the development of abilities and competencies to promote intellectual capital [10]–[12].

Thus, the new advances in e-learning have created lifelong learning/teaching, the transfer of knowledge [13] and introduced new concepts like Mobile Learning (m-learning) with the increase of technology [14]. Moreover, the raising of public awareness about environmental problems demands new competencies in sustainability. Hence, a few years ago e-learning had been used to promote and improve the quality of life-long education through the acquirement of knowledge, skills, and values for Sustainable Development [15].

2) ACADEMIC RESEARCH

Academic research has been transformed by information technology due to the rapid, widespread diffusion of electronic papers, digitalization of libraries and journals, web access to information and repositories among other facilities [6]. worldwide, researchers and enterprise leaders collaborate from different perspectives in diverse projects thanks to the capability for remote exchange and communication [16]. the innovative advances on use of data (technical use) and elaboration and presentation of projects (academic use) enhance teachers' curriculum and university's visibility [17]. furthermore, information technologies serve as a control mechanism for academic misconduct like plagiarism, self-plagiarism, coercive citations, and questionable reviewing [18].

3) QUALITY MEASURE

E-learning service providers have improved their mechanism to assure the quality of their products and services [19].

Quality in education has been discussed in one of the following themes: measure the impact of knowledge [3], [20]–[22], teaching quality [23], [24], assessment quality [25] and timetabling quality [26]. Higher Educational Institutions must guarantee quality for all stakeholders: students, teachers, directors, government and society. In computer assisted-learning, the possibilities of rendering information are more numerous [27], which improves the outcomes and allows that the data insight the educational field supports quality measurement. Besides it leads us the third ICT usage at Higher Education Institutions [3], [28].

4) DECISION-MAKING PROCESS

In order to improve the decision-making process, the information first needed to be automatized, even small transactional operations such as attendance lists. Decisions made at HEIs have an administrative and academic nature. Thus, universities have computational systems to aid mostly administrative operations. Most of these are separated by departments, such as the Accounting Information System or the Academic Information System, where data can be extracted from the different silos to support the decision-making process. ICT at universities serves to help management (supporter) but also serves to improve (enabler) the decision-making process. In the organizational context some examples using Information and Communication Technologies are applied in Accounting Systems [29] and Enterprise Resource Planning [30], as well as the academic management [31].

The impact of Information and Communication Technologies depends on its infrastructure, accessibility, and the intensity of use. Although the computational advancement in processing speed and algorithms designed, has shown significant progress, more efficient and user-friendlier applications are needed when it comes to the decision-making process. Opportunely, Machine Learning arises with different algorithms that learn from data to support various task in this field.

B. MACHINE LEARNING ALGORITHMS IN THE EDUCATIONAL FIELD

Applications of Machine Learning (ML) in the academic field focused mainly on using supervised algorithms to predict students' behaviors with the possibility of early intervention. Some authors have covered different educational problems using ML such as course planning, institutions' and teachers' quality, intervention and prediction, and learning product selection. Although we focus on Higher Educational Institutions applications, other works using and comparing Machine Learning algorithms in public and private schools have been developed [3].

Educational activities datasets, such as web-log files traced from Leaning Management Systems (LMS) or Massive Open Online Couse (MOOC's) are increasingly being used to analyze students' learning behavior. Interesting examples come from various universities around the world including Massachusetts Institute of Technology [32], [33], University of Vigo [34], University of Liége [35], Open University of China [36], University of Alagoas [37] along with others, which used information from LMS and other repositories and applied different Machine Learning algorithms to predict students' performances. Table 1 highlights some successful ML works at some of the mentioned Higher Educational Institutions.

TABLE 1. Machine learning currently projects at higher educational institutions.

HEI	Goals	Stakeholder	ML Algorithm Used
University of Liége [35]	To predict freshmen´s failure. They used data from school records and environmental factors to efficient remediation and or study mediation	Students and Advisors	RF, LR, ANN
University of Alagoas [37]	To predict student's failure rates in introductory programming courses to solve educator's questions regarding these aspects	Teachers and tutors	DT, SVM, ANN, NB
Babcock University [38]	To predict students' performance and recommend necessary action through a framework of an intelligent recommender system, based on background factors	Advisors and Intelligent tutors	DT, RF, ANN
University of Jordan [22]	To predict and analyze knowledge level of learners in order to adapt content presentation and evaluation	Teachers and curriculum designers	RF, ANN, SVM, NB
Massachus etts Institute of Technolog y [32]	To predict and intervene in final course scores using process-level information. They analyze the overall prediction accuracy as well as the chronological progression of prediction accuracy	Teachers and intelligent tutors	LIR, LR, ANN, SVM
Amrita School of Engineerin g [39]	To predict the placement performance of students and propose a recommendation system to identify and pay attention to specific students' skills	Students and Teachers	DT, LR

RF = Random Forest, LR= Logistic Regression, ANN = Artificial Neural Network, DT = Decision Tree, SVM = Support Vector Machine, NB = Naïve Bayes, LIR = Linear Regression

Table 1 shows authors that successfully used various Machine Learning algorithms to compare the accuracy of them. As illustrated in Figure 1, the most popular algorithms used were Artificial Neural Network (ANN) and Support Vector Machine (SVM), followed by Decision Trees (DT), Random Forest (RF) and Logistic Regression (LR). Due to in our previous worked [5], we have already compared the use of SVM and ANN when predicting graduation rates in a public University in Colombia, with the aim of continuing that research, in this work we analyze DT, RF, and LR. The choose of these last three algorithms is done because it suits our prediction objective and have shown excellent outcomes

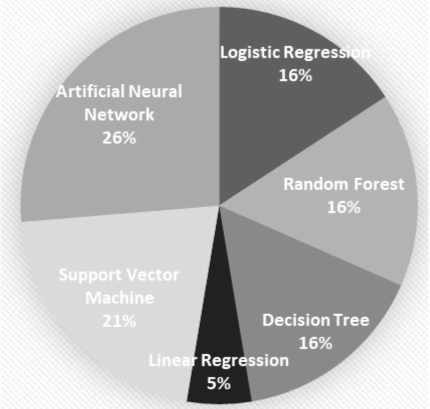


FIGURE 1. Machine learning algorithms used in related works.

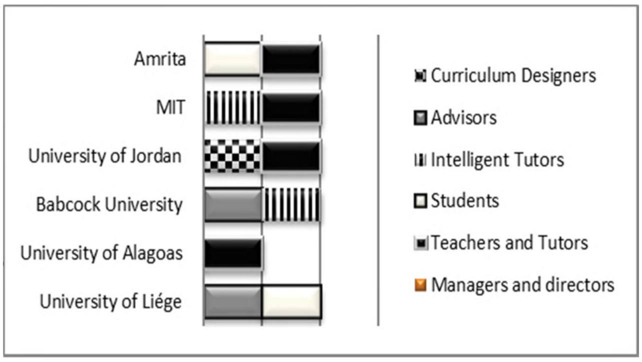


FIGURE 2. Stakeholders at reviewed works using ML executed projects.

on related works, where it shows are in the trend of usage. Furthermore, it is observed that recent studies are not focused on directors or administrators as most of the research was limited to specific stakeholders, mainly students and teachers.

As illustrated in Figure 2., although the same classification algorithms as related works are used, they differ from our research. (I) Our data comes from a face-to-face educational model. (II) Due to more features are included, algorithms architecture differ. (III) The stakeholders are directors and deans from HEIs which have particular visualization results needs, and have not been adressed before. (IV) Strategic decisions are supported when the right information is given to the high chain management as would be exposed in Section 3.

In a public university in Brazil [37] they used four prediction techniques: Support Vector Machine, Decision Tree, Neural Network and Naive Bayes to predict students' failure rates in introductory programming courses. To this aim, data was extracted from distance education. After applying data preprocessing and algorithm fine-tuning, the effectiveness of these algorithms was improved. First, they reduce the number of attributes and balance the information by applying the Synthetic Minority Over-sampling Technique. Then they fine-tune the data according to the parameters in each algorithm. Like them, we compare Machine Learning algorithms using the effectiveness metric to predict student's failure rates. Besides, our research is allocated on predictive models from educational data. However, the focus in our research is on graduation rates, involving the whole curriculum rather than a particular subject. Moreover, stakeholders and educational model are also different. Preprocessing data considering the number attributes is held to this aim.

This analysis of the related work also corroborated what was stated in our prior work [2], [5], [40], [41]; reviewed researches haven't solved Directors and Managers necessities when it comes to making a decision. Their primary goal is not focused on supporting the strategic level in universities. Although few of the works involved Machine Learning algorithms in their development, the stakeholders are mainly students and teachers and seek to fulfill their requirements and not the academic as a whole.

Therefore, in the next sections, we focus on decisions' classification at HEIs, and we use Machine Learning algorithms in a real case study to set a baseline to support directors at higher educational institutions during the decision-making process regarding issues of graduation rates.

III. DECISIONS' CLASSIFICATION AT HIGHER EDUCATIONAL INSTITUTIONS

Higher Educational Institutions are a particular type of organization of the tertiary sector. They currently hold primary responsibility for the governance and management of their finances, activities, and personnel by retaining the autonomy to decide their organizational operations. In higher education how decisions are made about institutional priorities, strategies, goals, and resource allocations, and who is held accountable for these decisions, are all functions of institutional governance [42].

Three types of governance is observed at HEIs, which influenced the operational and managerial behavior: Academic, Bureaucratic, and Corporate [43].

A. ACADEMIC

Faculty members work to retain authority and decision-making powers in areas such as teaching, curriculum, academics, and administration.

B. BUREAUCRATIC

University retains hierarchy layers with divisions of labor characterized by procedures, fixed administrations, and direct orders by higher leaders.

C. CORPORATE

Viewing the education as a service makes the students the core costumers, which is a natural consequence of taking marketing in higher education? These main marketing activities are in support of recruiting and retention efforts [44]. University follows the practice of enterprise to highlight customer needs and market competition.

Although in real world governance exhibits variations a mixture between these three categories, it is quite usual universities to prefer and implement the academic and bureaucracy styles [43]. Among HEIs are specific differences in the mission and management strategies, for instance, private universities are more market-oriented and action-oriented while public universities focus on the roles of students and alumni in the society. However, all of them work on behalf of students' success, and HEIs make their decisions to ensure it. Therefore, in order to classify the kind of decisions at HEIs, and considering their primary goal between the types of HEIs, we follow the hierarchic structure to divide work vertically according to decisions' responsibilities, which resemble an organizational pyramid structure [45].

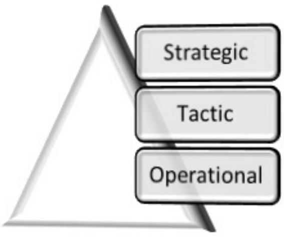


FIGURE 3. Decisions' structure levels at HEIs.

1) STRATEGIC

The uppermost level defines the policies and strategies for the organization integrating the primary goals and actions into a cohesive whole. The higher level of the institutions is the more ambitious in their strategic planning [36]. The managing positions are frequently represented by the governing board, rector and deans [46].

The managing positions are frequently represented by the governing board, rector and deans. They discuss the critical factors in strategic planning and provide guidelines for its execution. Decision-making at this level impacts the entire university. For instance, one strategic decision at this level is the number of freshmen accepted each semester. The decision affects the university's resource allocation (e.g., Budget, teachers, and facilities), as well as the society as more people, might access professional programs. At this level, according to relative works analysis, machine learning algorithms have not been used to aid this corporate stage.

2) TACTIC

The purpose of the tactic level is to identify and execute the detailed plans made at the strategic level. Generally, deans work together with the head of departments or programs directors to achieve the above planning. Intermediate directors coordinate resources usage efficiently, providing management and planning at specific times. Once the strategic planning is accepted, the tactic level is in charge of its implementation and control. Thus, quality assurance is an essential task at this level. This middle management performs decisions such as the number of students per teacher or curriculum changes. Algorithms such as Naïve Bayes and Artificial Neural Network have been used to ensure teaching, assessment and timetabling quality.

3) OPERATIONAL

The lowest level is in charge of everyday processes and through their job they sustain the whole structure. Specific tasks and transactional activities are performed to support the operations of the institution. This level holds the majority of Information Technology requires by HEIs. IT governance in this stage works as an instrument to control and manage the IT resources such as infrastructure technology and people [47]. Collaborators as teachers, advisors, tutors, programs assistants, and secretaries, among others, execute their task according to the guidelines provided by the strategic and tactic level. Although decision-making at this level affects a smaller population within the university, it might impact students' success (i.e., schedule and timetabling evaluation) and operational mechanisms (i.e., subject registration process). At the state-of-the-art examination, we found that most of the works done using machine learning on the educational field are a focus on the operational stage. Some of the prominent algorithms used are Artificial Neural Networks and Support Vector Machine.

We observed that each of the levels in the pyramid have a decision-making process that, from top to bottom, affect a more significant portion of the community. Although some software applications can support decision-making processes, higher levels generally work with the information provided by the operational stage and data is not analyzed and visualized easily to support decisions at high stages. Norman and Ahmed identified the central software misfit in Higher Education planning software. Some of the cases they stated are poor consultant effectiveness, poor-reliance on heavy customization, reduced IT infrastructure, poor project management effectiveness, poor management support, too tight project schedule and poor knowledge transfer [30].

