Machine learning classification analysis for a hypertensive population as a function of several risk factors.

Fernando López Martínez; Aron Schwarcz; Edward Rolando Núñez Valdez; Vicente García Díaz. Machine learning classification analysis for a hypertensive population as a function of several risk factors. Expert Systems with Applications. 110, pp. 206 - 215. Elsevier, 2018.

DOI: http://dx.doi.org/10.1016/j.eswa.2018.06.006

Machine Learning Classification Analysis for a Hypertensive Population as a Function of Several Risk Factors

Fernando López-Martínez[a,1,*], Aron Schwarcz.MD[a,2,**], Edward Rolando Núñez-Valdez[b,3,**], Vicente García-Díaz[b,4,**]

[a] *350 Engle Street,Englewood Hospital, Englewood,NJ 07631, USA*
[b] *C/ Federico Garca Lorca, University of Oviedo, 33007, Oviedo, Spain*

Abstract

This research presents a prediction model to evaluate the association between gender, race, BMI, age, smoking, kidney disease and diabetes using logistic regression. Data were collected from NHANES datasets from 2007 to 2016. An unbalanced sampling dataset of 19.709 with (83%) non-hypertensive individuals and (17%) hypertensive individuals. Some risk factors were categorized, and indicator variables were created to transform the continuous variables to a binary form to have consistent predictors with the outcome. The results show a sensitivity of 77%, a specificity of 68%, precision on the positive predicted value of 32% in the test sample and a calculated AUC of 73%. The model also confirms that individuals with obesity, age range between 71 and 80 years old, race non-Hispanic black and male have higher odds of having hypertension. Diabetes, kidney disease and smoking habits do not affect odds of the outcome.

Keywords: `Logistic Regression`, Hypertension, BMI, Diabetes, Blood Pressure, Cardiovascular disease, NHANES

[*] Corresponding author
[**] Co-Authors
Email addresses: `uo259897@uniovi.es` (Fernando López-Martínez), `aron.schwarcz@ehmchealth.org` (Aron Schwarcz.MD), `nunezedward@uniovi.es` (Edward Rolando Núñez-Valdez), `garciavicente@uniovi.es` (Vicente García-Díaz)
[1] Student. Department of Computer Science, Oviedo University. Phone: +1 5515870112
[2] Board Certified in Cardiovascular Disease and Nuclear Cardiology. Phone: +1 9175969186
[3] Lecturer. Department of Computer Science, Oviedo University. +34 671114773
[4] Lecturer. Department of Computer Science, Oviedo University. +34 985103326

1. Introduction

Machine Learning methods have shown the potential to improve outcomes in many domains of research. Health monitoring, patterns recognition, text analytics, knowledge discovery and web semantic analysis are just some domains of action, and Sophisticated and powerful Machine Learning algorithms are already well understood and accessible (Hijazi et al., 2016). According to the definition of Machine Learning “The field of study that gives computers (or machines) that ability to learn without being explicitly programmed” (Samuel, 1959), transforming complex tasks or problems in easily resolvable algorithms, allow us to implement solutions for a variety of situations were learning from data is the primary goal. Supervised machine learning algorithms have been used in traditional disease risk models (Chen et al., 2017) for patients with specific conditions and the use of large amounts of data available allows to improve the accuracy of the classification models(Frunza et al., 2011). The use of Logistic regression has increased rapidly nowadays in biometrics, and clinical research to classify data efficiently (Song et al., 2017) and Hypertension condition is one of the cardiovascular diseases where logistic regression with a binary outcome can be applied.

World Health Organization estimates that 40 million deaths occurred due to noncommunicable diseases in the world in 2015, the majority of those deaths were caused by the cardiovascular disease with 17.7 million (World Health Organization, 2017). In the United States, cardiovascular disease is the leading cause of death despite the existence of effective and inexpensive treatments (National Center for Health Statistics, 2017). In 2015 the number of deaths in adults due to conditions of heart based on death certificates was 63,3164 deaths in adults of 20 years old and over. Hypertension or high blood pressure is one of the most critical risk factors for cardiovascular disease among U.S. adults (Nwankwo et al., 2013) and, it has significant public and economic implications. The health care expenditure associated with high blood pressure in 2011 in the US were $46 billion, and the projected total cost for 2030 is $274 billion (Mozaffarian & Benjamin, 2015).

The prevalence of hypertension in the U.S. adults population is high and increasing in recent years as can be seen in the National Health and Nutrition Examination Survey (NHANES) conducted by the National Center for Health Statistics. This survey has been the main source of tracking the burden of hypertension in the U.S. population (Committee on Public Health Priorities to Reduce and Control Hypertension in the U.S. Population., 2010). This increment in hypertensive population probably shows us that the current approaches to predict cardiovascular disease fail to identify people who would benefit from preventive hypertensive treatment. The use of machine learning tools in medicine has proven significant outcomes in clinical decision support and patient care, specifically in hypertension diagnosis (Kublanov et al., 2017). For this reason, this research in collaboration with clinical experts will present the use of several factors to create an accessible and interpretable logistic regression prediction model to classify hypertensive patients and study the relevance of each variable in the presence of the others using national health data from the National Health and Nutrition Examination Survey NHANES. The risk variables were selected after reviewing a compressive review of studies describing equations to predict hypertension (Echouffo-Tcheugui et al., 2013).

The structure of this paper is as follows: Section 2 presents related work and literature research of several risk models that used machine learning. In section 3, we describe the development of the machine learning model, describing the studied population and the data source, and validation of the model. Section 4 presents the statistical and clinical analysis of the outcomes. Section 5 discusses the results of the model and its limitations. Finally, section 6 covers conclusions and future work.

2. Related work

We conducted a literature research to identify several risk models that have used machine learning techniques to predict the occurrence of hypertension among different populations. Muhamad Amir (Ahmad et al., 2014) shows that age, BMI, and systolic blood pressure were significantly related to hypertension and the AUC calculated was 71.8%. Another study, performed by the department of community medicine in Kathmandu medical college (Manandhar, 2016), found a significant association between the variables age, smoking, religion, occupation, alcohol consumption and family history of hypertension while sex, education status and exercise were found to be insignificant. In China, a study was developed and found that age, Fasting Plasma Glucose (FPG), the number of people for dinner, alcohol consumption, illiterate and smoking are risk factors for hypertension (Zheng et al., 2013). Other studies in the University of medical sciences in Tehran, Iran examined distinct patterns of metabolic risk factors and their association with cardiovascular disease risk (Ramezankhani et al., 2017). This study found BMI, Fasting Plasma Glucose (FPG), high total cholesterol (TC) and low glomerular filtration rate (GFR) to have a significant association with hypertension. Another study performed in Nigeria in the department of mathematical and physical sciences (Olubiyi, 2013) shows that just age had a meaningful relationship with hypertension. A Swedish risk model (Fava et al., 2013) used logistic regression to study anthropometric, anamnestic and metabolic features and their association with the prevalence of hypertension. Age, sex, BMI, and heart rate were highly significantly associated with hypertension, other variables, such as glycolipids parameters and other anamnestic elements were included in the model, and the association was somewhat attenuated but remained significant. The area under the curve for this study was 66%. A hypertension risk score was developed in the department of medicine, UNC Kidney Center and Division of Nephrology and Hypertension, School of Medicine in the U.S. (Kshirsagar et al., 2010) to determine which demographic and medical variables predict and evaluate the development of hypertension.

Age, smoking, family history of hypertension, the presence of diabetes and BMI were associated with developing hypertension. Female sex, family history of diabetes, and exercise did not reach a general significant level of p$\leq$0.05 with an AUC of 75% to 78% in the testing sample. Machine learning has been widely used in recent years with medical data for medical discovery and detection of patterns in complex diseases (Seffens et al., 2015). These studies show that by using supervised machine learning algorithms, we can build a classification model to perform risk assessments for hypertension with acceptable levels of accuracy. Our model uses age, gender, ethnicity, BMI, smoking history, kidney disease and diabetes as categorical variables with values at the same range to reduce the scale difference between them while all other models used a combination of categorical and continuous variables. Neither one of the studied models used the groups of independents variables used in this research and investigated their significance with Hypertension without including systolic blood pressure and diastolic blood pressure. In the final analysis, we can observe a clear relationship between these non invasive independent variables and their significance with the outcome.

