Prediction of Hypertension Risk

Sidar YÜK
Middle East Technical University
Ankara, Turkey
sidar.yuk@metu.edu.tr

Abstract— The main aim of this project is to estimate the risk status of hypertension. To better understand the dataset, exploratory data analysis and confirmatory data analysis were performed, and various visualization and statistical tests were performed while performing these analyses. The classification and regression trees (CART) method was applied for imputation. A more balanced classification in the dataset needed to be made. Using Caret and Rose libraries in R, Under-sampling, Over-sampling, Rose, and Smote methods were compared, respectively, and metrics such as accuracy and sensitivity were used as performance metrics, and Smote gave the best result for the imbalance problem. After the dataset was cleaned, the prediction of risk status was estimated with different machine learning methods, such as support vector machine, random forest, and artificial neural network. The performance of the created models was evaluated based on accuracy, sensitivity, specificity, and F1 score. All these operations were done through R-studio.

Keywords—Xgboost,Smote, Neural Network, Classification, Hypertension

I. INTRODUCTION

Hypertension is an insidious disease that often does not cause symptoms. That is why many hypertension patients are unaware of their disease. This is why it is called the "silent killer." Hypertension, when it reaches a dangerous level, leads to severe complications such as heart attack, stroke, heart failure, or kidney disease. Therefore, accurate estimation of risk is fundamental.

In this study, individuals were classified as low and high hypertension risk. Different machine learning methods were used to find the most accurate prediction: logistic regression support vector machine, random forest, decision tree, xgboost, and artificial neural network, respectively. All models' accuracy, sensitivity, specificity, and F1 scores were calculated and compared.

II. LITERATURE REVIEW

Statisticians and researchers have many studies on hypertension.[1] Firstly, four different models are applied in this study: random forest, CatBoost, MLP neural network, and logistic regression. When these four models were compared, the most accurate result was obtained with random forest.[2] Secondly, in a study conducted in China, the LightGMB model is recommended for hypertension prediction.[3] Finally, in the research conducted in Bangladesh, support vector machine recursive feature elimination (SVMRFE), artificial neural network, decision tree, random forest, and gradient boosting methods are applied for hypertension prediction, and the most accurate result is obtained with the SVMRFE method.

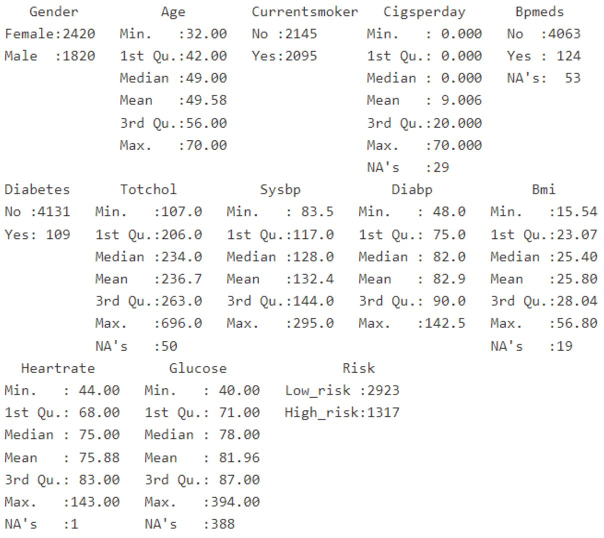
III. METHODOLOGY

A. Dataset

Hypertension risk model data was taken from Kaggle. It consists of 13 variables in total. Five of the variables are categorical, and 8 of them are numerical. The "Risk, identified as the dependent variable for this study, is determined by a range of covariates." The dataset contains 4240 observations in total. In addition, there is no duplicated problem, but 1% of the dataset consists of NA values. NA values were imported. There are various packages in R for this process. The details of this process will be seen in more detail in the following steps.

•**"Gender":** Gender of individual-Categoric variable
•**"Age":** Age of person - Continuous variable
•**"Currensmoker":** Smoking status – Categoric variable
•**"Cigsperday":**Daily cigarette count – Continuous variable
•**"BPMeds":** Blood pressure meds use – Categoric variable
•**"Diabetes":** Diabetes status – Categoric variable
•**"Totchol":** Total cholesterol level – Continuous variable
•**"Sysbp":** Systolic blood pressure – Continuous variable
•**"Diabp":** Diastolic blood pressure – Continuous variable
•**"Bmi":** Body mass index – Continuous variable
•**"Heartrate":** Heart rate – Continuous variable
•**"Glucose":** Glucose level – Continuous variable
•**"Risk":** Hypertension risk status – Categoric variable