Two global statistics are frequently cited as measures of student success: the cohort graduation rate and the freshman-to-sophomore retention rate. Thus, Faculty decisions should focus on their enhancement. Students' persistence to complete their educational goals are a key gauge of student success, and therefore institutional success [44]. Hence, in the next section, we propose the usage of Machine Learning algorithms to predict graduation rates and collaborate with academic decisions on behalf of student's success. Moreover, we set a baseline to support decision making at the strategic level at HEIs according to directors' needs analyzed in our previous work.

IV. METHOD

One of the main goals of this study is to compare the effectiveness of existing Machine Learning algorithms in predicting graduate rates that will support decision making at the strategic level. Thus, we will be classifying student's academy performance to predict the number of graduated and not graduated students, being this our objective variable.

First Subsection presents data characterization, cleaning and preparation. Subsequently, subsection B contextualize the three Machine Learning algorithms used in this work. We present their basics, method, architecture, and configuration used. The tools and metrics used are respectively indicated in subsection C and D.

A. DATA COLLECTION AND PREPARATION

The experiment was conducted with real data from a public university in Colombia. The data source contains information from 6100 engineer students. We analyzed students from five different engineer programs such as industrial, cadastral, systems, electrical and electronical engineering, enrolled during the ten years period of 2004-2014. This number of students is assumed after disregarding cases of missing data such as students who leave during the first semester and students who enter the university after 2009. Students who enter after 2009 will not graduate before 2014 because engineering careers take ten semesters to graduate and therefore will not address the supervised algorithms needs. More than 55200 records were available to analyze.

As our case study is a public university, data policies are strict. Although our research was restricted by their data-protection policies, and we lack information about students' gender or age among other socio-demographic data, for the most part we use students' academic records to held the graduation rates prediction. We believe the inclusion of socio-demographic and socio-economic data would be worth to analyze in the future. However, in this study the academic information obtained is efficient to analyze the insightful outcomes.

The students' academic features include in all three algorithms are stated in Table 2.

Once the classification objective was set (i.e., graduated and not graduated students) and data was acquired, we conducted the following steps to build every algorithm model:

i. Using the scaling method, we transformed data by giving them values from a range [0, 1] 0 as a minimum and a maximum of 1.

ii. Set initial hyperparameters for each algorithm.

iii. Using a stratified sampling technique, split the dataset into two subsets 70% for training and 30% for prediction, to keep the data distribution. The sampling method alleviate the effect of class imbalance problem as one of the most employed method [48].

iv. From the training subset in step three, we use 5-fold cross validation technique to tune the hyperparameters in each algorithm.

v. We execute the algorithms with the initial settings. In each k-fold we save the accuracy obtained as well as the values entered in each hyper parameter, to adjust them in each run and encounter the most suitable values for them until the accuracy reached the expectations.

vi. Finally the algorithms are executed to train the whole training set using the best values obtained for the hyperparameters in each algorithm. Hyperparameters set in each algorithm as well as the architecture and contextualization of each algorithm is exposed in the next subsection.

TABLE 2. Variable definitions and type of measurement.

Variable	Measurement
Arithmetic mean of the final grade of all subjects coursed	Grade from 0.0 to 5.0
Subjects enrolled in each semester	Quantity
Subjects satisfactory approved in each semester	Quantity
Subjects failed in each semester	Quantity
Subjects that were failed but after an extra summary exam become approved (validated)	Quantity
Subjects enrolled in each semester	Quantity
Median grade	Grade from 0.0 to 5.0
Maximum grade	Grade from 0.0 to 5.0
Minimum grade	Grade from 0.0 to 5.0
The range of the grades	Quantity
The middle grade between the smallest grade and the median of the data set	Quantity
Median grade	Quantity
The middle grade between the median and the highest grade of the data set	Quantity
The square root of the average of the squared deviations of the values subtracted from their average value	Quantity
Summary of the subjects enrolled during the whole career	Quantity
Summary of the satisfactory approved subjects during the whole career	Quantity
Summary of the failed approved subjects during the whole career	Quantity
Summary of the subjects that become approved after the validation exam during the whole career	Quantity
The middle grade between the smallest grade and the median of the data set	Quantity
Median grade	Quantity
The middle grade between the median and the highest grade of the data set	Quantity
The square root of the average of the squared deviations of the values subtracted from their average value	Quantity
Summary of the subjects enrolled during the whole career	Quantity
Summary of the satisfactory approved subjects during the whole career	Quantity
Summary of the failed approved subjects during the whole career	Quantity
Summary of the subjects that become approved after the validation exam during the whole career	Quantity
Socio-economic status according to the students' residence address	{1,2,3,4,5,6}

B. MACHINE LEARNING ALGORITHM ARCHITECTURE

1) DECISION TREE (dt)

It is a highly used classifiers due to its simplicity for understanding and interpretation. It requires little data preparation, handles numerical and categorical data, and performs very well with large data set in a short time [49]. Additionally, the hierarchical tree structure resembles a human way of decision-making, providing extending information about the sequence to classify and individually into a class, discovering rules in a more comprehensible manner [50]. In our case study, the classification falls either into "graduated" or "not graduated." DT is a type of supervised learning algorithm that is mostly used for classification problems. Surprisingly, it works for both categorical and continuous dependent variables. Although there are many specific decision tree algorithms (e.g., ID3, C4.4, C5.0, CART, and CHAID), we worked with the most popular developed by Quinlan. C5.0 is significantly faster and more efficient than its predecessors C4.5 and ID3. C5.0 supports boosting, which gives the trees more accuracy.

Furthermore, it allows the weighting of different attributes and misclassification types and separates the data automatically to help reduce noise [20]. When constructing decision trees, it is essential to find the best splitting point measurement (i.e., information gain, gain ratio, Gini index, and entropy measure). The selection of the split attribute should directly decide the learning trend [51]. Gini index is used as a split measure for choosing the most appropriate splitting attribute for each node. The split function has this form:

$$I_g(p) = 1 - \sum\nolimits_{i=1}^{J} p_i^2 \tag{1}$$

Gini index is mathematically shown above; whit J classes suppose I ∈ {1, 2, 3, ...J} and pi is the fraction of items labeled with class 'i' in the set. The data set is portioned into two subsets. The split can be made at the highest value of the lower partition, at the lowest value of the lower partition, or at the average of the two. We choose these last split point to force the model only binary splits on nominal columns. On the other hand, the pruning method used is the Minimum Description Length (MDL) which reduces the rate of misclassification, provides high accuracy and fast execution times.

2) RANDOM FOREST

RF is a classifier that combines the performances of numerous Decision Trees algorithms to predict the value of a variable [52].

Each tree in the forest gives a classification and "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest). RF Regression predictor has the form:

$$f_{rf}^{K}(x) = \frac{1}{K}\sum\nolimits_{K=1}^{k} T(x) \tag{2}$$

When RF receives and (x) input vector, made up of the values of different evidential features analyzed for a given training area, RF builds a number K of regression trees and averages the results. After K such trees $\mathrm{T(x)}_1^{\mathrm{K}}$ are grown. About the architecture used, the split attribute chosen was Gain Radio since it did not represent a significant difference among the other options provided by knime (Information Gain, Information Gain Radio y Gini Index). Moreover, the number of models used was one hundred mainly because of the dataset size and testing results. The number of decision trees to learn of number of models set is 100.

Additionally, the use of static random seed is required y knime to start the prediction, and the one automatically generated by the software was used (i.e., 1508210392822).

3) LOGISTIC REGRESSION

This classification algorithm could be confusing by its name. It is used to estimate discrete values (e.g., binary values) based on a given set of independent variables.

It is known as logit regression because it predicts the probability of an occurrence, of any event, by fitting data into a logit function [53]. Logistic regression gives linear class boundaries. Due to the fact it uses an 'S'-shaped curve instead of a straight line it is a natural fit for dividing data into groups.

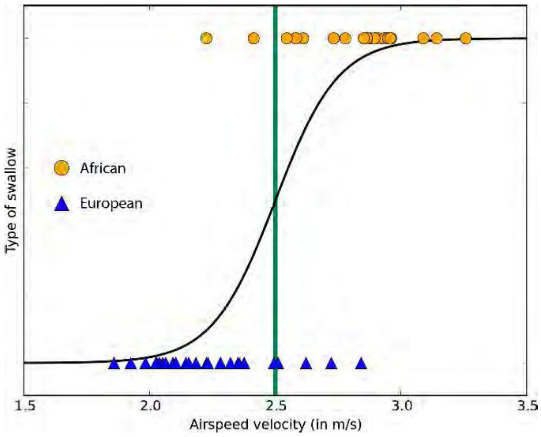


FIGURE 4. Logistic regression representation of two-class data with just one feature [54].

Figure 4 shows an example took from Azure of logistic regression to two-class data with just one feature. The class boundary is the point at which the logistic curve is just as close to both classes [54].

Target column used was Objective (i.e., graduated or not graduated) with true as the reference category. Stochastic Average Gradient (SAG) was used as the solver mainly because it minimizes the negative of the Log Likelihood function and supports regularization. It relies on the idea of the gradient descent method meaning that in each interaction the algorithm moves into a descending direction of the negative of Log Likelihood function with step size Δ., which is called the learning rate. The learning rate strategy was fixed to a step size of 0.2. Finally, the termination conditions were set at a maximal number of epochs (i.e., 200) and Epsilon (i.e., $\varepsilon = 0.001$).

C. INSTRUMENTATION

To perform the pre-processing of data and all the Machine Learning algorithms we used KNIME 3.4.0 (Konstanz Information Miner) analytic platform [55]. KNIME is open-source software, developed in Java which allows ETL processes (Extraction, Transformation, and Loading) in addition to various modular components for machine learning and data mining.

D. EFFECTIVENESS METRICS

To evaluate the performance of the compared Machine Learning algorithms, we use the area under the curve (AUC) as the evaluation criteria. AUC is a popular measure for ranking class performance of the learned classifiers [56] AUC is calculated as follows:

$$E = \frac{P_0 - \frac{t_0(t_0+1)}{2}}{t_0 t_1} \tag{3}$$

where t0 and t1 are numbers of negative and positive instances respectively. P0 = ∑ri denoting the rank of the nth negative instance in the ranked list. This equation can only handle a two-level class problem corresponding to the prediction sought (graduated or not graduated).

Moreover, the use of a confusion matrix is useful to present the prediction results of the test. If the data set contains 'n' distinct classes, the confusion matrix is an (n×n) matrix [57]. Since this case examines only two types of data (graduated or not graduated), we have a (2×2) confusion matrix indicated in Table 3.

TABLE 3. Confusion matrix.

Prediction	Graduated	Not graduated
Actual		
Graduated	A	C
Not Graduated	D	B

From the confusion matrix data, we defined the overall accuracy, the precision rate and the recall rate as follows [58]: The precision rate of the graduated class $= A/(A + D)$ the precision rate of the not graduated class $= B/(B + C)$. The recall rate of the graduated class $= A/(A + C)$ and the recall rate of the not graduated class $B/(B + D)$ and the overall accuracy $(A + B)/(A + B + C + D)$.

V. RESULTS

Once the 55200 records were divided into the training set (i.e., 70%) and test set (i.e., 30%) three machine learning methods were used to process the test set (i.e., Decision Trees, Random Forests and Logistic Regression). First, we used the Receiver Operating Characteristic Curve (ROC) as a standard metric for the binary outcome expected (graduated or not graduated.) AUC helps to reduce the ROC curve to a single value, representing the expected performance of the classifier.

The x-axis indicates the false positive rate, while the y-axis indicates the true positive rate. It is clear that the AUC for Logistic Regression (i.e., 0.9028) represented in Fig. 5 is slightly higher than Random Forest's AUC (i.e., 0.8994) and Decision Trees (i.e., 0.8830). One of the goals of this study is to identify the potential graduated students. Comparatively, RF was the most effective and more precise in predicting student graduates.

Table 4 shows the results of the methods mentioned in section 4.3. In terms of the precision rate and the recall rate of the graduated class, there was little difference in the performance of the three prediction models: Random Forest has the highest recall rate (91.93%), followed by Decision Tree (91.38%) and Logistic Regression with the lowest recall rate (90.935 %). However, regarding the precision rates of the same class RF is the lowest (86.44%), followed by DT (86.61%) and LR (87.04%) with the highest precision rate.

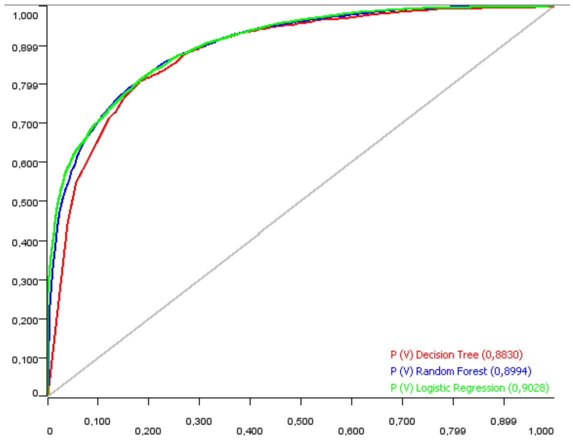


FIGURE 5. Shows ROC curves for the three algorithms compared.

TABLE 4. Evaluation of prediction results.