3. Model Development and Validation

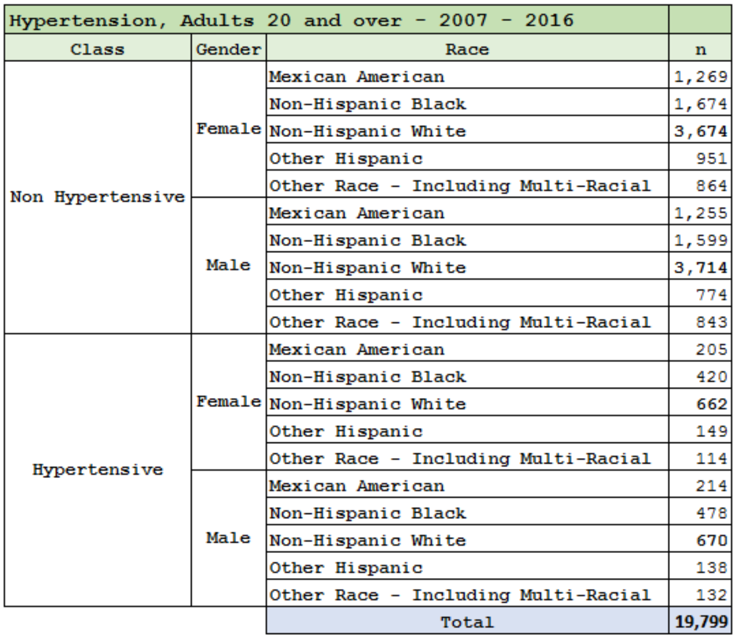
We have applied the logistic regression classification model in this paper, and we will discuss and analyze how the model has been implemented and evaluated.

3.1. Data Source

We utilized The National Health and Nutrition Examination Survey (NHANES) datasets from NHANES 2007-2008 to NHANES 2015-2016. Demographic variables, examination data like blood pressure and body measures and questionnaire data including questions related to diabetes, smoking cigarette use and kidney conditions. NHANES data was released as SAS transport files and imported for the study by using xport 2.0 Python reader.

3.2. Study Population and Analysis

Data of the National Health and Nutrition Examination Survey (NHANES), collected during 2007 - 2016, were used to train and test the prediction model. This model was created to evaluate the importance of the risk factor variables and their relationship with the prevalence of hypertension among a nationally representative sample of adults $\geq$ 20 years in the United States (n=19,799). Table 1. shows the distribution of the samples by hypertensive patients, gender and race after the data cleaning process where only records with non-null values where used.



Hypertension, Adults 20 and over - 2007 - 2016			
Class	Gender	Race	n
Non Hypertensive	Female	Mexican American	1,269
		Non-Hispanic Black	1,674
		Non-Hispanic White	3,674
		Other Hispanic	951
		Other Race - Including Multi-Racial	864
	Male	Mexican American	1,255
		Non-Hispanic Black	1,599
		Non-Hispanic White	3,714
		Other Hispanic	774
		Other Race - Including Multi-Racial	843
Hypertensive	Female	Mexican American	205
		Non-Hispanic Black	420
		Non-Hispanic White	662
		Other Hispanic	149
		Other Race - Including Multi-Racial	114
	Male	Mexican American	214
		Non-Hispanic Black	478
		Non-Hispanic White	670
		Other Hispanic	138
		Other Race - Including Multi-Racial	132
		Total	**19,799**

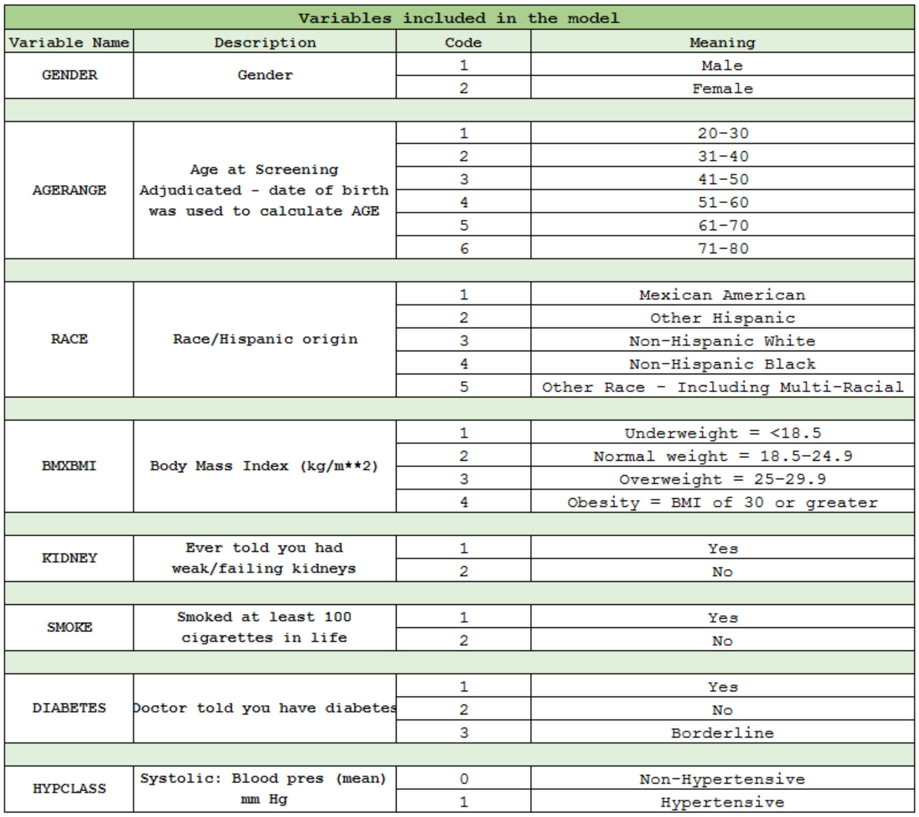
Table 1: Number of samples by Hypertensive class, gender and race.

3.3. Risk Factor variables

Based on the Healthy People initiative of the Office of Disease Prevention and Health Promotion to improve cardiovascular health and reduction in deaths from cardiovascular disease, the leading controllable risk factors for heart disease are high blood pressure, high cholesterol, cigarette smoking, diabetes, unhealthy diet and physical inactivity, and overweight and obesity (Mozaffarian & Benjamin, 2016). For this study, blood pressure was used to create the dichotomous dependent variable for the model; where hypertension is defined as a mean systolic blood pressure of 140 mmHg. The independent variables or features selected were Age, Race, Body Mass Index (BMI), cigarette smoking, kidney disease presence and diabetes. Kidney disease was not present in the Healthy People Initiative for heart disease and stroke, but the kidney plays a key role in keeping the blood pressure in a normal range (Tedla et al., 2011). Awareness of chronic kidney disease (CKD) was defined as "yes" response to the question "Have you ever told by a health care provider you have weak or failing kidneys?" during the interview, and for NHANES 2015-2016, CKD was defined as a glomerular filtration rate (GFR) $\leq$ 60 ml/min/1.73 m^2 (Miller, 2009) and albumin-creatinine ratio $\geq$ 30 mg/g (Center for Disease Control and Prevention, 2015). Participant were considered to have diabetes if they answered "Yes" or "Borderline" to the question "Doctor told you have diabetes?" (U.S. Department of Health and Human, 2005). Cigarette smokers were defined as adults who reported having smoked $\geq$ 100 cigarettes during their lifetime and who currently smoke every day or some days (Centers for Disease Control, 2016). Age and BMI were categorized and transformed from continues variables to Categorical variables to help understand the real relationship between the variables and to reduce the scale difference between them. Table 2. shows all the independent variables selected.