Evaluation index	DT	RF	LR
Recall rate of graduated class	91.38%	91.93%	90.93%
Recall rate of not graduated class	65.91%	65.21%	67.33%
Precision rate of graduated class	86.61%	86.44%	87.04%
Precision rate of not graduated class	76.01%	77.00%	75.46%
Overall Effectiveness|	83.92%	84.11%	84.02%

Among the three prediction models, RF had the highest overall accuracy (84.11%), followed by LR (84.02%), while the DT had the lowest accuracy rate (83.92%). The three models had a relatively high overall accuracy rate that exceeded 83%. With the aim of comparing the impact of the first features include in each training algorithm, we test them using 6, 15 and 19 features as shown in Table 5. We observe that the overall accuracy increases as the number of features tested increase. Revealing more features leads to obtaining greater accuracy. For instance, Random forest accomplishes an accuracy enhancement of 4.4%. Considering 10000 students and increase from 79.71% to 81.11% would represent more than 400 students correctly classified. Algorithm's improvement also relies on data distribution and data set.

Compare to related works mentioned in Section 2 [22], [32], [35], [37]–[39]; our results indicate an accuracy rate above of the 83%, which is more than four percentage point higher accuracy from other researchers. Moreover, an additional comparison of including more features in the initial training set was developed to evaluate whether it is possible to obtain higher accuracy results. Table 5., shows the augment of the overall accuracy when more data features are involved.

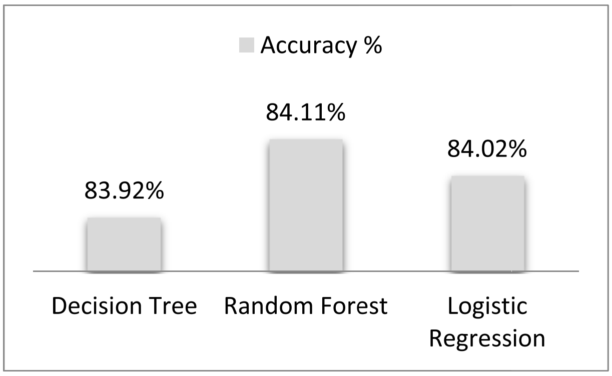


FIGURE 6. Comparative overall accuracy of the algorithms tested. First.

TABLE 5. Overall accuracy comparison by initial features tested.

	6 Features	9 Features	16 Features
Decision Tree	79,57	81,34	83,92
Random Forest	79,71	81,34	84,11
Logistic Regression	79,55	81,14	84,02

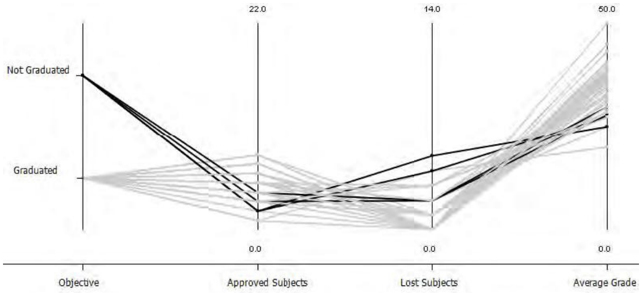


FIGURE 7. Relevant features to Graduate. The figure shows three relevant academic features: student's graduation: approved subjects, lost subjects, and average grade.

Visualization of the information is an imperative task in any analytics process. Through KNIME data visualization is available, including scatter plots, pie charts, box plots, histograms, and others. From the data collected at District University Francisco Jose de Caldas (JFC), a parallel coordinate are plotted to analyze multivariate data: features tested regarding grades.

The Parallel Coordinates plots just 50 records were taken randomly to this plot. Black color represents not graduated students, while gray color graduated students. First vertical parallel line from left to right, represents the objective sought (graduated or not graduated), subsequently is approved subjects, followed by lost subjects and average grade. We observe that graduated students (gray lines) reach a better average grade and more approved subjects. By contrast, not graduated students (black lines) approve fewer subjects and reach lower average grades. These patterns are easily recognized thanks to machine learning algorithm.

Depict insightful information through an intentional graphic, improve not just the data' understanding through the recognition of trends and hidden patterns, but enhance the actions taken from the data processed. For instance, Figure 7, allows HEIs Administrators to observe how the average grade and the number of lost subjects impact the fact that students get graduated or not, to mention some of the features that are analyzed from the data set. Although some graphics provide by KNIME are useful, we propose in the next section significant improvements to succeed in the information launch and discuss potential efficient metrics that could be added.

VI. CONCLUSION AND DISCUSSIONS

Even though at educational field Machine Learning is still emerging, its effectiveness to analyze information is notorious. Through the analysis, predictions, and visualizations of information, HEIS' directors obtain a greater understanding of the different variables involved when making a decision. Machine Learning supports this process providing various algorithms suitable to the different kinds of data and the different kinds of predictions required.

We employ three supervised classification algorithms: Decision Trees, Random Forests and Logistic Regression, where Random Forest performs the best outcomes.

From a recent literature review [3], [34], we observed not just their efficiency but also their acceptance on the research field when contributing to decision processes. Results indicate that the three tested Machine Learning algorithms can identify at least an 83% accuracy rate, which is a slightly higher rate reached compare to related works (i.e., $\overline{X} = 79\%$). Using these experiments the early identification of students likely to not graduate is highly effective, although from this prediction other aspects such as students' academic performance and dropout rates could be analyzed.

Continuing with the research done in previous work [5] we compare different Machine Learning algorithms in this paper as well as we analyze decision's structure at HEIs and how they are managed according to the institutional governance. Having these students recognized early can allow HEIs governance strategic planning abilities are respecting students' exclusion policies, students' dropout rates, retention rates, strengthen programs, and a whole host of others.

The comparison of ROC, accuracy, precision rates and recall rates were conducted. We observed that the overall accuracy is prevailing in Random Forest although the area under the curve (AUC) is slightly superior in Logistic Regression. However, as effectiveness metric, accuracy is more significance than ROC curve due to ROC being insensitive to data sets with unbalanced proportion classes [59]. Regarding precision rate and recall rate, the three algorithms are similar.

Therefore, a more in-depth analysis of the number of features tested, or data normalization, will be presented in a forthcoming paper. Future research will also include the analysis of other effectiveness metrics such as F-Measure or Specificity as well as the comparison with other classification algorithms, that would be worth to analyze in other Higher Educational Institutions. Moreover, would be worth to include socio-demographic and socio-economical information about the students when analyzing the variables that might affect the graduation rates at Higher Educational Institutions,

Performing this recognition has managerial implications not just for reducing processing time but increasing reliability on its prediction.

This last is essential due to help during the decision-making process with insightful information that cooperates improving the decision effectiveness even in the different educational field such as resource planning, teacher's management, curriculum design, and related factors. Although the visualization nodes provided by KNIME are efficient to represent the results for this study, a future goal for this research includes the development of a computational tool for deans and university administrators. Specific technician knowledge is required to understand the reports and graphics on KNIME, and a more straightforward interpretation of the academic data and predictions using Human Computer Interaction techniques would significantly support the decisions making process.

REFERENCES

[1] A. Lašáková, L. Bajzíková, and I. Dedze, "Barriers and drivers of innovation in higher education: Case study-based evidence across ten European Universities," *Int. J. Educ. Develop.*, vol. 55, pp. 69–79, May 2017.

[2] Y. V. N. Acevedo and C. E. M. Marín, "System architecture based on learning analytics to educational decision makers toolkit," *Adv. Comput. Sci. Eng.*, vol. 13, no. 2, pp. 89–105, 2014.

[3] A. R. T. Góes, M. T. A. Steiner, and P. J. S. Neto, "Education quality measured by the classification of school performance using quality labels," *Appl. Mech. Mater.*, vols. 670–671, pp. 1675–1683, Oct. 2014.

[4] H. Lounis and T. Fares, "Using efficient machine-learning models to assess two important quality factors: Maintainability and reusability," in *Proc. Joint Conf. 21st Int. Workshop Softw. Meas. 6th Int. Conf. Softw. Process Product Meas.*, Nov. 2011, pp. 170–177.

[5] Y. Nieto, V. García-Díaz, C. Montenegro, and R. G. Crespo, "Supporting academic decision making at higher educational institutions using machine learning-based algorithms," *Soft Comput.*, vol. 23, no. 12, pp. 4145–4153, 2018.

[6] K. V. Pincus, D. E. Stout, J. E. Sorensen, K. D. Stocks, and R. A. Lawson, "Forces for change in higher education and implications for the accounting academy," *J. Accounting Educ.*, vol. 40, pp. 1–18, Sep. 2016.

[7] A.-P. Pavel, A. Fruth, and M.-N. Neacsu, "ICT and e-learning—Catalysts for innovation and quality in higher education," *Procedia Econ. Finance*, vol. 23, pp. 704–711, Jan. 2015.

[8] F. V. Elena, A. M. Manuel, and G. G. S. Carina, "Which use give teachers at La Laguna University to ICTs?" *Procedia-Social Behav. Sci.*, vol. 93, pp. 1646–1651, Oct. 2013.

[9] H. M. Vo, C. Zhu, and N. A. Diep, "The effect of blended learning on student performance at course-level in higher education: A meta-analysis," *Stud. Educ. Eval.*, vol. 53, pp. 17–28, Jun. 2017.

[10] I. M. Taucean and M. Tamasila, "Research challenges for eLearning support in engineering and management training," *Procedia-Social Behav. Sci.*, vol. 124, pp. 210–218, Mar. 2014.

[11] E. Faham, A. Rezvanfar, S. H. M. Mohammadi, and M. R. Nohooji, "Using system dynamics to develop education for sustainable development in higher education with the emphasis on the sustainability competencies of students," *Technol. Forecast. Social Change*, vol. 123, pp. 307–326, Oct. 2017.

[12] F. A. Goni, A. G. Chofreh, M. Mukhtar, S. Sahran, S. A. Shukor, and J. J. Klemeš, "Strategic alignment between sustainability and information systems: A case analysis in Malaysian public higher education Institutions," *J. Clean. Prod.*, vol. 168, pp. 263–270, Dec. 2017.

[13] I. González-González and A. I. Jiménez-Zarco, "Using learning methodologies and resources in the development of critical thinking competency: An exploratory study in a virtual learning environment," *Comput. Hum. Behav.*, vol. 51, pp. 1359–1366, Oct. 2015.

[14] F. Moreira, M. J. Ferreira, C. P. Santos, and N. Durão, "Evolution and use of mobile devices in higher education: A case study in Portuguese higher education Institutions between 2009/2010 and 2014/2015," *Telematics Inform.*, vol. 34, no. 6, pp. 838–852, Sep. 2017.

[15] U. M. Azeiteiro, P. Bacelar-Nicolau, F. J. P. Caetano, and S. Caeiro, "Education for sustainable development through e-learning in higher education: Experiences from Portugal," *J. Clean. Prod.*, vol. 106, pp. 308–319, Nov. 2015.

[16] K. H. Wang, N. J. Ray, D. N. Berg, A. T. Greene, G. Lucas, K. Harris, A. Carroll-Scott, B. Tinney, and M. S. Rosenthal, "Using community-based participatory research and organizational diagnosis to characterize relationships between community leaders and academic researchers," *Preventive Med. Rep.*, vol. 7, pp. 180–186, Sep. 2017.

[17] A. Valentín, P. M. Mateos, M. M. González-Tablas, L. Pérez, E. López, and I. García, "Motivation and learning strategies in the use of ICTs among University students," *Comput. Educ.*, vol. 61, pp. 52–58, Feb. 2013.

[18] C. Hopp and G. A. Hoover, "How prevalent is academic misconduct in management research?" *J. Bus. Res.*, vol. 80, pp. 73–81, Nov. 2017.

[19] S. E. Alptekin and E. E. Karsak, "An integrated decision framework for evaluating and selecting e-learning products," *Appl. Soft Comput.*, vol. 11, pp. 2990–2998, Apr. 2011.

[20] D. Delen, H. Zaim, C. Kusey, and S. Zaim, "A comparative analysis of machine learning systems for measuring the impact of knowledge management practices," *Decis. Support Syst.*, vol. 54, pp. 1150–1160, Jan. 2013.

[21] N. Cohen, J. Gattuso, and K. MacLennan-Brown, *CCTV Operational Requirements Manual*, no. 28. London, U.K.: Home Office Scientific Develpoment Branch, 2009.

[22] N. Ghatasheh, "Knowledge level assessment in e-learning systems using machine learning and user activity analysis," *Int. J. Adv. Comput. Sci. Appl.*, vol. 6, no. 4, pp. 107–113, 2015.

[23] R. Rodríguez and G. Rubio, "Teaching quality and academic research," *Int. Rev. Econ. Educ.*, vol. 129, pp. 10–27, Sep. 2016.

[24] X.-Y. Liu, "Private colleges teachers evaluation system based on support vector machine (SVM)," in *Proc. Int. Conf. Appl. Sci. Eng. Innov. (ASEI)*, 2015, pp. 1918–1921.

[25] K. J. Gerritsen-van Leeuwenkamp, D. Joosten-ten Brinke, and L. Kester, "Assessment quality in tertiary education: An integrative literature review," *Stud. Educ. Eval.*, vol. 55, pp. 94–116, Dec. 2017.

[26] A. Muklason, A. J. Parkes, E. Özcan, B. McCollum, and P. McMullan, "Fairness in examination timetabling: Student preferences and extended formulations," *Appl. Soft Comput.*, vol. 55, pp. 302–318, Jun. 2017.

[27] I. Smeureanu and N. Isaila, "Information technology, support for innovation in education sciences," *Procedia-Social Behav. Sci.*, vol. 15, pp. 751–755, Jan. 2011.

[28] H. Lakkaraju, E. Aguiar, C. Shan, D. Miller, N. Bhanpuri, R. Ghani, and K. L. Addison, "A machine learning framework to identify students at risk of adverse academic outcomes," in *Proc. Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2015, pp. 1909–1918.

[29] S. M. Shuhidan, N. Mastuki, and W. M. N. W. M. Nori, "Accounting information system and decision useful information fit towards cost conscious strategy in Malaysian higher education institutions," *Procedia Econ. Finance*, vol. 31, no. 15, pp. 885–895, 2015.

[30] A. Y. Noaman and F. F. Ahmed, "ERP systems functionalities in higher education," in *Proc. Int. Conf. Commun., Manage. Inf. Technol.*, vol. 65, pp. 385–395, Jan. 2015.

[31] T. Anastasios, S. Cleo, P. Effie, T. Olivier, and M. George, "Institutional research management using an integrated information system," *Procedia-Social Behav. Sci.*, vol. 73, pp. 518–525, Feb. 2013.

[32] A. J. Stimpson and M. L. Cummings, "Assessing intervention timing in computer-based education using machine learning algorithms," *IEEE Access*, vol. 2, pp. 78–87, 2014.

[33] S. Boyer, B. U. Gelman, B. Schreck, and K. Veeramachaneni, "Data science foundry for MOOCs," in *Proc. IEEE Int. Conf. Data Sci. Adv. Anal.*, Oct. 2015, pp. 1–10.

[34] C. Gonzalez, E. Elhariri, N. El-Bendary, and A. Fernandez, "Machine learning based classification approach for predicting students performance in blended learning," in *Proc. 1st Int. Conf. Adv. Intell. Syst. Inform. (AISI)* (Advances in Intelligent Systems and Computing), vol. 407. Switzerland: Springer, 2016, pp. 47–56.

[35] A.-S. Hoffait and M. Schyns, "Early detection of University students with potential difficulties," *Decis. Support Syst.*, vol. 101, pp. 1–11, Sep. 2017.

[36] M. Tan and P. Shao, "Prediction of student dropout in e-Learning program through the use of machine learning method," *Int. J. Emerg. Technol. Learn.*, vol. 10, no. 1, pp. 11–17, 2015.

[37] E. B. Costa, B. Fonseca, M. A. Santana, F. F. de Araújo, and J. Rego, "Evaluating the effectiveness of educational data mining techniques for early prediction of students' Academic failure in introductory programming courses," *Comput. Hum. Behav.*, vol. 73, pp. 247–256, Aug. 2017.

[38] M. Goga, S. Kuyoro, and N. Goga, "A recommender for improving the student academic performance," *Procedia-Social Behav. Sci.*, vol. 180, pp. 1481–1488, May 2015.

[39] S. K. Thangavel, P. D. Bkaratki, and A. Sankar, "Student placement analyzer: A recommendation system using machine learning," in *Proc. 4th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS)*, Jan. 2017, pp. 1–5.

[40] Y. V. Nieto, "Modelo de un sistema de software basado en las técnicas de learning analytics como herramienta de apoyo en la toma de decisiones Académico-administrativas en las Instituciones públicas de educación superior," Univ. Distrital Francisco José de Caldas, Bogotá, Colombia, Tech. Rep. 131115, 2015.

[41] Y. V. Nieto, V. G. Diaz, and C. E. Montenegro, "Academic decision making model for higher education institutions using learning analytics," in *Proc. 4th Int. Symp. Comput. Bus. Intell. (ISCBI)*, Sep. 2016, pp. 27–32.

[42] A. Clark, "IT Governance: Determining who decides," *EDUCAUSE Center Anal. Res. Bull.*, vol. 2005, no. 24, pp. 1–13, Nov. 2005.

[43] S. J. Chan and C. Y. Yang, "Governance styles in Taiwanese universities: Features and effects," *Int. J. Educ. Develop.*, vol. 63, pp. 29–35, Nov. 2018.

[44] M. Guilbault, "Students as customers in higher education: The (controversial) debate needs to end," *J. Retail. Consum. Services*, vol. 40, pp. 295–298, Jan. 2018.

[45] G. R. Jones, *Organizational Theory, Design, and Change*, 7th ed. Upper Saddle River, NJ, USA: Prentice-Hall, 2011.

[46] J. Hu, H. Liu, Y. Chen, and J. Qin, "Strategic planning and the stratification of Chinese higher education institutions," *Int. J. Educ. Develop.*, vol. 63, pp. 36–43, Nov. 2018.

[47] I. S. Bianchi and R. D. Sousa, "IT Governance mechanisms in higher education," *Procedia Comput. Sci.*, vol. 100, pp. 941–946, Jan. 2016.

[48] L. Nanni, C. Fantozzi, and N. Lazzarani, "Coupling different methods for overcoming the class imbalance problem," *Neurocomputing*, vol. 158, pp. 48–61, Jun. 2015.

[49] A. Trabesli, Z. Elouedi, and E. Lefevre, "Decision tree classifiers for evidential attribute values and class labels," *Fuzzy Sets Syst.*, vol. 366, pp. 46–62, Jul. 2019.

[50] S. Helal, J. Li, L. Liu, E. Ebrahimie, S. Dawson, D. J. Murray, and Q. Long, "Predicting academic performance by considering student heterogeneity," *Knowl.-Based Syst.*, vol. 161, pp. 134–146, Dec. 2018.

[51] H. Sun and X. Hu, "Attribute selection for decision tree learning with class constraint," *Chemom. Intell. Lab. Syst.*, vol. 163, pp. 123–129, Apr. 2017.

[52] V. Rodriguez-Galiano, M. Sanchez-Castillo, M. Chica-Olmo, and M. Chica-Rivas, "Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines," *Ore Geol. Rev.*, vol. 71, pp. 804–818, Dec. 2015.

[53] S. K. Singh, R. W. Taylor, M. M. Rahman, and B. Pradha, "Developing robust arsenic awareness prediction models using machine learning algorithms," *J. Environ. Manage.*, vol. 211, pp. 125–137, Apr. 2018.

[54] M. Azure. (2017). *How to Choose Machine Learning Algorithms*. Accessed: Jan. 17, 2018. [Online]. Available: https://docs.microsoft.com/es-es/azure/machine-learning/studio/algorithm-choice

[55] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, and B. Wiswedel, "KNIME: The Konstanz information miner," in *Data Analysis, Machine Learning and Applications* (Studies in Classification, Data Analysis, and Knowledge Organization). Germany: Springer, 2007.

[56] Y. Zhang, J. Wu, C. Zhou, and Z. Cai, "Instance cloned extreme learning machine," *Pattern Recognit.*, vol. 68, pp. 52–65, Aug. 2017.

[57] D. J. Yu, J. Hu, Q. M. Li, Z. M. Tang, J. Y. Yang, and H. B. Shen, "Constructing query-driven dynamic machine learning model with application to protein-ligand binding sites prediction," *IEEE Trans. Nanobiosci.*, vol. 14, no. 1, pp. 45–58, Jan. 2015.

[58] M. R. K. Darzi, S. T. A. Niaki, and M. Khedmati, "Binary classification of imbalanced datasets: The case of CoIL challenge 2000," *Expert Syst. Appl.*, vol. 128, pp. 169–186, Aug. 2019.

[59] G. Hackeling, *Mastering Machine Learning With Scikit-Learn*. Birmingham, U.K.: Packt Publishing, 2014.

YURI NIETO received the M.Sc. degree in computer science and communications from Francisco Jose de Caldas District University. She is currently pursuing the Ph.D. degree with the Computer Science Program, University of Oviedo. In 2012, she was an Industrial Engineer with the AXON and GIIRA Investigation Groups, where she is currently a member. Her research interests include machine learning, decision support systems, learning analytics, distributed systems, and virtualization.

VICENTE GACÍA-DÍAZ received the Ph.D. degree in computer science from the University of Oviedo, in 2011, where he is currently a Software Engineer and an Associate Professor with the Department of Computer Science. He has supervised over 60 academic projects and published over 70 research papers in journals, conferences, and books. His research interests include machine learning, natural language processing, model-driven engineering, and domain-specific languages. He is also part of the Editorial and Advisory Board of several journals and has been an Editor of several special issues in books and journals.

CARLOS MONTENEGRO received the Diploma of Advanced Studies degree from the Pontifical University of Salamanca, in 2008, the M.Sc. degree in information and communication systems from the Universidad Distrital Francisco José de Caldas, and the Ph.D. degree in systems and computer services for the Internet from the University of Oviedo, Asturias, Spain, in 2012. He is currently a Systems Engineer. His research interest includes object-oriented.

CLAUDIO CAMILO GONZÁLEZ received the M.Sc. degree in urban development from San Buenaventura University, Colombia. He is currently pursuing the Ph.D. degree in educational technology with Lleida University. He is also a System Engineer and a member of the Byte and Desing Research Group, UNAD University, Colombia. His research interests include education, e-learning, and decision support systems.

RUBÉN GONZÁLEZ CRESPO received the Ph.D. degree in computer science engineering. He is currently the Dean of the Higher School of Engineering, UNIR, and the Director of the AENOR (Spanish Association for Standardization and Certification) Chair in Certification, Quality and Technology Standards. He is also a member of different committees at the ISO Organization. He is also an Advisory Board Member of the Ministry of Education at Colombia and an Evaluator of the National Agency for Quality Evaluation and Accreditation of Spain (ANECA).

De: Dr. Imran Bajwa
A: Miguel de la Uz Cuervo; VICENTE GARCIA DIAZ; cemontenegrom@udistrital.edu.co
Asunto: [J.UCS] Manuscript ID: jucs-18007 - Accepted for Publication
Fecha: jueves, 23 de mayo de 2019 9:24:54

Dear Yuri Vanessa,

We are pleased to inform you that the following paper has been officially accepted for publication:

Manuscript ID: jucs-18007
Type of manuscript: Article
Title: Decision-making model at Higher Educational Institutions based on Machine Learning
Authors: Yuri Vanessa Nieto, Vicente García-Díaz, Carlos Enrique Montenegro

Received: 21 December 2018
E-mails: uo250052@uniovi.es, garciavicente@uniovi.es, cemontenegrom@udistrital.edu.co
Submitted to section: Intelligent Computing for Society,
http://www.jucs.org/ujs/jucs/info/special_issues/JUCS_Intelligent_Computing_for_Society_2019_CfP.html
Journal of Universal Computing Science
http://www.jucs.org/jucs_25_2

We will start the final preparations for publication now.

Kind regards,
Dr. Imran Sarwar Bajwa
Managing Guest Editor - J.UCS Special Issue
E-Mail: imran.sarwar@iub.edu.pk

Decision-making model at Higher Educational Institutions based on Machine Learning

Yuri Vanessa Nieto
(University of Oviedo, Oviedo, Spain
uo250052@uniovi.es)

Vicente García-Díaz
(University of Oviedo, Oviedo, Spain
garciavicente@uniovi.es)

Carlos Enrique Montenegro
(District University Francisco José de Caldas, Bogotá, Colombia
cemontenegrom@udistrital.edu.co)

Abstract: At Higher Educational Institutions (HEI) the high hierarchical managers and directors face many challenges during the decision-making process, that sometimes are rely on intuition, and past experiences, leading not just to delays but the low impact in the whole academic community. A decision-making model for managers and administrator of HEIs is presented. We propose a detailed methodology when academic prognosis is taking place. The comparison between five robust Machine Learning algorithms is executed accomplishing outperformed results by Support Vector Machine. As a validation experiment, we executed the proposed decision model in a face-to-face public university in Colombia, showing the results in a developed web platform prototype with its correspondent architecture. Moreover, we discuss the social implication of low graduation rates.

Keywords: Decision Support System, Decision-making, Machine Learning, Support Vector Machine, Classification algorithms
Categories: I.2, I.5, L.2, L.3.

1 Introduction

The primary mission of a Higher Educational Institution is to educate highly competent professionals that support the advancement of the region where they developed. Hence, the cohort graduation rate has been a worldwide accepted indicator of the students and university success [Guilbault 2017]. Education plays an important role to prevail a society, therefore when a student graduate the whole society is benefited.

Deciding the high hierarchical management position of an HEI has significant implications in the whole academic community. Policymakers, managers, and subordinates of HEIs are disengaged from students when deciding [La, Bajzíková and Dedze 2017]. Since managers and directors' decisions impact student success and most of their decisions rely on past experiences or intuition [Stefanova and Kabakchieva 2018], our research goal is to support this complex task through the development and test of a decision-making model that includes the latest and high reliable computational algorithms to this aim.

Machine Learning (ML) has been a growing trend when analysis information. Its popularity is due to its high accuracy reached, lowest processing time and the range of available algorithms that fit either classification or regression problems. Although many related works have been conducted using ML to identify and classify students, they collect information about distance education, and their prediction is reduced to single subjects or courses. Moreover, they lack the implementation and actions taken about their results probably, because they are focussed on students and teachers as their primary stakeholders.

Firstly, we propose a methodology for data-driven academic prognosis that identifies the steps required when aiming taking action over data. Moreover, in order to include the outperformed algorithm in our decision-making model, we compared Support Vector Machine, Decision Trees, Random Forest, Artificial Neural Network and Logistic Regression. These five classifiers are popular in related literature due to its high accuracy reached and efficiency as classification techniques in the educational field.