3.4. Machine Learning Model

Machine learning models have contributed to a lot of improvements in epidemiological studies and patient care (Nilashi et al., 2017). Logistic regression classification analysis was used in this paper to investigate the relationship between the independent variables described above and the dependent variable hypertension. The implementation of this risk prediction models will be valuable for the identification of individuals at risk of developing hypertension based on the risk factors presented in the Healthy People initiative and the association



Variables included in the model			
Variable Name	Description	Code	Meaning
GENDER	Gender	1	Male
		2	Female
AGERANGE	Age at Screening Adjudicated - date of birth was used to calculate AGE	1	20-30
		2	31-40
		3	41-50
		4	51-60
		5	61-70
		6	71-80
RACE	Race/Hispanic origin	1	Mexican American
		2	Other Hispanic
		3	Non-Hispanic White
		4	Non-Hispanic Black
		5	Other Race - Including Multi-Racial
BMXBMI	Body Mass Index (kg/m**2)	1	Underweight = <18.5
		2	Normal weight = 18.5-24.9
		3	Overweight = 25-29.9
		4	Obesity = BMI of 30 or greater
KIDNEY	Ever told you had weak/failing kidneys	1	Yes
		2	No
SMOKE	Smoked at least 100 cigarettes in life	1	Yes
		2	No
DIABETES	Doctor told you have diabetes	1	Yes
		2	No
		3	Borderline
HYPCLASS	Systolic: Blood pres (mean) mm Hg	0	Non-Hypertensive
		1	Hypertensive

Table 2: Risk factor variables and the dependent variable.

and significance between the variables.

Logistic Regression is used to assess the probability of having hypertension or not as a function of risk factors. The most useful way to model this relationship is expressed by the logit function. The logit function is also called the link function because it links the value of the independent variables to the probability of occurrence of the event defined by the dependent variable (Sainani, 2014).

$$logit(p) = ln(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_i X_i$$

we can express the chance of success as:

$$p = \frac{e^{(\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_i X_i)}}{1+e^{(\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_i X_i)}}$$

Where p is the predicted probability of having hypertension, X_i are the risk factors or independent variables, β_i are the coefficients that will be estimated by using the method of maximum likelihood and will allow us to calculate the odds that for every unit increase in X_i, the odds of having hypertension changes by e^{β}. The estimation of the parameters intercept and coefficients will be obtained using the logistic regression model in the package sklearn of Python programing language(Pedregosa et al., 2012). Once the coefficients have been estimated, we can calculate the probability of an individual having hypertension as:

$$p = \frac{e^{(\beta_0+\beta_1 gender+\beta_2 age+\beta_3 race+\beta_4 bmi+\beta_5 kidney+\beta_6 smoke+\beta_7 diabetes)}}{1+e^{(\beta_0+\beta_1 gender+\beta_2 age+\beta_3 race+\beta_4 bmi+\beta_5 kidney+\beta_6 smoke+\beta_7 diabetes)}}$$

3.5. Logistic Regression Classifier

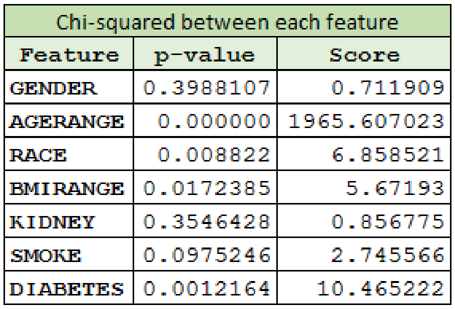
As explained in the previous section the probabilities describing the possible outcomes are modeled using a logistic function. We have used the implementation of logistic regression in scikit-learn from Python to fit a binary, multivariable logistic regression that uses the LIBLINEAR classification algorithm (Fan et al., 2008) and predicts the positive class. The cost function that is minimized by the L2 penalized logistic regression problem and implements a trust region Newton method (Lin et al., 2008) is:

$$\min_{w,c} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \log(\exp(-y_i(X_i^T w + b)) + 1)$$

Where C is the regularization strength, smaller values specify stronger regularization. These hyper-parameters where determined by using a grid search and two-fold cross-validation on the sampling data set.

3.6. Variables selection

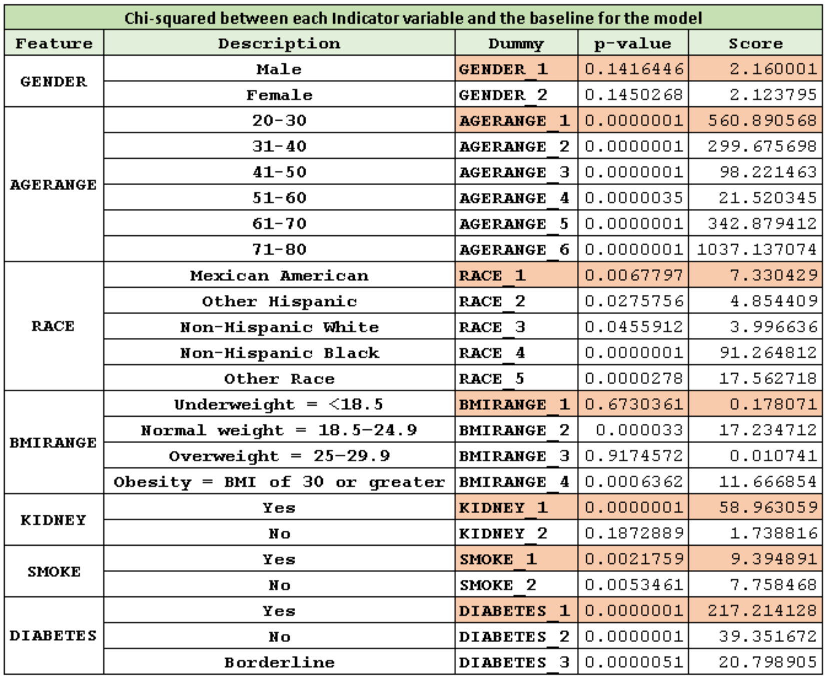
The process of selecting variables for the final model is important to include risk factors with a significant relationship with the dependent variable. However, clinical importance has also been relevant in addition to the statistical significance to select the final independent variables included in the model. We have examined the p-value associated with each variable, and if it was less than the level of significance (0.05) (Uriel, 2013), it could be concluded that the parameter is significant for our classification model (Skelly, 2011). Mostly, the significance level of a test is the probability of making a Type I error, and common values for this significance level in clinical research are 0.10, 0.05 and 0.01. Level of significance of 5% or 0.05 is most commonly used in medicine by consensus of researchers. We have computed chi-squared between each independent variable, and the dependent variable by using Chi2 from the sklearn feature selection package to indicate the strength of evidence that there are some association (Mchugh, 2013), and Table 3 shows the result of this test on the independent variables and Table 4 shows the p-values and score for the indicator variables and the variable baseline or reference variable. Indicator variables have been added to the model to represent better the categorical variables as the binary logic of the predictors is more consistent with the binary outcome and decision making (Garavaglia et al., 1998). The indicator variables provide valuable information about the categories of the independent variables and tend to increase the probability of events, resulting in a more powerful and stable model. For our model, in order to eliminate the redundancy or multicollinearity generated by the indicator variables, we have eliminated the first indicator variable in some cases, and the variable with the indicator ”No” in others. These variables will be our reference variables.



Chi-squared between each feature		
Feature	p-value	Score
GENDER	0.3988107	0.711909
AGERANGE	0.000000	1965.607023
RACE	0.008822	6.858521
BMIRANGE	0.0172385	5.67193
KIDNEY	0.3546428	0.856775
SMOKE	0.0975246	2.745566
DIABETES	0.0012164	10.465222

Table 3: Chi2 test and p-value for the independent variables

As is shown in Table 3 and Table 4 and based on the probability value of each variable, the null hypothesis is rejected for any $0.05 \geq$ p-value, while the null hypothesis is not rejected when $0.05 <$ p-value.



Chi-squared between each Indicator variable and the baseline for the model				
Feature	Description	Dummy	p-value	Score
GENDER	Male	GENDER_1	0.1416446	2.160001
	Female	GENDER_2	0.1450268	2.123795
AGERANGE	20-30	AGERANGE_1	0.0000001	560.890568
	31-40	AGERANGE_2	0.0000001	299.675698
	41-50	AGERANGE_3	0.0000001	98.221463
	51-60	AGERANGE_4	0.0000035	21.520345
	61-70	AGERANGE_5	0.0000001	342.879412
	71-80	AGERANGE_6	0.0000001	1037.137074
RACE	Mexican American	RACE_1	0.0067797	7.330429
	Other Hispanic	RACE_2	0.0275756	4.854409
	Non-Hispanic White	RACE_3	0.0455912	3.996636
	Non-Hispanic Black	RACE_4	0.0000001	91.264812
	Other Race	RACE_5	0.0000278	17.562718
BMIRANGE	Underweight = <18.5	BMIRANGE_1	0.6730361	0.178071
	Normal weight = 18.5–24.9	BMIRANGE_2	0.000033	17.234712
	Overweight = 25–29.9	BMIRANGE_3	0.9174572	0.010741
	Obesity = BMI of 30 or greater	BMIRANGE_4	0.0006362	11.666854
KIDNEY	Yes	KIDNEY_1	0.0000001	58.963059
	No	KIDNEY_2	0.1872889	1.738816
SMOKE	Yes	SMOKE_1	0.0021759	9.394891
	No	SMOKE_2	0.0053461	7.758468
DIABETES	Yes	DIABETES_1	0.0000001	217.214128
	No	DIABETES_2	0.0000001	39.351672
	Borderline	DIABETES_3	0.0000051	20.798905

Table 4: Chi2 test and p-value for all Indicator variables and the baseline for the model

Even though the p-values for the variables GENDER, BMIRANGE_1, BMIRANGE_3, and KIDNEY_2 are not statistically significant at 0.05 level of significance, the clinical importance of these variables in the model for interpretation allows us to include them. We have run the model with the variables and without them, and there were no significant changes in the accuracy score, positive predicted value rate, and true positive rate.