In addition, we considered HEIs' managers' academic concerns that were identified in our previous work, as long as their information visualization needs, with the aim of providing a decision-making model that support strategic decisions at HEIs, based on predictions made on graduation rates by the machine learning algorithm presented. The insightful information provided by this model leads to various potential actions. For instance is useful to i) increase students retention, ii) prioritize intervention efforts, iii) create strategies to mitigate early failure and strategic plans through the knowledge of the futurity, iv) diminish students dropout, v) increase the HEI quality indicators among others.

This paper is structured as follows: Section 2 presents the methodology overview, the model for decision-making, the algorithm of the classification technique, and the data used in the experiment. Section 3 exhibits the results where we compare the Machine Learning algorithms used, the toolkit developed for decision making and the social implication of graduation rates. In Section 4 we revise related works. Finally, in Section 5 and 6 we conduct discussions, conclusion and future work.

2 Machine Learning Model to predict graduation of Students

2.1 Problem statement

Decisions made at the strategic level of Higher Educational Institutions (HEIs) are affected in face-to-face educational model mainly because of (a) the disengagement of the stakeholders, (b) difficulties in data acquisition, formatting and centralization and (c) the lack of using efficient computational algorithms to support these complex processes[Nieto, García-Díaz, Montenegro and Crespo 2018].

Therefore, managers and HEIs directors struggle during the decision-making process leading to rely their decisions randomly and not based in logical analysis [Saeed and Dixit 2015], or based on intuition or past experiences [Stefanova and Kabakchieva 2018], delays, limited observation of the "whole picture", limited academic impact of the decision, among others.

2.2 Methodology overview

The model proposed in this paper is focused on support academic decision-making through the prognosis of the number of graduate students, which will enable high level authorities to take action over data through the obtention of a greater understanding when making a strategic decision regarding to dropout, students' retention, resource planning, curriculum design and teacher's management. Moreover, this approach reduces processing time, increases reliability on predictions and promotes hypothesis formulation and hidden patterns understand.

The overall methodology showed in Figure 1. used in this study is built upon the standard process cycle of data mining and data-driven approach devising a set of steps along a workflow.

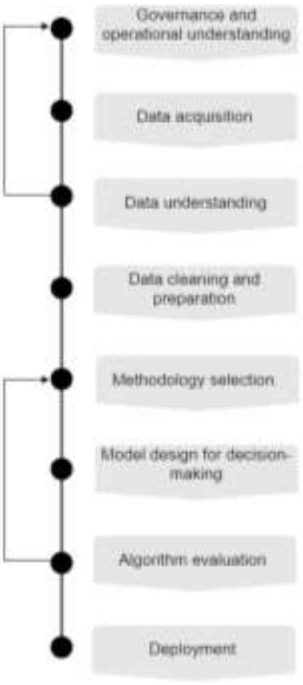

Governance and operational understanding
Data acquisition
Data understanding
Data cleaning and preparation
Methodology selection
Model design for decision-making
Algorithm evaluation
Deployment

Figure 1: Methodology for data-driven academic prognosis

(1) *Governance and operational understanding:* The objectives of the academic prediction should be set in order to identify the stakeholder needs and the possible algorithms to execute. The environment where the HEIs is developed should be evaluated to understand how extern policies affect their decisions. Their current academic situation along with their strategies, processes, people in charge, times and sources should be examined.

(2) *Data acquisition*: Data sources supply the information needed in any analytical purpose. Several HEIs with a face-to-face educational model finds the data accessing

a handicap for decision-making. Data should be harvest, integrated and storage in a coherent database, although it might be in different formats or store in various silos.

(3) *Data understanding:* According to the HEI' educational model and processes operation, data is organized, formatted and stored in different manners. Data should be review and describe within a context. The prognosis relies on the data quantity and quality.

(4) *Data cleaning and preparation:* Data should be selected, cleaned and formatted towards the feature selection. Important characteristics are extracted from the data. The infrastructure that will store and manage the data should be set.

(5) *Methodology selection*: The Machine Learning algorithm is selected. This step is the core functionality that learns from past data and generalizes into the future [Stecto, Dinmohammadi, Zhao and Robu 2019]. Following the main goal of the prediction, if the output variable is categorical, a classification algorithm must be selected. Otherwise, a regression algorithm should be used.

(6) *Model design for decision-making:* Apply learning algorithms to the model data. Depend on the algorithm selected in step above, extra steps as data normalization should be carried out. In this step the algorithm is trained on selected data, tested and validated.

(7) *Algorithm Evaluation*: Is highly recommended to perform algorithms comparison in order to choose the most suitable according to the objective goals. The validity of the model can be estimated through different metrics such as ROC, accuracy, F-measure, recall, and precision.

(8) *Deployment:* Human-Computer Interaction techniques should be considered in order to present the information to managers and directors in a friendly and easily understandable way. Prediction reports should be available on an online toolkit as defined in users' profile. The main results from this step are reliable, supported academic decisions, inducement to hypothesis formulations and a greater insight into the HEI situation.

2.3 Classification technique

Support Vector Machine is a widely used classification technique [Liu, Wang, Wang, Lv and Konan 2017] that, given a set of objects belonging to one of two categories, constructs a hyperplane in a high dimensional space that separates those categories [Miguéis, Freitas, Garcia and Silva 2018]. For nonlinear problems, by using Kernel functions, also known as the Kernel trick, the data is embedded to a higher dimensional space, where it becomes linearly separable, which makes SVM more potent due to is not restricted to linear decision surfaces [Wittek 2014].

Traditional kernel functions include linear, polynomial, Gauss, Sigmoid and Fourier series. The choice of kernel abruptly alters the nature of the decision boundary.

Notably, Gauss Kernel (radial formulation) is stated given its simplicity, high efficiency, easy access and less computation [Wang, Huang and Cheng 2016] [Huang, Maier, Hornegger and Suykens 2017]. Therefore, such a kernel is conveniently selected.

Radial Basis Function (RBF) is a commonly used kernelized learning algorithm in SVM. It has not explicitly defined embedding, operating in an infinite dimensional space. It has the form:

$$\boldsymbol{K}(\boldsymbol{x_m}, \boldsymbol{y_m}) = \boldsymbol{exp}(-\boldsymbol{\gamma}\|\boldsymbol{x_m}, \boldsymbol{y_m}\|^{2}) \quad (1)$$

Where γ>0 is the parameter that controls the radius of the basis function, which serves to control the dispersion of the kernel in the input space. The kernel function $\boldsymbol{K}(\boldsymbol{x_m}, \boldsymbol{y_m})$ intends to measure the "similarity" between $\boldsymbol{x_m}$ and $\boldsymbol{y_m}$ (the larger, the more similar). In terms of the square Euclidean distance, we assume how close are those points to each other.

The binary classification process (i.e., graduated, and not graduated students) imply to fitting a hyperplane (decision boundary) to only two separable classes. Thus, the hyperplane becomes in a straight line separating two classes [Athani, Kodli, Banavasi and Hiremath 2018]. γ and C parameters play an essential role in this RBF classification vector.

In a preliminary training phase, the parameter γ has to be determined [Fischetti 2016]. If this parameter is too small, the model might be under fitted, by contrast, if it is too large the model it might be overfitting.

In order to avoid any of these two phenomena, we take a portion of the 70% of the dataset to training. If we used the whole data to complete the training, the chance that the parameters capture noisy is enhanced, leading to overfitting. We determine the best values for C and γ using 5-fold cross validation, in each k-fold run we evaluate the performance of the algorithm. We used the best performing parameter for creating the final model and testing.

Moreover, we use the C SVM. This type of Support Vector Machine trades off correct classification of training examples using a parameter C. With the aim of helping to improve the accuracy of the output; this parameter implements a penalty on the misclassification that is performed while separating the classes, working as a regularization parameter. Konstanz Information Miner platform (KNIME) was the tool used to execute the Machine Learning algorithm. The extension to SVM along with the cross-validation, aid to encounter the most suitable values for γ (i.e., γ= 0.1) and C (i.e., C=10).

Support Vector Machines is the classification technique choose due to its superior accuracy, compact and comprehensive resulting [Stoean and Stoean 2014], it provides a better decision boundary in any classification problem [Kaneda, Pei, Zhao and Liu 2014][Wang et al. 2016], its remarkable accuracy obtained using similar academic features[Costa, Fonseca, Santana, de Araújo and Rego 2017] [Ghatasheh 2015] and are less prone to overfitting than other models [Czibula, Gergely and Gaceanu 2014]

[Singh, Taylor, Rahman and Pradha 2018]. For more details of the parameters set, observe the algorithm in the next subsection.

2.4 Model approach for decision-making at HEIs

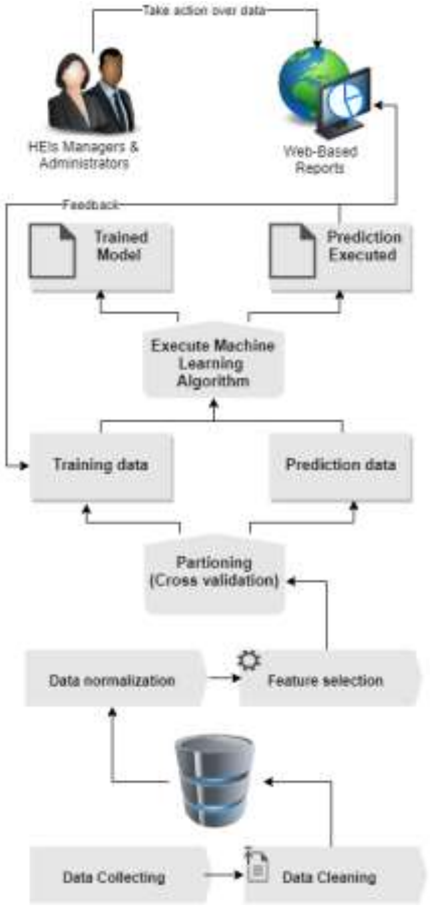


Figure 2: Model proposed for decision-making at HEIs

The model in Figure 2 presents the core goal of this study. The bottom-up approach depicts the integration of human and computer intelligence used within three main phases along with a HEIs' managers-centered objective function.

The first phase requires acquisition and data cleaning. A relatively large labeled data set is needed to create a reliable model [Ihalagedara, Kithuldeniya, Weerasekara and Deegalla 2015]. In academic prognosis, primary information arouses from systems of educational institutions[Tan and Shao 2015]. Educational models as distance and blended rely on more centralized information which differs with the struggles that face-to-face educational models have to face when gathering information.

The inputs listed in Table 1 comes from real data where our model is validated.

ID	**Name**	**Description**	**Measurement**
Inputs provided by the university administration			
1	Residential Stratum	Socio-economic status according to the students' residence address	{1,2,3,4,5,6}
2	Average of the taken subjects	Arithmetic mean of the final grade of all subjects coursed	Grade from 0.0 to 5.0
4	Subjects took in the semester	Subjects enrolled in each semester	Quantity
5	Subjects approved in the semester	Subjects satisfactory approved in each semester	Quantity
6	Subjects failed in the semester	Subjects failed in each semester	Quantity
7	Subjects validated in the semester	Subjects that were failed but after an extra summary exam become approved	Quantity
Inputs calculated from the data given to including in the study			
8	Median grade	Value of the grade separating the higher half from the lower half of the grades	Grade from 0.0 to 5.0
9	Maximum grade	Maximum value from all over final grades	Grade from 0.0 to 5.0
10	Minimum grade	Minimum value from all over final grades	Grade from 0.0 to 5.0
11	The range of the grades	Difference between maximum and the minimum grade	Quantity
12	First quartile of the grades	The middle grade between the smallest grade and the median of the data set	Quantity
13	Second quartile of the grades	Median grade	Quantity
14	Third quartile of the grades	The middle grade between the median and the highest grade of the data set	Quantity
15	The standard deviation of the grades	The square root of the average of the squared deviations of the values subtracted from their average value	Quantity
16	Accumulated subjects took in the whole career	Summary of the subjects enrolled during the whole career	Quantity
17	Accumulated subjects approved in the whole career	Summary of the satisfactory approved subjects during the whole career	Quantity

18	Accumulated subjects failed in the whole career	Summary of the failed approved subjects during the whole career	Quantity
19	Accumulated subjects validated in the whole career	Summary of the subjects that become approved after the validation exam during the whole career	Quantity

Table 1: Description and values of the student features used in the study

In the second phase, we constructed a Machine Learning algorithm to predict the number of students to graduate. Learning in a model translates into fitting a model's parameter to a specific dataset, interactively them updating with several passes through the data [Stecto et al. 2019]. From ML perspective, we use a supervised learning model that rely the prognosis on a dataset of fault events held in the past, from which a learning algorithm finds patterns. It correlates the captured data from the monitored asset to a target variable. [Diez-Olivan, Del Ser, Galar and Sierra 2019].

The main steps carry out to build the model are the following:

(1) Normalize data given them values from [0-1] with the transformation: $p^n = (p - p^{mean})./p^{std}$ where $p^{mean}$is the average of the input vectors in the data set, and $p^{std}$is the vector containing the standard deviations of each element of the input vectors.

(2) Set the accuracy desired (i.e., β>83%)

(3) Set initial settings of the Support Vector Machine (SVM)

(3.1) Set type of SVM, i.e., C-SVC

(3.2) Set type of kernel function, i.e., Radial Basis Function (RBF)

$K(x_m, y_m) = exp(- \gamma \| x_m, y_m \|^2)$ where γ>0

Settings of the Kernel:

(a) Set gamma in kernel function, i.e., γ=1.0

(b) Set C, i.e., C=10.0

(4) Split dataset into two subsets 70% for training and 30% for prediction using a stratified sampling technique to obtain homogenous groups

(5) Split the training subset from step (4) into 5-fold groups to later execute 5-fold cross-validation

(6) Compute the algorithm with initial settings in step (3) Search for $K(x_m, y_m)$ where (x_m, y_m) are Boolean variables according to the original problem formulation, ie., x=graduated, y=not graduated

(7) Store of the accuracy obtained in each k-fold, as well as the values of γ and C, entered in each run

(8) Repeat step (6) to find the best parameters of γ and C

(8.1) Adjust settings in each run until the accuracy is satisfactory for the researches

(9) Compute the algorithm in step (6) to train the whole **training set** using the best values obtained for γ and C

(10) Compute $K(x_m, y_m)$ in step (5) on the **prediction set**

(11) If the accuracy desired is reached approved the last saved solution, if not such a threshold exists, reduce β

Is important to note that our work driver is Managers and Directors of HEIs concerns about graduation rates when making an academic decision [Nieto et al. 2018]. Therefore, instead of determining the exact student performance in a semester, we instead search to segment them, by classifying their academic performance at the time taking to conclude their degree into two groups *(x,y)*, graduated or not graduate students respectively. Processing the students' graduates prediction using ML has managerial implications by enhancing reliability on results and reducing processing time.

We propose in phase three to expose prediction results in a web platform toolkit where HEIs deans and administrative managers can access the information in an easy to understand manner. We believe that the insightful information provided by this model opens an opportunity for different potential actions. For instace to a) create strategic plans through the knowledge of the futurity [Hu, Liu, Chen and Qin 2017] [Martínez et al. 2009], b) prioritize intervention efforts [Aguiar et al. 2015], c) diminish students dropout [Hamoud, Hashim and Awadh 2018], d) increase students retention, e) create strategies to mitigate early failure, f) better manage resources[Medina, García and Olguín 2018], g) increase the HEI quality indicators, among others.

2.5 Data

To illustrate our methodology in a real case study, we perform an experiment using data from an engineering faculty, belonging to a face-to-face educational model in a public University in Colombia. The engineering faculty has approximately 6000 students, 400 professors, whom more than 150 are researches, and offers five undergraduate programs and more than ten graduated programs (specialization, masters, and doctorate) in several fields such as Electrical Engineering and Cadastral Engineering.

The primary data set includes in this study refers to undergraduate students' information for those enrolled between 2004 to 2014, i.e., 12477. After disregarding cases of missing data, our final sample counts 6103 students from whom we have 55220 records. In the data cleaning phase, we consider that engineer programs last ten semesters. Thus, students enrolled after 2009 would not graduate before 2014. Therefore, their data was removed since they are not worthy to train a supervised algorithm. The description of the attributes used in this study is presented in Table 1.

While the model proved to be promising there are limitations that we come through during the data collection phase:

(1) The data provided by the institution used as a case study was very limited because of their data-protection policies. Therefore, relevance attributes such as socio-demographic data, socio-economic status, high school background, and enrollment process information are missing. Students' academic records are evaluated for the most part.

(2) Consequently, the feature selection process is limited by the short number of attributes. Thus, to cope with this limitation, we derivate data that come from the Oracle Database accessed by calculating extra attributes that are shown in the second part of Table 1.

(3) Despite the fact that the data structures were provided, the governance and operational understanding were necessary to process and clean data. For

instance, grades of each subject were provided, but policies to approved or validate a subject were necessary to understand in order to classify students into the objective variable. Moreover, the changes on the syllabus in each career during the observation window represented a challenge to train the data set.

From the limitations stated above, a challenging question arises: Can one predict whether a student graduated or not graduated with these attributes? Classification learning algorithms used in our study address this problem, despite the limitations given by the institutions' data policy. In section 4 we observe the notorious accuracy that these algorithms achieve and how this data highlights the academic situation of a student.

2.6 Evaluation criteria

A 5-fold-cross-validation method was used to obtain error metrics. The data set was split in 5-fold at random, and using 4-folds to train the model, predicting the remaining fifth fold, considered as new data, and calculating the prediction error metrics. The process is repeated five times to predict each fold.

Moreover, as we propose in the methodology above, is highly recommended to perform algorithms evaluation and comparison to choose the one that fits and provides better results according to the objective goals using generalized methods.

Therefore, it is essential to establish the metrics to compare the algorithms. To this aim, in this experiment, we include the measure of F_1-score, due to if provides us the harmonic mean between precision and recall [Costa et al. 2017] described in equation 3 and 4 respectively.

$$F_1 score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{2}$$

$$Precision = \frac{TP}{FP+TP} \tag{3}$$

$$Recall = \frac{TP}{FN+TP} \tag{4}$$

Where TP (True Positives) is the number of positive instances correctly classified as positive. FP (False Positives) is the number of negative instances incorrectly classifies as positive. FN (False Negatives) is the number of positive instances incorrectly classified as negative.

Recall denotes the classifier performance concerning false negatives (the quantity we miss). On the other hand, precision uses the false positive the quantity we caught) to give us information about the classifier performance.

F_1-score also known as F-measure represents in a single score recall and precision. Although the harmonic mean behaves like an average when x is equal to y, when they are different F_1-score is closer to the smaller number as compared to the larger number, giving the algorithm and appropriate score rather than just an arithmetic mean.

3 Results

3.1 Comparing Machine Learning Algorithms

Our primary objective is to design a model for decision making at HEI supported on Machine Learning algorithms. First, to define the algorithm for our model, we conducted in previous works a comparison and analysis of the most used and relevant algorithms for classification in the educational field.

In order to get better predictability, we consider a total of five robust classifiers: Support Vector Machine (SVM), Artificial Neural Network (ANN), Decision Trees (DT), Random Forest (RF) and Logistic Regression (LR). Thus, using the effectiveness metrics defined, we identified and chose the algorithm that best predicts the number of graduated and not graduated students.

However, a previous step before comparing the algorithms is exposed to illustrate and support how we manage the handicaps presented during the experiment.

Since we have a reduced the attributes' quantity given by the case study, and we are addressing a face-to-face educational model where data is more limited in contrast to distance, and blended education were researchers have to clean data by removing, for instance, loggings or time expended in the platform, we tested if data given was enough to represent students' academic situation by running the algorithms mentioned above with a different quantity of data features.

The differences across the scenarios where the number of features (N) vary are shown in Figure 3.

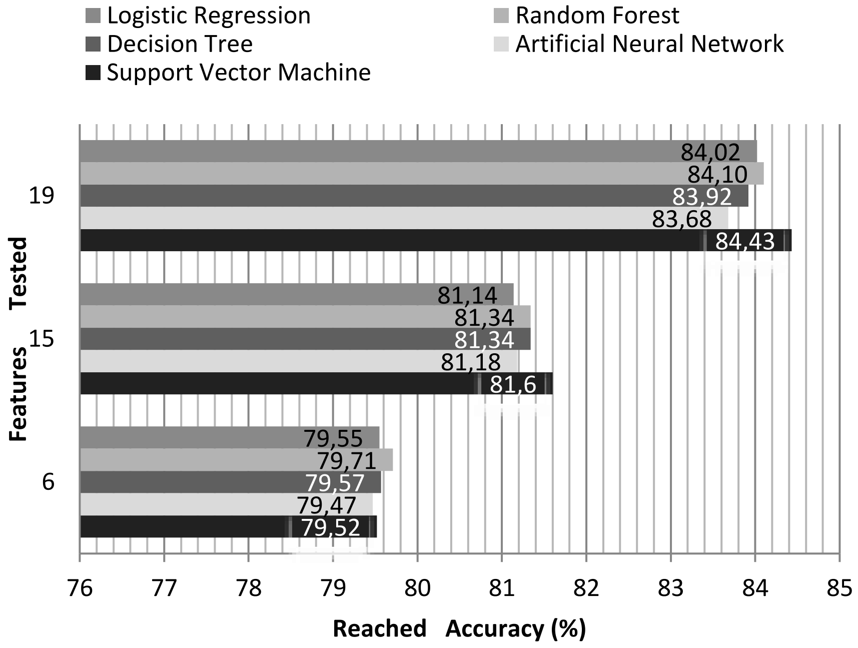


Figure 3: Accuracy reached by the algorithms with different quantity of features

The x-axis in Figure 3 indicates the percentual reached accuracy by the algorithms tested, while the y-axis indicates the three scenarios executed where the number of features varies. From the accuracy results of this experiment, we observe it was relevant to add more features to the dataset by calculating them. The accuracy increases as more features are integrated into the model. The accuracy enhancement is approximately 5% when using 19 features comparing when just 6 features are used, which regarding students correctly classified, it means more than 300 students, becoming a significant finding.

The results listed in Table 2 show the evaluation metrics. Although our evaluation criteria are based on the F-Score, we exposed the results given by metrics such as accuracy, precision, recall, Area Under the Curve (AUC), and Cohen Kappa.

	Support Vector Machine	**Artificial Neural Network**	**Decision Tree**	**Random Forest**	**Logistic Regression**
F-Score	0,8951569	0,8866559	0,889378	0,8910678	0,8894826
Precision	0,8599858	0,8598123	0,866170	0,8644938	0,8704747
Recall	0,9333276	0,9152296	0,913863	0,9193273	0,9093392
Accuracy	0,8454062	0,8345406	0,839249	0,8410600	0,8402148
Error	0,1545937	0,1654593	0,160750	0,1589399	0,1597851
AUC	0,8972	0,8986	0,883	0,8994	0,9028
Cohen Kappa	0,601	0,596	0,596	0,598	0,602

Table 2: Description and values of the student features used in the study

When comparing SVM against ANN, RF, LR, and DT, we found SVM provide slightly better results regarding F-score (values vary between 0.00577 and 0.00408). Every classifier was tested using the same data set and data features using the open analytic platform KNIME. SVM stands out with better most of the metrics. It shows his superior precise reaching the best F- Score (89.51%).

Feasibly, the reason for SVM attains the best performance is because classifies efficiently non-linear separable data when using the appropriate kernel function, is highly tolerant of overfitting and highly accurate (i.e., 89.72%). Similarly occurs with RF which is known to be a better classifier when multiple categorical variables are presented. The combination of these two algorithms could enhance the overall accuracy of the model.

In terms of Area Under the Curve (AUC) that represents the expected performance of the classifier LR achieve the best result (i.e., 0.9028) with 0.0056 points over SVM and Kappa value (i.e., 0.602). By contrast, Artificial Neural Network had the worse overall metrics with the lowest kappa value (i.e.0.596) and accuracy (i.e., 83.45%).

Since both of our binary outcomes (graduated and not graduated) are equally important for our model goal, the F-Score is the suitable metric to seek a balance between Precision and Recall. Thus, and for the outstanding results in the other metrics evaluated, Support Vector Machine is the Machine Learning algorithm used in our proposed model.

3.2 Toolkit for decision-making

The results of the prediction themselves are difficult to access for Higher Educational managers and directors because they would have to get the technical knowledge to search for the outcomes into the programs or platforms where the analysis is held. Besides, the results presented by these platforms in tables and few graphics are not easy to visualize, understand and manage.

Thus, we go beyond the model itself, presenting a web-based toolkit for managers and directors that enables to predict the number of graduated and not graduated students, exposing the results in a friendly and easy to understand manner. The main challenge is to allow the upload of new data and predictions and make it transparent to users. Although the machine learning algorithm is still executed in KNIME, in our prototype we propose a three-layer architecture to integrate the infrastructure as shown in Figure 4.