3.7. Model Training

The study population (19,759) was split into a training dataset and a test dataset. The training dataset was derived from a random sampling of 70% (13,831) of the extracted study population and the test sampling the remaining 30% (5,928). We ran the logistic regression model on the entire data set in order to verify the accuracy score of the model. The model accuracy score was 69% (mean accuracy on the given dataset and class labels) with a number of correct predictions of 13,632. This accuracy score is based on the Jaccard similarity coefficient score (Seifoddini & Djassemi, 1991). The Jaccard similarity is used for comparing the similarity and diversity of sample sets and is defined as the size of the intersection divided by the size of the union of the sample sets. After this initial evaluation of the model with the entire dataset, we will train and evaluate the model with the 30% test sampling specified before and the accuracy score was 70% with a number of correct predictions of 4,170. Due that we have an imbalance dependent variable, we have calculated the theoretical accuracy of random guessing on a two-classification problem of 72% that is expressed by:

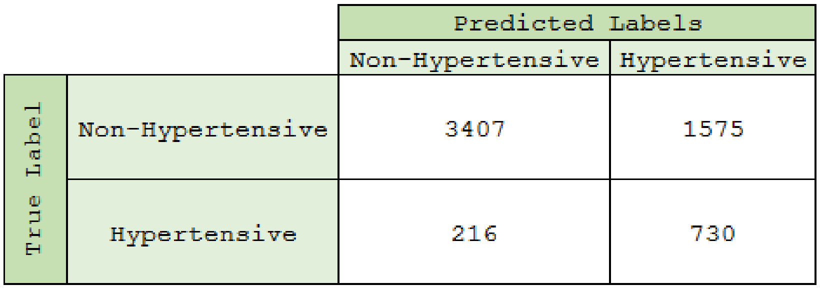
$$P(\hat{y} = y) = P(\hat{y} = 1)P(y = 1) + P(\hat{y} = 0)P(y = 0)$$

Because of the imbalance of the independent variable, and the results of the model accuracy for the entire sample and the test sample no being better than the random guessing, we have avoided using the accuracy score to evaluate the performance of the model. As a result, we need to prove that the model we built is significantly better than random guessing. ROC AUC Score, confusion matrix, and classification report will be used to evaluate the model.

3.8. Model Evaluation

After splitting the data into a training set and a testing set, we have generated some evaluation metrics to evaluate our classifier. In Figure 1, the confusion matrix shows the tabular classification results summary of the actual dependent variable or class label vs. the predicted ones. True positive value (730), True Negative value (3407), False Negative (216) and False Positive value (1575).

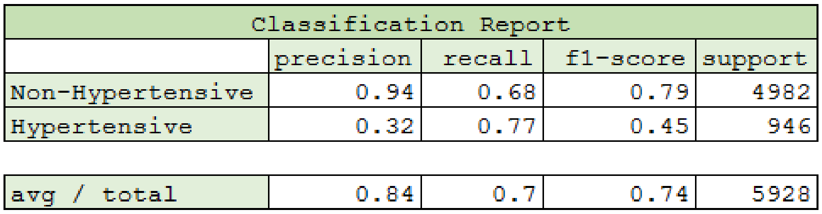
In Figure 2, the classification report for our classifier shows the calculated precision or positive predicted value, the sensitivity or recall and the harmonic mean of precision and sensitivity. The low precision of the hypertensive class value is due to the large number in the false positives category of the confusion matrix, the sensitivity of the model is acceptable, and this is reflected in the



		Predicted Labels	
		Non-Hypertensive	Hypertensive
True Label	Non-Hypertensive	3407	1575
	Hypertensive	216	730

Figure 1: Classifier Confusion Matrix

low number of false negatives, and the values of the F measure indicates the balance between the precision and the sensitivity.



Classification Report				
	precision	recall	f1-score	support
Non-Hypertensive	0.94	0.68	0.79	4982
Hypertensive	0.32	0.77	0.45	946
avg / total	0.84	0.7	0.74	5928

Figure 2: Classification Report

The confusion matrix and the classification report allowed us to see the errors in the predictions for the test data and the exactness (precision) and completeness(recall) of our model.

We have also calculated the performance of our binary classifier by computing the receiver operating characteristic curve (ROC Curve) (Yang et al., 2017). It was created by plotting the true positive rate (0.7716) vs. the false positive rate (0.3161) at various threshold settings. The calculated AUC was 0.73 or 73%, and it is interpreted as the probability that the classifier will assign a higher score to a randomly chosen positive example than to a randomly chosen negative example (Tom, 2005).

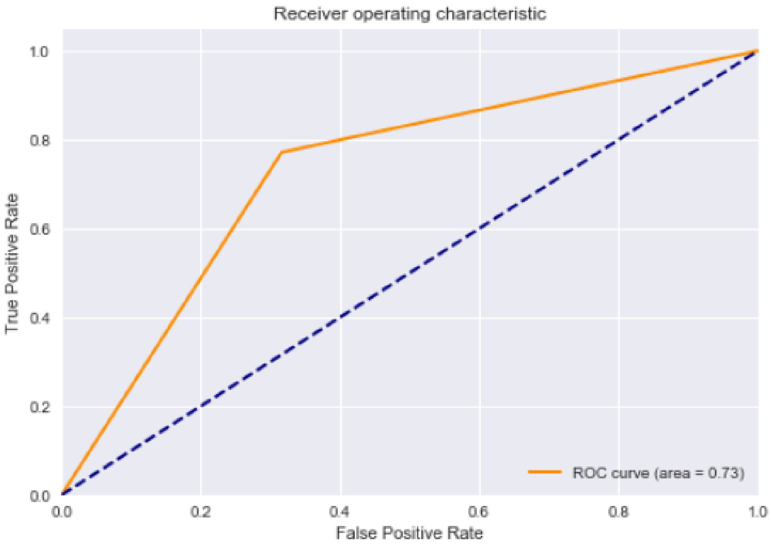


Figure 3: ROC Curve

3.9. Predicting the Probability

265 Our model can be used to estimate the probability that a person will belong to the hypertensive or non-hypertensive class depending on the values of all the independent variables. Using the equation described before to calculate the probabilities of an individual having hypertension:

$$p = \frac{e^{(\beta_0+\beta_1 gender+\beta_2 age+\beta_3 race+\beta_4 bmi+\beta_5 kidney+\beta_6 smoke+\beta_7 diabetes)}}{1+e^{(\beta_0+\beta_1 gender+\beta_2 age+\beta_3 race+\beta_4 bmi+\beta_5 kidney+\beta_6 smoke+\beta_7 diabetes)}}$$

we can define each variable as:

$$gender1 = \left\{ \begin{array}{c} \text{1 if ith person is male} \\ \text{0 if ith person is female} \end{array} \right\}$$

$$gender2 = \left\{ \begin{array}{c} \text{1 if ith person is female} \\ \text{0 if ith person is male} \end{array} \right\}$$

$$Kidney1 = \left\{ \begin{array}{c} \text{1 if ith person has CKD} \\ \text{0 if ith person do not have CKD} \end{array} \right\}$$

$$Kidney2 = \left\{ \begin{array}{c} \text{1 if ith person do not have CKD} \\ \text{0 if ith person is have CKD} \end{array} \right\}$$

If we want to predict the probability of an individual female, of 35 years old, Non-Hispanic white, overweight, with no CKD, no smoking, and borderline diabetes. The values of the variables are:

gender2(0)
agerange2(1) agerange3(0) agerange4(0) agerange5(0) agerange6(0)
race2(0) race3(1) race4(0) race5(0)
bmirange2(0) bmirange3(1) bmirange4(0)
kidney2(1)
smoke2(1)
diabetes2(0) diabetes3(1)

And, our prediction model after rearranging the equation based on the coefficients shown in Table 5 will be:

$$p = \frac{e^{(-0.01996732-0.742861-0.199558-0.083449-0.060981-0.027108+0.068509)}}{1+e^{(-0.01996732-0.742861-0.199558-0.083449-0.060981-0.027108+0.068509)}}$$

Thus, the predicted probability of having hypertension for our test individual will be:

$$p = \frac{0.387}{1.052} = 0.37 = 37\%$$

Then we compare this value with our chosen threshold of 0.5:

$$p(HavingHypertension = Yes) > 0.5$$

For this reason, 0.37 > 0.5 is False, and we can conclude that the individual is not hypertensive for our model.