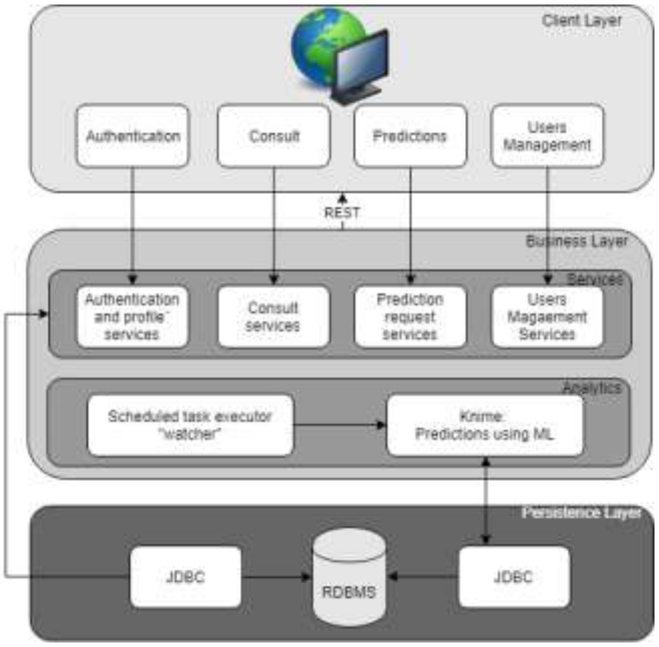


Figure 4: Overall architecture of the prototype

(1) Client Layer*:* After a secure logging, the web toolkit offers four functional modules: *Authentication* (manage profile settings and updating), *Consult* (allow the visualization of the requests done such as tables, filters, and graphics rendered from the business layer), *Predictions* (manage the prediction request and obtain the data) and *Users management* (Allows to create, update and delete users). The modules access to the business layer through REST scheme based on the HTTP protocol, which is part of the

business layer. The development is done in java using the libraries React and Flux to create the user interface. The library d3 is used to visualize graphics,

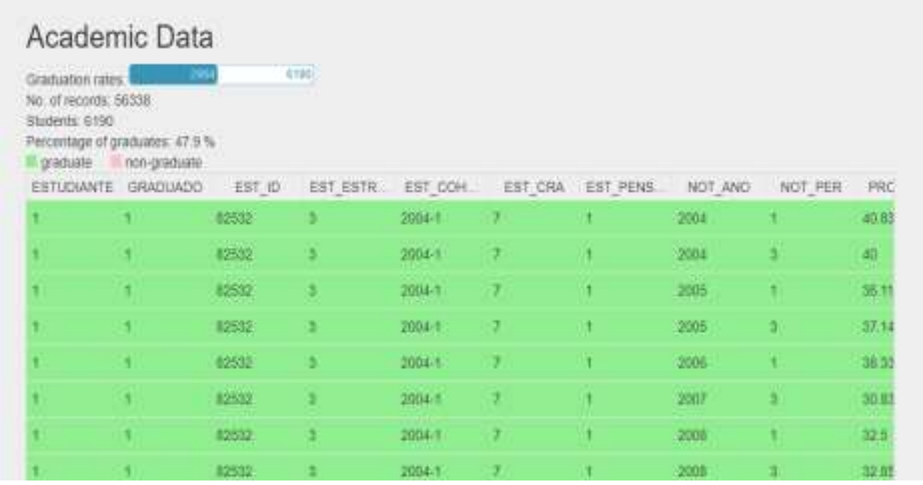
(2) Business Layer: Comprises the business logic and is represented in two sub-layers: the services layer and the analytics layer.

   a. Service sub-layer: As well as the client layer, is divided into modules that expose the four services according to its functionality. The Object Relational Mapping Hibernate applies the persistence interface of Java JPA. Liquibase is used as a migratory system to fill the base information in the persistence layer.

   b. Analytics sub-layer: The server application called “watcher” uses the spring-schedule extension to execute an instance task periodically from Knime. This last contains the workflows to process the training and prediction functionalities. The communication between these two components is done with command lines that call on the Knime instance specifying the correspondent workflow to run.

(3) Persistence layer*:* It uses MySql as the Relational Data Base Management System (RDBMS). The data sources supply the information required in the toolkit. Users can upload new data in CSV format. A template is provided to ensure the data contains all the features required to execute the prediction.

With the aim of providing a friendly web platform to access the predictions done by or model, support the specific needs of the Higher Educational Institution that serves as a case study and, overcome the high costs of Knime server, we developed this prototype. Some of the results are shown in Figure 5 and 6.



Academic Data

Graduation rates: 2964 6190

No. of records: 56338

Students: 6190

Percentage of graduates: 47.9 %

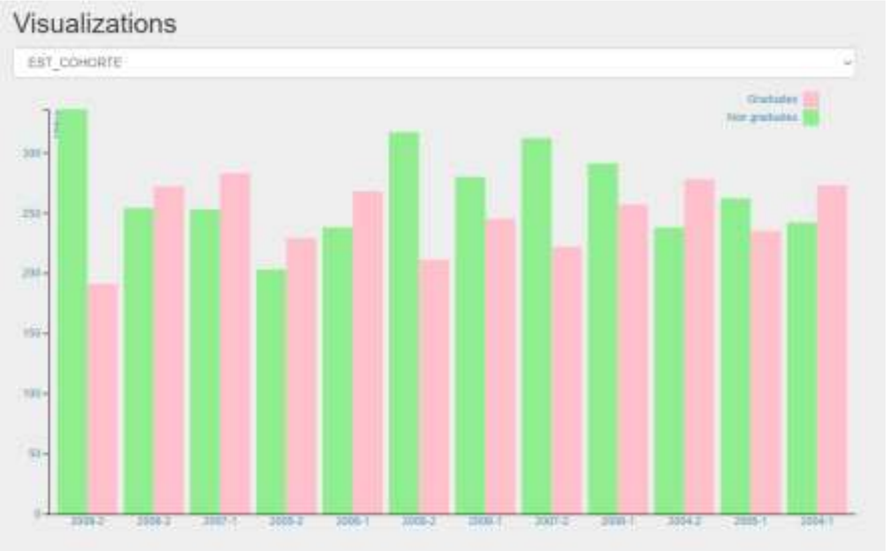
🟩 graduate 🟥 non-graduate

ESTUDIANTE	GRADUADO	EST_ID	EST_ESTR...	EST_COH...	EST_CRA	EST_PENS...	NOT_ANO	NOT_PER	PRC
1	1	82532	3	2004-1	7	1	2004	1	40.83
1	1	82532	3	2004-1	7	1	2004	3	40
1	1	82532	3	2004-1	7	1	2005	1	36.11
1	1	82532	3	2004-1	7	1	2005	3	37.14
1	1	82532	3	2004-1	7	1	2006	1	38.33
1	1	82532	3	2004-1	7	1	2007	3	30.83
1	1	82532	3	2004-1	7	1	2008	1	32.5
1	1	82532	3	2004-1	7	1	2008	3	32.85

Figure 5: Consultation of academic information in the prototype

A consultation on the graduation rate is shown in Figure 5. Green color represents graduated students while red color represents not graduated students. The graduation rate is calculated in a 47.9% with more than 2940 graduated students.

Furthermore, between the various graphics the prototype plots, we consider a potential outcome from Figure 6. That was described by the District University engineer Dean and programs directors as “suitable towards retention decisions” because it is “a simple way to understand information that was unrevealed before.”


Visualizations

EST_COHORTE

Figure 6: Graduated vs. Not graduated students by cohort

The bars in Figure 6 represent the graduate students (green bars) and not graduated students (red bars) in each cohort. The x-axis indicates the cohort, while the y-axis indicates the number of students. Although the green color prevails in most of the cohorts observed, the slight difference with the not graduated students should boost high hierarchical administrator of the HEIs to formulate hypothesis and take action on this concern.

From this point of view, the toolkit prototype shows the results of our proposed model and supports managers and directors given them insightful and highly reliable information. Thus, new opportunities emerge to achieve more favorable decisions.

3.3 Students graduation and social implications

District University Francisco José de Caldas is a public face-to-face Higher Educational Institution. We consider the implications of students’ low rate graduations in our case study due to their remarkable implications in Colombian economy and society in general.

From high school graduate students, just 48% access to professional education and less than 10% have the opportunity to study in a public university. One student cost to the national state 3500 dollars semester, and accordingly to the World Bank and coherently to our study outcomes (i.e., 47.9%), in Colombia just 50% of students that access to professional education finish their studies and graduate [Granja 2017].

From a general society point of view, where all Colombian support with their taxes to the National General Budget, one can believe that the funds invested in Higher Education, half of it is not generating yields.

Education plays a vital role in a country development, and its reputation [Vo and Nguyen 2012], In Colombia low graduation rates affects a) the young people, due to the time expended with not tangible rewards and also have to accept underpaid jobs, b) working people that contribute from their salary to education without experimenting a national enhancement economy, c) industry sector that lack from graduates in some expertise areas, and invest less in scholarships due to the low trust factors, among numerous hidden implication that this phenomenon has in our society.

The cohort graduation rate is a crucial measure of student, and institutional success [Guilbault 2017] and our results expose alarming statistics from District University that are worth to mention.

4 Related work

Decision Support Systems (DSS) as a concept arose in the seventies as the systematization of the decision-making process. Since then, countless applications have been published related to it. Renowned decision analysis methods include influence diagrams, cost/benefits analysis, multi-attribute utility models, analytic hierarchy process, statistics, operation research, among other methods that work with dynamic, uncertain, and multi-criteria aspects of a decision.

In the educational domain, a further evolution link to the computational advances of the 2000s emerged by the name of Educational Data Mining. Recent work on DSS as the Mohamed [Abdahllah 2015] uses operation research to propose a Decision Support model that assist students for long course planning. His mathematical optimization balances students' preferences and advisors' recommendations while maintaining regulation policies.

Livereis [Livereis, Mikropoulos and Pintelas 2016] proposes a DSS that predicts students' performance concerning the final examination in Mathematics in a private secondary school. They utilized Naïve Bayes, Decision Trees, Sequential Minimal Optimization, Back Propagation, 3NN, and Ripper algorithm combine through the voting methodology. A decision support platform is presented as a prototype; however, methodology or a model itself is missing.

Several studies have been published in the literature that use Machine Learning algorithms to help place students within an organization [Thangavel, Bkaratki and Sankar 2017], predict students behavioural intentions , and mainly to identify and segment students that are likely to fail or dropout [Stimpson, Cummings and Member 2014] [Ghatasheh 2015] [Shanthini, Vinodhini and Chandrasekaran 2018] [Sandoval, Gonzalez, Alarcon, Pichara and Montenegro 2018].

The approach proposed at the University of Liege in Belgium [Hoffait and Schyns 2017] used Artificial Neural Network, Decision Trees and Random Forests to identify first-year students profiles with high failure risk. They propose to add an "uncertain class" to increase the accuracy of the predictions. Thus, their prediction results were distant from in just in 4% higher. However, they include many other data, besides academic sources. It is confirmed that some attributes such as socio-demographic data from students represent an enhancement on the prediction. Despite this, their contribution to Higher Educational Institutions are just the prediction numbers itself, due to not any action is proposed from these results. Probably because their focus is on students and their decisions related to this concern are limited.

The work in [Costa et al. 2017]follows a similar machine learning comparison. They compare Support Vector Machine, Decision Tree, Neural Network, Naïve Bayes, to predict students likely to fail in a specific introductory subject. Agreeing with our results, SVM outperforms the other algorithms significantly. However, it does not spcify data feature, and apart from the results, it does not propose any improvement because of the alarming statistics. Besides, it does not help during the decision-making process.

Even though, these studies have shown the high usability of Machine Learning algorithms in the educational field; they lack the implementation and actions taken about their results perhaps, because their stakeholders are mainly students and teachers. Therefore, their results conclude just with the statistics. Moreover, they extract information from the distance education arena, although their hold representative amount of data, their experiments are aimed at small courses or subjects.

Besides, from recent years publications review, and to our knowledge so far, we observed that Machine Learning had not been used to help decision-making process for high hierarchical managers of HEIs or even schools, neither there are focused on fulfilling managers or directors' requirements.

As discussed before, our novel contribution includes a step by step methodology to execute an academic prediction at HEIs. We exposed the features include in our model and specify the mathematical algorithm include in the decision- making model proposed. Additionally, we overcome the limitations of the University that serves a case study and present the results achieved in different scenarios. Finally, to exhibit the decision-making model scope, we developed a prototype to display results to managers and directors that help them to get a whole picture of the students' academic situation regarding graduation rates.

5 Discussions

In this study, we have used Support Vector Machine as the algorithm which our decision model relies on to make academic predictions. Despite its high accuracy and highlighted metrics results, this classifier can be inefficient and slow due to its computational difficulties as well as the model complexity with a more massive scale training set [Liu et al. 2017].

Time complexity in SVM is usually between $O(m^2n)$ and $O(m^3n)$ where m is the number of instances, and n is the number of features [Stecto et al. 2019], but none of the mentioned studies have used a data set more extensive than ours. Although, we believe that if a larger data set is analyzed, time processing can become in a handicap of the model without affecting its high reliability.

In the present study, we have overcome the lack of data attributes from the case study as our information provider. After various experiments, we found the algorithms reached a significant accuracy enhancement when more attributes were added by calculating them from the raw data.

The academic prediction was determined by the authors in previous research as our work driver after a survey conducted to Deans and higher management positions at HEIs where they stated their concern about graduation rates when making an academic decision.

During the experiments, we used the analytic platform that in the beginning suits our needs, but some different analytics platforms are also possible. Alternatively to Knime, platforms such as TensorFlow, RapidMiner, Alteryx among others allows through nodes or libraries access to Machine Learning tools. Since we have to face a high cost for Knime server and also have to fulfill specific needs from the case study HEI, we develop our toolkit that allows to run predictions and visualize results.

Data used in the case study and the workflows of the algorithms executed in Knime are available at https://github.com/vicegd/decision.making.higher.education.

6 Conclusions and Future work

The choice of the Machine learning algorithm depends on the problem to solve. Therefore, the methodology proposed is an essential outcome of the research due to specifies important hints when attempting academic prognosis, starting from governance and operational understanding to the deployment.

Additionally, this paper provides the comparison of five robust Machine Learning algorithms (SVM, ANN, DT, RF, LR) and their ability to recognize a balance segmentation of graduated and not graduated students. It was revealed that Support Vector Machine has the best recognition F- Score (i.e., 89,51%) and Accuracy (i.e., 84,54%) among them.

In the near future, a hybrid model proposal between Support Vector Machine and Random Forest would be worth attempting when seeking graduation rates due to its outperformed at overall prediction classification problem.

One of the main limitations of this study was data acquisition. Thus, it would be interesting to execute the same model proposed with larger data attributes that include socio-economic, and demographic data from students. For instance, students' gender, tuition fee, graduated high school GPA, monthly income, people in charge, and city where they came from, to mention few attributes that would be worth to analyze. When having more input parameters, the use of an optimization parameter such as gradient descendent [Jeon, Park and Lee 2018] might help to optimize our problem.

Performing the model proposed offers reliable predictions' results and insightful information about the academic situation of a HEIs. Additionally, the achieved outcomes along the model' execution, and the positive feedback received from the case study Dean and program directors about the prototype ensures us that is possible to support academic decision-making through our model efficiently. Nevertheless, a web-based toolkit that meets Human-Computer Interaction requirements would represent a significant enhancement of our prototype and the way the model exposes information.

This research represents an innovative model that supports high hierarchical administrators of HEIs during academic decision- making regarding graduation rates. We care about stakeholder that were not addressed whom significantly impact the whole institutional community when deciding. Furthermore, hypothesis formulation, and decision respecting retention rates, students' exclusion policies, students' dropout rates, and strengthen programs, are encouraging to be held from the decision- making model proposed.

References

[Abdahllah 2015] Abdahllah, M.: ‘A decision support model for long-term course planning’; Decision Support Systems, Vol. 74 (2015), pp. 33–45. https://doi.org/http://dx.doi.org/10.1016/j.dss.2015.03.002

[Aguiar, Lakkaraju, Bhanpuri, Miller, Yuhas and Addison 2015] Aguiar, E., Lakkaraju, H., Bhanpuri, N., Miller, D., Yuhas, B., Addison, K. L.: ‘Who, when, and Why: A Machine Learning Approach to Prioritizing Students at Risk of Not Graduating High School on Time’; In Proceedings of the Fifth International Conference on Learning Analytics And Knowledge - LAK ’15. ACM (2015), pp. 93–102. https://doi.org/10.1145/2723576.2723619

[Athani, Kodli, Banavasi and Hiremath 2018] Athani, S. S., Kodli, S. A., Banavasi, M. N., Hiremath, P. G. S.: ‘Student performance predictor using multiclass support vector classification algorithm’; In IEEE International Conference on Signal Processing and Communication, ICSPC 2017 (2018), pp. 341–346. https://doi.org/10.1109/CSPC.2017.8305866

[Costa, Fonseca, Santana, de Araújo and Rego 2017] Costa, E. B., Fonseca, B., Santana, M. A., de Araújo, F. F., Rego, J.: ‘Evaluating the effectiveness of educational data mining techniques for early prediction of students’ academic failure in introductory programming courses’; Computers in Human Behavior, Vol. 73 (2017), pp. 247–256. https://doi.org/10.1016/j.chb.2017.01.047

[Czibula, Gergely and Gaceanu 2014] Czibula, G., Gergely, I., Gaceanu, R.: ‘A support vector machine model for intelligent selection fo data representations’; Applied Soft Computing, Vol. 18 (2014), pp. 70–81. https://doi.org/http://dx.dor.org/10.1016/j.asoc.2014.01.026

[Diez-Olivan, Del Ser, Galar and Sierra 2019] Diez-Olivan, A., Del Ser, J., Galar, D., Sierra, B.: ‘Data fusion and machine learning for industrial prognosis: Trends and perspectives towards Industry 4.0’; Information Fusion, Vol. 50 (2019), pp. 92–111. https://doi.org/https://doi.org/10.1016/j.inffus.2018.10.005

[Fischetti 2016] Fischetti, M.: ‘Fast training of Support Vector Machines with Gaussian kernel’; Discrete Optimization, Vol. 22 (2016), pp. 183–194. https://doi.org/10.1016/j.disopt.2015.03.002

[Ghatasheh 2015] Ghatasheh, N.: ‘Knowledge Level Assessment in e-Learning Systems Using Machine Learning and User Activity Analysis’;, Vol. 6, No. 4 (2015).

[Granja 2017] Granja, S.: ‘Colombia mejora en acceso a la educación superior pero falta calidad’; El Tiempo. Bogotá (2017, June 5), pp. 1–5. Retrieved from https://www.eltiempo.com/vida/educacion/acceso-y-calidad-de-educacion-superior-en-colombia-segun-el-banco-mundial-95456

[Guilbault 2017] Guilbault, M.: ‘Students as customers in higher education : The (controversial) debate needs to end’; Journal of Retailing and Consumer Services (2017), pp. 8–11. https://doi.org/http://dx.doi.org/10.1016/j.retconser.2017.03.006

[Hamoud, Hashim and Awadh 2018] Hamoud, A. K., Hashim, A. S., Awadh, W. A.: ‘Predicting Student Performance in Higher Education Institutions Using Decision Tree Analysis’; International Journal of Interactive Multimedia and Artificial Intelligence, Vol. 5, No. 2 (2018), p. 26. https://doi.org/10.9781/ijimai.2018.02.004

[Hoffait and Schyns 2017] Hoffait, A., Schyns, M.: 'Early detection of university students with potential difficulties'; Decision Support Systems (2017). https://doi.org/http://dx.doi.org/10.1016/j.dss.2017.05.003

[Hu, Liu, Chen and Qin 2017] Hu, J., Liu, H., Chen, Y., Qin, J.: 'Strategic planning and the stratification of Chinese higher education institutions'; International Journal of Educational Development, , No. 2016 (2017), pp. 1–8. https://doi.org/10.1016/j.ijedudev.2017.03.003

[Huang, Maier, Hornegger and Suykens 2017] Huang, X., Maier, A., Hornegger, J., Suykens, J.: 'Indefinite kernels in least squares support vector machines and principal component analysis'; Applied and Computational Harmonic Analysis, Vol. 43 (2017), pp. 162–172. https://doi.org/http://dx.doi.org/10.1016/j.acha.2016.09.001

[Ihalagedara, Kithuldeniya, Weerasekara and Deegalla 2015] Ihalagedara, K., Kithuldeniya, R., Weerasekara, S., Deegalla, S.: 'Feasibility of Using Machine Learning to Access Control in Squid Proxy Server'; (2015), pp. 491–494.

[Jeon, Park and Lee 2018] Jeon, Y., Park, Y., Lee, S.: 'Machine Learning Optimization of Parameters for Noise Estimation'; Journal of Universal Computer Science, Vol. 24, No. 9 (2018), pp. 1271–1281.

[Kaneda, Pei, Zhao and Liu 2014] Kaneda, Y., Pei, Y., Zhao, Q., Liu, Y.: 'Study on the effect of learning parameters on decision boundary making algorithm'; In EEE International Conference on Systems, Man and Cybernetics (2014), pp. 705–710. https://doi.org/10.1109/SMC.2014.6973992

[La, Bajzíková and Dedze 2017] La, A., Bajzíková, Ľ., Dedze, I.: 'Barriers and drivers of innovation in higher education : Case study-based evidence across ten European universities'; International Journal of Educational Development, Vol. 55, No. May (2017), pp. 69–79. https://doi.org/10.1016/j.ijedudev.2017.06.002

[Liu, Wang, Wang, Lv and Konan 2017] Liu, C., Wang, W., Wang, M., Lv, F., Konan, M.: 'An efficient instance selection algorithm to reconstruct training set for support vector machine'; Knowledge-Based Systems, Vol. 116 (2017), pp. 58–73. https://doi.org/10.1016/j.knosys.2016.10.031

[Livereis, Mikropoulos and Pintelas 2016] Livereis, I., Mikropoulos, T., Pintelas, P.: 'A Decision Support System for Predicting Student Performance'; Themes in Science and Technology Education, Vol. 9, No. 1 (2016), pp. 43–57. https://doi.org/10.15680/IJIRCCE.2014.0212015

[Martínez, Franco, Rodriguez, Crespo, G-Bustelo and Baena 2009] Martínez, O. S., Franco, E. T., Rodriguez, H. C., Crespo, R. G., G-Bustelo, B. C. P., Baena, L. R.: 'Viabilidad de la aplicación de Sistemas de Recomendación a entornos de e-learning'; In V Simposio Pluridisciplinar sobre Diseño y Evaluación de Contenidos (2009). Retrieved from http://www.web.upsa.es/spdece08/contribuciones/157_SPDECE.pdf

[Medina, García and Olguín 2018] Medina, M., García, C., Olguín, M.: 'Planning and Allocation of Digital Learning Objects with Augmented Reality to Higher Education Students According to the VARK Model'; International Journal of Interactive Multimedia and Artificial Intelligence, Vol. 5, No. 2 (2018), p. 53. https://doi.org/10.9781/ijimai.2018.02.005

[Miguéis, Freitas, Garcia and Silva 2018] Miguéis, V. ., Freitas, A., Garcia, P. J. V, Silva, A.: 'Early segmentation of students according to their academic performance: A predictive modelling approach'; Decision Support Systems, Vol. 115 (2018), pp. 36–51. https://doi.org/https://doi.org.10.1016/j.dss.2018.09.001

[Nieto, García-Díaz, Montenegro and Crespo 2018] Nieto, Y., García-Díaz, V., Montenegro, C., Crespo, R. G.: ‘Supporting academic decision making at higher educational institutions using machine learning-based algorithms’; Soft Computing (2018), pp. 1–9. https://doi.org/10.1007/s00500-018-3064-6

[Saeed and Dixit 2015] Saeed, F., Dixit, A.: ‘A decision support system approach for accreditation & quality assurance council at higher education institutions in Yemen’; In IEEE 3rd International Conference on MOOCs, Innovation and Technology in Education (MITE). IEEE (2015), pp. 163–168. https://doi.org/10.1109/MITE.2015.7375308

[Sandoval, Gonzalez, Alarcon, Pichara and Montenegro 2018] Sandoval, A., Gonzalez, C., Alarcon, R., Pichara, K., Montenegro, M.: ‘Centralized student performance prediction in large courses based on low-cost variables in an institutional context’; The Internet in Higher Education, Vol. 37 (2018), pp. 76–89. https://doi.org/https://doi.org/10.1016/j.iheduc.2018.02.002

[Shanthini, Vinodhini and Chandrasekaran 2018] Shanthini, A., Vinodhini, G., Chandrasekaran, R. .: ‘Predicting students’ academic preformance in the university using meta decision tree classifiers’; Journal of Computer Science, Vol. 14, No. 5 (2018), pp. 654–662. https://doi.org/https://doi.org/10.3844/jcssp.2018.654.662

[Singh, Taylor, Rahman and Pradha 2018] Singh, S., Taylor, R., Rahman, M., Pradha, B.: ‘Developing robust arsenic awareness prediciton models using machine learning algorithms’; Journal of Enviromental Management, Vol. 211 (2018), pp. 125–137. https://doi.org/10.1016/j.jenvman.2018.01.044

[Stecto, Dinmohammadi, Zhao and Robu 2019] Stecto, A., Dinmohammadi, F., Zhao, X., Robu, V.: ‘Machine learning methods for wind turbine condition monitoring: A review’; Renewable Energy, Vol. 133 (2019), pp. 620–635. https://doi.org/https://doi.org/10.1016/j.renene.2018.10.047

[Stefanova and Kabakchieva 2018] Stefanova, K., Kabakchieva, D.: ‘Educational data mining perspectives within university big data environment’; In 2017 International Conference on Engineering, Technology and Innovation: Engineering, Technology and Innovation Management Beyond 2020: New Challenges, New Approaches, ICE/ITMC 2017. IEEE (2018), pp. 264–270. https://doi.org/10.1109/ICE.2017.8279898

[Stimpson, Cummings and Member 2014] Stimpson, A. J., Cummings, M. L., Member, S.: ‘Assessing Intervention Timing in Computer - Based Education Using Machine Learning Algorithms’; IEEE Access, Vol. 2 (2014), pp. 78–87. https://doi.org/10.1109/ACCESS.2014.2303071

[Stoean and Stoean 2014] Stoean, C., Stoean, R.: ‘Post-evolution of variable-length class prototypes to unlock decision majing within support vector machines’; Applied Soft Computing, , No. 25 (2014), pp. 159–173. https://doi.org/http://dx.doi.org/10.1016/j.asoc.2014.09.017

[Tan and Shao 2015] Tan, M., Shao, P.: ‘Prediction of student dropout in E-learning program through the use of machine learning method’; International Journal of Emerging Technologies in Learning, Vol. 10, No. 1 (2015), pp. 11–17. https://doi.org/10.3991/ijet.v10i1.4189

[Thangavel, Bkaratki and Sankar 2017] Thangavel, S. K., Bkaratki, P. D., Sankar, A.: ‘Student placement analyzer: A recommendation system using machine learning’; In 2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS). Colmbatore (2017), pp. 1–5. https://doi.org/10.1109/ICACCS.2017.8014632

[Vo and Nguyen 2012] Vo, T. N. C., Nguyen, H. P.: ‘A Knowledge-Driven Educational

Decision Support System'; In 2012 IEEE RIVF International Conference on Computing & Communication Technologies, Research, Innovation, and Vision for the Future. IEEE (2012), pp. 1–6. https://doi.org/10.1109/rivf.2012.6169819

[Wang, Huang and Cheng 2016] Wang, X., Huang, F., Cheng, Y.: 'Computational performance optimization of support vector machine based on support vectors'; Neurocomputing, Vol. 211 (2016), pp. 66–71. https://doi.org/10.1016/j.neucom.2016.04.059

[Wittek 2014] Wittek, P.: 'Quantum Machine Learning'; Quantum Machine Learning (1st ed.). Elsevier (2014). https://doi.org/https://doi.org/10.1016/C2013-0-19170-2