4. Syntheses and Analysis

Statistical and clinical assessment is important to validate the feasibility and effectiveness of the model. We will discuss the statistical and clinical analysis of the results.

4.1. Statistical Analysis

The test sampling of 5,928 individuals contains 4,982 (84%) non-hypertensive individuals and 946 (16%) hypertensive individuals. The model yields a sensitivity of 730/946 = 77% (positives that were correctly identified) and a specificity of 3407/4982 = 68% (negatives that are correctly identified). The precision of the model or the positive predicted value was 730/2305 =32% and the negative predicted value 3407/3623 =94%. The miss rate or false negative rate of the model was 216/946 = 22% and a fall out rate or false positive rate of 1575/4982 = 31%. The false discovery rate or probability of false alarm of the model was 68%, and the likelihood ratio for negative individuals was 6%. The miss rate and the fall out rate are not very costly for our model due to the imbalance of the sampling. Our model was better at identifying individuals who will not develop hypertension than those that will develop hypertension.

4.2. Clinical Analysis

After calculating the coefficients for each dichotomous independent variable as shown in Table 5, we have converted this estimates onto the odds ratio scale by exponentiating the coefficient estimate e^{β}. In logistic regression the odds ratio represents the constant effect of an independent variable X, on the likelihood that one outcome will occur (Persoskie & Ferrer, 2017). This constant effect can be demonstrated by a general model:

Suppose we have a k independent variable model where k>2 and the variables are binary or continuous:

$$log(\frac{p}{1-p}) = \textbf{log odds of disease} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

Now, we fix values of $x_2, x_3, ..., x_k$, and we get:

$$\textbf{log odds of disease for } x_1 = \mathbf{c} : e^{\beta_0+\beta_1 c+\beta_2 x_2+...+\beta_k x_k}$$

$$x_1 = c+1 : e^{\beta_0+\beta_1 (c+1)+\beta_2 x_2+...+\beta_k x_k}$$

The odds ratio, increasing x_1 by 1 and holding $x_2, x_3, ..., x_k$ fixed at any values is:

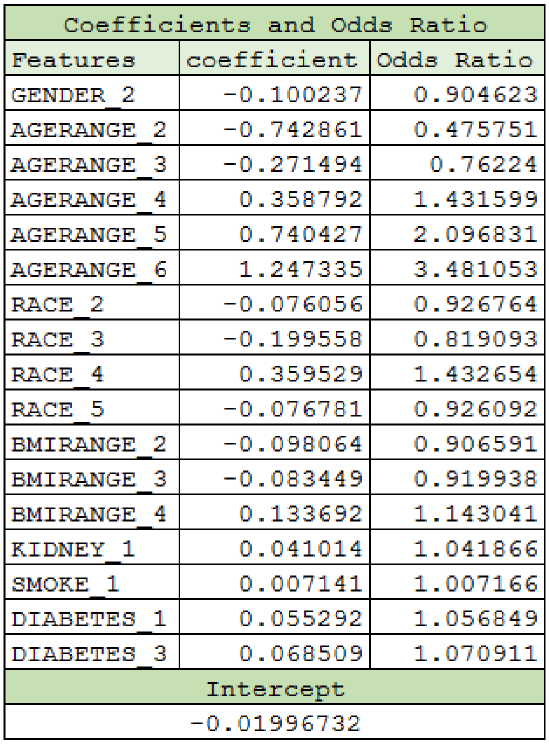
$$\mathbf{OR} = \frac{e^{\beta_0+\beta_1 (c+1)+\beta_2 x_2+...+\beta_k x_k}}{e^{\beta_0+\beta_1 c+\beta_2 x_2+...+\beta_k x_k}} = e^{\beta_1}$$

As a result, e^{β_1} is the increase in odds of disease obtained by increasing x_1 by 1 unit, holding $x_2, x_3, ..., x_k$ fixed. x_1 needs to be binary or continuous, and none of the remaining effects $x_2, x_3, ..., x_k$ can be a product effect.

Consequently, for fixed values of age, race, BMI, kidney, smoke, and diabetes, the odds of having hypertension are approximately 0.9 times higher for female individuals than for the base category male (Table 5). The odds for individuals with age between 71 and 80 years old are 3.4 times higher than the base range 20 to 30 years old. Individuals who identified as non-Hispanic black have 1.4 times higher odds of having hypertension than the rest of race categories. The model also confirms that individuals with obesity have 1.14 times higher odds of having hypertension than individuals with BMI <30. Individuals with diabetes and borderline diabetes have 1.05 and 1.07 times higher odds, respectively, of hypertension compared to those individuals with no diabetes. Based on the odds ratio of individuals that smoke, this independent variable does not affect odds of the outcome.

5. Discussion and Limitations

We have used in our research a set of seven independent variables that were also used across various models we have studied, but not all these variables were included at the same time in either of these models. Six of these variables are included in the Healthy People 2020 Initiative of the Office of Disease Prevention of the U.S. We identified that for our model some independent variables had



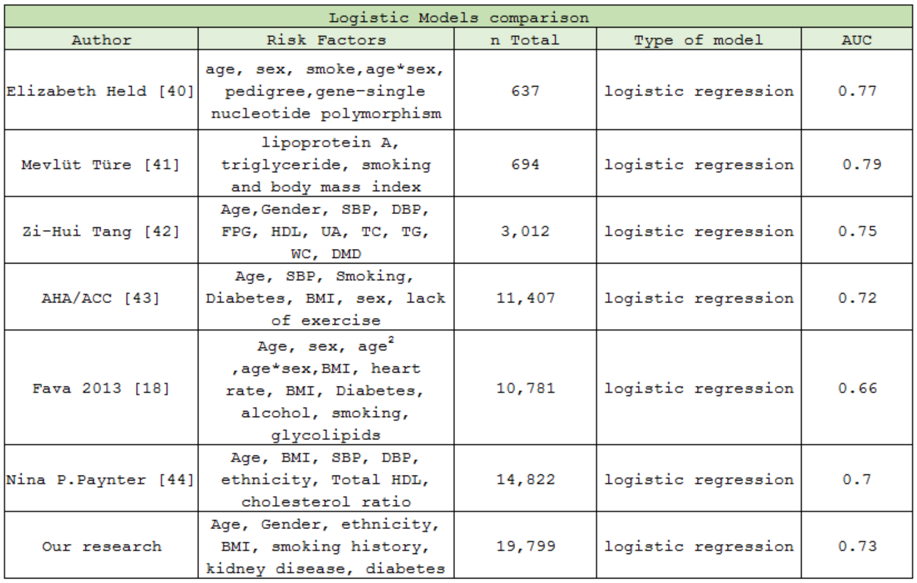
Coefficients and Odds Ratio		
Features	coefficient	Odds Ratio
GENDER_2	-0.100237	0.904623
AGERANGE_2	-0.742861	0.475751
AGERANGE_3	-0.271494	0.76224
AGERANGE_4	0.358792	1.431599
AGERANGE_5	0.740427	2.096831
AGERANGE_6	1.247335	3.481053
RACE_2	-0.076056	0.926764
RACE_3	-0.199558	0.819093
RACE_4	0.359529	1.432654
RACE_5	-0.076781	0.926092
BMIRANGE_2	-0.098064	0.906591
BMIRANGE_3	-0.083449	0.919938
BMIRANGE_4	0.133692	1.143041
KIDNEY_1	0.041014	1.041866
SMOKE_1	0.007141	1.007166
DIABETES_1	0.055292	1.056849
DIABETES_3	0.068509	1.070911
Intercept		
-0.01996732		

Table 5: Coefficients and Odds Ratio

high range values and it was necessary to transform three of these variables into categories and these categories to values between 0 and 1. To our knowledge, this is the first study that uses these risk factors and its categories to evaluate with logistic regression the relationship and behavior between these risk factors and their incidence on the development of hypertension. As shown in Table 6, we reviewed previous studies of risk models to predict hypertension (Echouffo-Tcheugui et al., 2013) and other literature reviews. The number of people in these studies ranged from 637 to 11,407, with the age of participants ranging from less than 25 to 69 years or more and several of them only including a single gender. Age, sex, and smoking are present in almost all of them. For our model, gender was not statistically significant. However, we included it based on its clinical importance for our study. Our model showed the best AUC value 0.73

(95% CI[0.70 - 0.76]) for diagnostic accuracy, indicating fair agreement with the final diagnosis. We compared our model and other logistic regression models with the AUC for prediction of hypertension of the Framingham risk equation (Parikh et al., 2008) which is 0.78 (95% CI[0.75 - 0.81]) for diagnostic accuracy. The Framingham risk score is one of the risk models used for the prediction of cardiovascular risk. The logistic regression model of Elizabeth Held(Held et al., 2015) included age, sex, smoke,age*sex, pedigree and a genesingle nucleotide polymorphism and generated an AUC of 0.77. This model used a prospective cohort of 637 individuals and presented better true positive rate and better true negative rate than our model. Mevlüt Türe(Ture et al., 2005) developed a logistic regression model that excluded age, sex, and smoking and it calculated an AUC of 0.79. Zi-Hui Tang(Tang et al., 2013) developed a model which include waist circumference, fasting plasma glucose and systolic and diastolic blood pressure measurements in addition to the typical variables with an AUC of 0.75. The Swedish risk model(Fava et al., 2013) implemented by Cristiano Fava with an AUC of 0.66, the American Heart Association/American College of Cardiology AHA/ACC(Goff et al., 2014) model with an AUC of 0.72 (95% CI[0.70 - 0.74]) and the model implemented by Nina P.Paynter(Paynter et al., 2009) with AUC of 0.7. All the models studied models indicated fair agreement with the final diagnosis for AUC values between 0.7-0.8. Our model has similar predictive accuracy. Models for predicting hypertension have potential public health and clinical applications in the prevention of hypertension, but more work is needed to develop other models incorporating more statistically significant variables and to perform additional exploration with different balanced datasets.

Several limitations of our study deserve comment. By the time we finished the study, the American Heart Association, the American College of Cardiology and nine other groups redefined high blood pressure as a reading of 130 over 80, down from 140 over 90. The change, the first in 14 years, means that 46 percent of U.S. adults, many of them under the age of 45, now will be considered hypertensive. Under the previous guideline, 32 percent of U.S. adults had high blood pressure(Whelton et al., 2017). We selected the independent variables based



Logistic Models comparison				
Author	Risk Factors	n Total	Type of model	AUC
Elizabeth Held [40]	age, sex, smoke,age*sex, pedigree,gene-single nucleotide polymorphism	637	logistic regression	0.77
Mevlüt Türe [41]	lipoprotein A, triglyceride, smoking and body mass index	694	logistic regression	0.79
Zi-Hui Tang [42]	Age,Gender, SBP, DBP, FPG, HDL, UA, TC, TG, WC, DMD	3,012	logistic regression	0.75
AHA/ACC [43]	Age, SBP, Smoking, Diabetes, BMI, sex, lack of exercise	11,407	logistic regression	0.72
Fava 2013 [18]	Age, sex, age^2 ,age*sex,BMI, heart rate, BMI, Diabetes, alcohol, smoking, glycolipids	10,781	logistic regression	0.66
Nina P.Paynter [44]	Age, BMI, SBP, DBP, ethnicity, Total HDL, cholesterol ratio	14,822	logistic regression	0.7
Our research	Age, Gender, ethnicity, BMI, smoking history, kidney disease, diabetes	19,799	logistic regression	0.73

Table 6: Logistic Models Comparison

on experimental and clinical data that shows the ability of these risk factors to provide insights for predicting hypertension, but we did not include other important factors as family history, diet habits, alcohol intake, stress and other chronic conditions that can increase the risk of hypertension. The sampling race of our study was predominately white and non-hypertensive. Thus, our study may not be generalizable to other races. It is also known that an increase in the number of independent variables, increase the complexity of the model and this can cause over-fitting producing not expected results, we addressed this problem by pre-training the model to select the best hyper-parameters and regularization methods. We didn't use ensemble learning to construct a balanced dataset for our model to enhance prediction performance, but this option should be explored in future research.

6. Conclusions and Future work

This research has shown machine-learning classification algorithms have a lot of potential in helping to predict, evaluate and detect cardiovascular disease cases, increasing the number of patients detected that could benefit from the healthy people initiative 2020 of the government of the United States. In this case, our model is an explanatory model that allow a better understanding of these risk factors and how they are associated with the dependent variable. The use of non-intrusive risk factors allows creating programs to identify individuals at high risk for hypertension to direct them for treatment. There are other machine learning models based on SVMs, K-NN, decision trees and neural networks with high classification accuracies but we have chosen logistic regression for this research because of the easy interpretation of the results for clinical purposes. Another conclusion that emerges from this research was that removing gender, race and smoke factors from the data do not affect the accuracy either the AUC of the model. This logistic regression model needs to be trained and tested in a large and heterogeneous primary care patient population. Regardless the model showed the best AUC value 0.73 (95% CI[0.70 - 0.76]) indicating fair agreement with the final diagnosis, more work will continue on this model to improve the diagnosis accuracy.

Based on the present research we can use other algorithms to classify and predict hypertension with higher accuracy. Future research will use and compare different classification methods like k-nearest neighbors (kNN), Support vector machines (SVM), Decision tree classifier and Random forest to evaluate the association of several risk factors with hypertension. Future research also will incorporate more risk factors like Weight, Standing Height, Waist Circumference, Albumin creatinine ratio, Glycohemoglobin, Total Cholesterol, HDL-Cholesterol, Triglycerides and Potassium levels. The future model should include a dependent variable with more than two levels, in this case for hypertension, the levels will be calculated based on systolic and diastolic blood pressure. The levels will be normal, Prehypertension, Hypertension Stage 1, Hypertension Stage 2 and Hypertensive Crisis.

Acknowledgments

This document presents independent research funded by the company SYS-DEVELOPMENT LLC. The points of view expressed are those of the authors and not necessarily those of the NIHR, the NHS, the NHANES or the department of health. We thank Babafemi Thompson, Clinical Systems Analyst at Englewood Hospital who provided insight and expertise that greatly assisted the research.

References

Ahmad, W. M. A. W., Nawi, M. A. B. A., Aleng, N. A., Halim, N. A., Mamat, M., & Pouzi, M. (2014). Association of hypertension with risk factors using logistic regression. *Applied Mathematical Sciences*, *8*, 2563–2572. URL: http://www.m-hikari.com/ams/ams-2014/ams-49-52-2014/alengAMS49-52-2014.pdfhttp://www.m-hikari.com/ams/ams-2014/ams-49-52-2014/42130.html. doi:10.12988/ams.2014.42130.

Center for Disease Control and Prevention (2015). *Percentage with CKD stage 3 or 4 who were aware of their disease by stage and age 1999-2012*. Technical Report. URL: http://nccd.cdc.gov/ckd/detail.aspx?QNum=Q98.

Centers for Disease Control (2016). *Current Cigarette Smoking Prevalence Among Working Adults United States , 2004 2010*. Technical Report Morbidity and Mortality Weekly Report (MMWR). URL: https://www.cdc.gov/mmwr/preview/mmwrhtml/mm6038a2.htm.

Chen, M., Hao, Y., Hwang, K., Wang, L., & Wang, L. (2017). Disease Prediction by Machine Learning over Big Data from Healthcare Communities. *IEEE Access*, *5*, 8869–8879. URL: http://ieeexplore.ieee.org/document/7912315/. doi:10.1109/ACCESS.2017.2694446.

Committee on Public Health Priorities to Reduce and Control Hypertension in the U.S. Population. (2010). *A Population-Based Policy and Systems Change Approach to Prevent and Control Hypertension*. URL: http://www.nap.edu/catalog.php?record{_}id=12819. doi:10.17226/12819.

Echouffo-Tcheugui, J. B., Batty, G. D., Kivimäki, M., & Kengne, A. P. (2013). Risk Models to Predict Hypertension: A Systematic Review. URL: http://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0067370{&}type=printable. doi:10.1371/journal.pone.0067370.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., & Lin, C.-J. (2008). LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, *9*, 1871–1874. URL: `http://www.csie.ntu.edu.tw/{~}cjlin/papers/liblinear.pdf`. doi:`10.1038/oby.2011.351`.

Fava, C., Sjögren, M., Montagnana, M., Danese, E., Almgren, P., Engström, G., Nilsson, P., Hedblad, B., Guidi, G. C., Minuz, P., & Melander, O. (2013). Prediction of blood pressure changes over time and incidence of hypertension by a genetic risk score in swedes. *Hypertension*, *61*, 319–326. URL: `http://www.ncbi.nlm.nih.gov/pubmed/23232644`. doi:`10.1161/HYPERTENSIONAHA.112.202655`.

Frunza, O., Inkpen, D., & Tran, T. (2011). A machine learning approach for identifying disease-treatment relations in short texts. *IEEE Transactions on Knowledge and Data Engineering*, *23*, 801–814. URL: `http://ieeexplore.ieee.org/document/5560656/`. doi:`10.1109/TKDE.2010.152`.

Garavaglia, S., Sharma, A., & Hill, M. (1998). a Smart Guide To Dummy Variables : Four Applications and a Macro. *Entropy*, . URL: `https://stats.idre.ucla.edu/wp-content/uploads/2016/02/p046.pdfhttp://www.ats.ucla.edu/stat/sas/library/nesug98/p046.pdf`.

Goff, D. C., Lloyd-Jones, D. M., Bennett, G., Coady, S., D'Agostino, R. B., Gibbons, R., Greenland, P., Lackland, D. T., Levy, D., O'Donnell, C. J., Robinson, J. G., Schwartz, J. S., Shero, S. T., Smith, S. C., Sorlie, P., Stone, N. J., & Wilson, P. W. F. (2014). 2013 ACC/AHA guideline on the assessment of cardiovascular risk: A report of the American college of cardiology/American heart association task force on practice guidelines. doi:`10.1016/j.jacc.2013.11.005`.

Held, E., Cape, J., & Tintle, N. (2015). Comparing machine learning and logistic regression methods for predicting hypertension using a combination of gene expression and next-generation sequencing data. *BMC proceedings*, *To appear.*, 1–5. URL: `https://bmcproc.biomedcentral.com/track/pdf/10.`

1186/s12919-016-0020-2?site=bmcproc.biomedcentral.comhttp://dx.doi.org/10.1186/s12919-016-0020-2. doi:10.1186/s12919-016-0020-2.

Hijazi, S., Page, A., Kantarci, B., & Soyata, T. (2016). Machine Learning in Cardiac Health Monitoring and Decision Support. *Computer*, *49*, 38–48. URL: http://ieeexplore.ieee.org/document/7742297/. doi:10.1109/MC.2016.339.

Kshirsagar, A. V., Chiu, Y.-l., Bomback, A. S., August, P. A., Viera, A. J., Colindres, R. E., & Bang, H. (2010). A hypertension risk score for middle-aged and older adults. *Journal of clinical hypertension (Greenwich, Conn.)*, *12*, 800–8. URL: http://doi.wiley.com/10.1111/j.1751-7176.2010.00343.xhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3683833{&}tool=pmcentrez{&}rendertype=abstract. doi:10.1111/j.1751-7176.2010.00343.x. arXiv:NIHMS150003.

Kublanov, V. S., Dolganov, A. Y., Belo, D., & Gamboa, H. (2017). Comparison of Machine Learning Methods for the Arterial Hypertension Diagnostics. *Applied Bionics and Biomechanics*, *2017*, 1–13. URL: https://www.hindawi.com/journals/abb/2017/5985479/https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5555018/pdf/ABB2017-5985479.pdf{%}0Ahttps://www.hindawi.com/journals/abb/2017/5985479/. doi:10.1155/2017/5985479.

Lin, C.-j. J., Weng, R. C., & Keerthi, S. S. (2008). Trust Region Newton Method for Logistic Regression. *Journal of Machine Learning Research*, *9*, 627–650. URL: https://www.csie.ntu.edu.tw/{~}cjlin/papers/logistic.pdfhttp://portal.acm.org/citation.cfm?id=1390681.1390703.

Manandhar, N. (2016). Risk factors of Hypertension: Logistic regression analysis. *SCIREA Journal of Health*, . URL: http://www.scirea.org/journal/PMH.

Mchugh, M. L. (2013). The Chi-square test of independence Lessons in biostatistics. *Biochemia Medica*, *23*, 143–9. URL: http://www.ncbi.nlm.nih.gov/

pubmed/23894860http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3900058http://dx.doi.org/10.11613/BM.2013.018. doi:10.11613/BM.2013.018.

Miller, W. G. (2009). Estimating glomerular filtration rate. *Clinical Chemistry and Laboratory Medicine*, *47*, 1017–1019. URL: https://www.niddk.nih.gov/health-information/communication-programs/nkdep/laboratory-evaluation/glomerular-filtration-rate/estimating. doi:10.1515/CCLM.2009.264.

Mozaffarian, D., & Benjamin, E. J. (2015). Heart disease and stroke statistics-2015 update : A report from the American Heart Association. *Circulation*, *131*, e29–e39. URL: http://www.ncbi.nlm.nih.gov/pubmed/25520374. doi:10.1161/CIR.0000000000000152. arXiv:15334406.

Mozaffarian, D., & Benjamin, E. J. (2016). Heart disease and stroke statistics-2016 update a report from the American Heart Association. URL: http://www.ncbi.nlm.nih.gov/pubmed/26673558. doi:10.1161/CIR.0000000000000350. arXiv:NIHMS150003.

National Center for Health Statistics (2017). *Health, United States, 2016: With Chartbook on Long-term Trends in Health*. Technical Report. URL: https://www.cdc.gov/nchs/data/hus/hus16.pdf{#}053https://www.cdc.gov/nchs/data/hus/hus16.pdf{#}019.

Nilashi, M., bin Ibrahim, O., Ahmadi, H., & Shahmoradi, L. (2017). An analytical method for diseases prediction using machine learning techniques. *Computers & Chemical Engineering*, *106*, 212–223. URL: https://ac.els-cdn.com/S0098135417302570/1-s2.0-S0098135417302570-main.pdf?{_}tid=4f3bcfbe-adcd-11e7-a2cb-00000aab0f6c{&}acdnat=1507648478{_}e70b8948658627c0f093ee1e9e50461ahttp://linkinghub.elsevier.com/retrieve/pii/S0098135417302570. doi:10.1016/j.compchemeng.2017.06.011.

Nwankwo, T., Yoon, S. S., Burt, V., & Gu, Q. (2013). Hypertension among adults in the United States: National Health and Nutrition Examination Survey, 2011-2012. *NCHS data brief*, (pp. 1–8). URL: http://www.ncbi.nlm.nih.gov/pubmed/24171916. doi:10.1017/CBO9781107415324.004. arXiv:arXiv:1011.1669v3.

Olubiyi, A. (2013). On Statistical Analysis of Blood Pressure with respect to Some Demographic Factors. *International Journal of Health Sciences*, *1*, 53–61. URL: www.aripd.org/ijhs.

Parikh, N. I., Pencina, M. J., Wang, T. J., Benjamin, E. J., Lanier, K. J., Levy, D., D'Agostino, R. B., Kannel, W. B., & Vasan, R. S. (2008). A risk score for predicting near-term incidence of hypertension: The Framingham Heart Study. *Annals of Internal Medicine*, *148*, 102–110. doi:10.7326/0003-4819-148-2-200801150-00005.

Paynter, N. P., Cook, N. R., Everett, B. M., Sesso, H. D., Buring, J. E., & Ridker, P. M. (2009). Prediction of Incident Hypertension Risk in Women with Currently Normal Blood Pressure. *American Journal of Medicine*, *122*, 464–471. URL: http://www.sciencedirect.com/science/article/pii/S0002934308011832?via{%}3Dihub{#}bbib24. doi:10.1016/j.amjmed.2008.10.034.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2012). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. URL: http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdfhttp://dl.acm.org/citation.cfm?id=2078195{%}5Cnhttp://arxiv.org/abs/1201.0490. doi:10.1007/s13398-014-0173-7.2. arXiv:1201.0490.

Persoskie, A., & Ferrer, R. A. (2017). A Most Odd Ratio:: Interpreting and Describing Odds Ratios. *American Journal of*

Preventive Medicine, *52*, 224–228. URL: https://ac.els-cdn.com/S0749379716303087/1-s2.0-S0749379716303087-main.pdf?{_}tid=f2270912-b342-11e7-83e7-00000aacb35e{&}acdnat=1508248759{_}ac8486c10b2456e56c401aae46eff40b. doi:10.1016/j.amepre.2016.07.030.

Ramezankhani, A., Azizi, F., Hadaegh, F., & Eskandari, F. (2017). Sex-specific clustering of metabolic risk factors and their association with incident cardiovascular diseases: A population-based prospective study. *Atherosclerosis*, *263*, 249–256. URL: https://ac.els-cdn.com/S002191501731170X/1-s2.0-S002191501731170X-main.pdf?{_}tid=faa1155c-b594-11e7-8f57-00000aacb362{&}acdnat=1508503893{_}46fee2cecefd2f11df63095b0db599f2. doi:10.1016/j.atherosclerosis.2017.06.921.

Sainani, K. L. (2014). Statistically Speaking Logistic Regression. *PM&R*, *6*, 1157–1162. URL: https://ac.els-cdn.com/S1934148214014439/1-s2.0-S1934148214014439-main.pdf?{_}tid=ef0269fc-add4-11e7-a835-00000aab0f26{&}acdnat=1507651751{_}266e31c6208eced8bb26e2131315cf51. doi:10.1016/j.pmrj.2014.10.006. arXiv:arXiv:1305.6995v1.

Samuel, A. (1959). Some studies in machine learning using the game of checkers. *Ibm Journal*, *3*, 210. URL: http://researcher.watson.ibm.com/researcher/files/us-beygel/samuel-checkers.pdfhttp://pages.cs.wisc.edu/{~}dyer/cs540/handouts/samuel-checkers.pdf. doi:10.1016/0066-4138(69)90004-4.

Seffens, W., Evans, C., & Taylor, H. (2015). Machine Learning Data Imputation and Classification in a Multicohort Hypertension Clinical Study. *Bioinformatics and biology insights*, *9*, 43–54. URL: http://www.ncbi.nlm.nih.gov/pubmed/27199552http://www.pubmedcentral.

nih.gov/articlerender.fcgi?artid=PMC4862746http://journals.sagepub.com/doi/10.4137/BBI.S29473. doi:10.4137/BBI.S29473.

Seifoddini, H., & Djassemi, M. (1991). THE PRODUCTION DATA-BASED SIMILARITY COEFFICIENT VERSUS JACCARD'S SIMILARITY COEFFICIENT. *Computers ind. Engng*, *21*, 263–266. URL: https://ac.els-cdn.com/036083529190099R/1-s2.0-036083529190099R-main.pdf?{_}tid=7217d320-af62-11e7-98d8-00000aacb35d{&}acdnat=1507822482{_}24a86e9db2b0ebb43998b272bc5be645.

Skelly, A. C. (2011). Probability, proof, and clinical significance. *Evidence-based spine-care journal*, *2*, 9–11. URL: http://www.ncbi.nlm.nih.gov/pubmed/23230400http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3506143http://www.ncbi.nlm.nih.gov/pubmed/23230400{%}5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3506143. doi:10.1055/s-0031-1274751.

Song, K., Nie, F., Han, J., & Li, X. (2017). Rank-k 2-D Multinomial Logistic Regression for Matrix Data Classification. URL: http://ieeexplore.ieee.org/document/8010841/. doi:10.1109/TNNLS.2017.2731999.

Tang, Z.-H., Liu, J., Zeng, F., Li, Z., Yu, X., & Zhou, L. (2013). Comparison of Prediction Model for Cardiovascular Autonomic Dysfunction Using Artificial Neural Network and Logistic Regression Analysis. *PLoS ONE*, *8*, e70571. URL: http://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0070571{&}type=printablehttp://dx.plos.org/10.1371/journal.pone.0070571. doi:10.1371/journal.pone.0070571.

Tedla, F. M., Brar, A., Browne, R., & Brown, C. (2011). Hypertension in Chronic Kidney Disease: Navigating the Evidence. *International Journal of Hypertension*, *2011*, 1–9. URL: http://www.ncbi.nlm.nih.gov/pubmed/21747971http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3124254http://www.hindawi.com/journals/ijhy/2011/132405/. doi:10.4061/2011/132405.

Tom, F. (2005). An introduction to ROC analysis. *IRBM*, *35*, 299–309. URL: http://www.cs.bu.edu/faculty/betke/cs585/restricted/papers/Fawcett-ROC-2006.pdf. doi:10.1016/j.patrec.2005.10.010. arXiv:/dx.doi.org/10.1016/j.patrec.200.

Ture, M., Kurt, I., Turhan Kurum, A., & Ozdamar, K. (2005). Comparing classification techniques for predicting essential hypertension. *Expert Systems with Applications*, *29*, 583–588. URL: https://pdfs.semanticscholar.org/14ca/bb4509d6285a4c516f15c164a55a6bdd04d6.pdf. doi:10.1016/j.eswa.2005.04.014.

Uriel, E. (2013). Hypothesis testing in the multiple regression model. *Statistics*, (pp. 1–49). URL: https://www.uv.es/uriel/4Hypothesistestinginthemultipleregressionmodel.pdf. doi:http://popstats.unhcr.org/en/overview.

U.S. Department of Health and Human (2005). Awareness of Prediabetes United States, 20052010. *Centers for Disease Control & Prevention Source: Morbidity and Mortality Weekly Report Centers for Disease Control & Prevention*, *62*, 209–212. URL: https://www.cdc.gov/mmwr/preview/mmwrhtml/mm6211a4.htmhttp://www.jstor.org/stable/24846089{%}5Cnhttp://www.jstor.org/stable/24846089?seq=1{&}cid=pdf-reference{#}references{_}tab{_}contents{%}5Cnhttp://about.jstor.org/terms.

Whelton, P. K., Committee, W., Carey, R. M., Chair, V., Aronow, W. S., Committee Member, W., Casey, D. E., Collins, K. J., Dennison Himmelfarb, C., DePalma, S. M., Gidding, S., Jamerson, K. A., Jones, D. W., MacLaughlin, E. J., Muntner, P., Ovbiagele, B., Smith, S. C., Spencer, C. C., Stafford, R. S., Taler, S. J., Thomas, R. J., Williams, K. A., Williamson, J. D., & Wright, J. T. (2017). 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults. *Journal of the American College of Cardiology*, (pp. 735–1097). URL: http://linkinghub.elsevier.com/retrieve/pii/S0735109717415191. doi:10.1016/j.jacc.2017.11.006.

World Health Organization (2017). *World Health Statistics 2017 : Monitoring Health for The SDGs*. URL: http://apps.who.int/iris/bitstream/10665/255336/1/9789241565486-eng.pdf?ua=1. doi:10.1017/CBO9781107415324.004. arXiv:arXiv:1011.1669v3.

Yang, Z., Zhang, T., Lu, J., Zhang, D., & Kalui, D. (2017). Optimizing area under the ROC curve via extreme learning machines. *Knowledge-Based Systems*, *130*, 74–89. URL: https://ac.els-cdn.com/S0950705117302241/1-s2.0-S0950705117302241-main.pdf?{_}tid=c4d9f486-af84-11e7-98d8-00000aacb35d{&}acdnat=1507837224{_}02667824ac22abfc9dd08e8f77a6cf05. doi:10.1016/j.knosys.2017.05.013.

Zheng, Z., Li, Y., & Cai, Y. (2013). The Logistic Regression Analysis on Risk Factors of Hypertension among Peasants in East China & Its Results Validating, . *10*, 416–420. URL: https://www.ijcsi.org/papers/IJCSI-10-2-1-416-420.pdf.