B. Descriptive Statistics



```
   Gender          Age        Currentsmoker   Cigsperday      Bpmeds
 Female:2420   Min.   :32.00   No :2145      Min.   : 0.000   No  :4063
 Male  :1820   1st Qu.:42.00   Yes:2095      1st Qu.: 0.000   Yes : 124
               Median :49.00                 Median : 0.000   NA's:  53
               Mean   :49.58                 Mean   : 9.006
               3rd Qu.:56.00                 3rd Qu.:20.000
               Max.   :70.00                 Max.   :70.000
                                             NA's   :29
 Diabetes      Totchol          Sysbp           Diabp            Bmi
 No :4131   Min.   :107.0   Min.   : 83.5   Min.   : 48.0   Min.   :15.54
 Yes: 109   1st Qu.:206.0   1st Qu.:117.0   1st Qu.: 75.0   1st Qu.:23.07
            Median :234.0   Median :128.0   Median : 82.0   Median :25.40
            Mean   :236.7   Mean   :132.4   Mean   : 82.9   Mean   :25.80
            3rd Qu.:263.0   3rd Qu.:144.0   3rd Qu.: 90.0   3rd Qu.:28.04
            Max.   :696.0   Max.   :295.0   Max.   :142.5   Max.   :56.80
            NA's   :50                                      NA's   :19
   Heartrate          Glucose            Risk
 Min.   : 44.00   Min.   : 40.00   Low_risk :2923
 1st Qu.: 68.00   1st Qu.: 71.00   High_risk:1317
 Median : 75.00   Median : 78.00
 Mean   : 75.88   Mean   : 81.96
 3rd Qu.: 83.00   3rd Qu.: 87.00
 Max.   :143.00   Max.   :394.00
 NA's   :1        NA's   :388
```

Table 1 Descriptive Summary of Data

According to the summary chart, no categorical variables contain na value except bpmeds. Furthermore, there are some imbalances in the categorical levels. When we look at the diabetes and Bpmeds variables, there is a massive difference in numbers between the levels. When we examine the numeric variables, Na is mainly seen in glucose. A large part of the Na ratio of the dataset is from Na values in glucose. Another critical point is that in cigsperday, the minimum and median are equal to 0

C. Explaratory Data Analysis

It is imperative to understand the data correctly before making analyses and predictions. A total of 6 questions were created in this part. The questions created were evaluated with data visualizations and then interpreted with various statistical tests.

C.1 How is the relationship between Diabp and Sysbp, and is there any observable pattern when considering the Risk status?

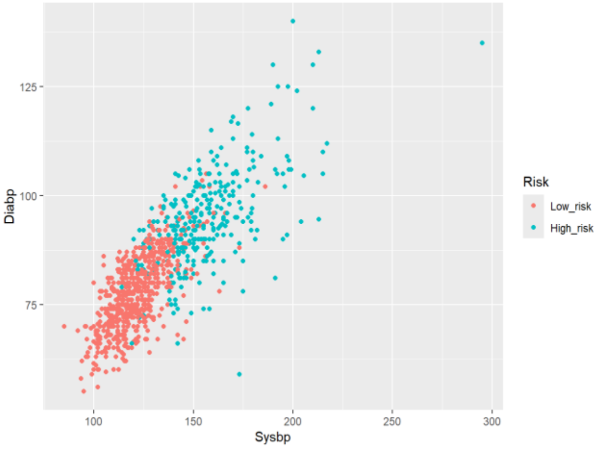


Figure 1 Bubble Plot for C.1

There is a strong relationship between Diabp and Syspb. Therefore, The risk of hypertension appears to be low when diabp and sysbp are low. We observe that as Diabp and Sysybp values increase, the risk of hypertension changes from Low risk to High risk.

glm(formula = Risk ~ Sysbp + Diabp, family = binomial, data = complete_data)
Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -22.365005 0.727157 -30.76 <2e-16
Sysbp 0.111551 0.004768 23.39 <2e-16
Diabp 0.074927 0.007153 10.47 <2e-16
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 5241.3 on 4239 degrees of freedom
Residual deviance: 2509.3 on 4237 degrees of freedom
AIC: 2515.3
Number of Fisher Scoring iterations: 6

Table 2 Result of Logistic Regression for C.1

When we look at the summary, all predictors are statistically significant. A multicollinearity test was performed, and all values < 10, so there is no multicollinearity problem. Then, the likelihood ratio test was performed, giving us the following result: at least one of the explanatory variables is significant. In addition, the ANOVA test was performed, and the ANOVA test showed us that Sysbp and Diabp played an essential role in the model ($p < 0.05$). As a result, the overall model was significant, and Sysbp and Diabp significantly affected the log odds.

C.2 Is there any relationship between Bpmeds and risk?

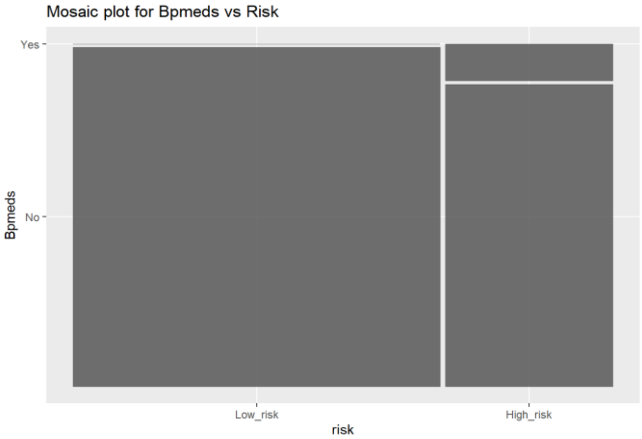


Figure 2 Mosaic Plot for C.2

When we look at the risk situations of blood pressure medication usage, those who use medication have a lower risk of hypertension than those who do not.

Some groups have less than five frequencies, so Fisher’s exact test must be implemented.

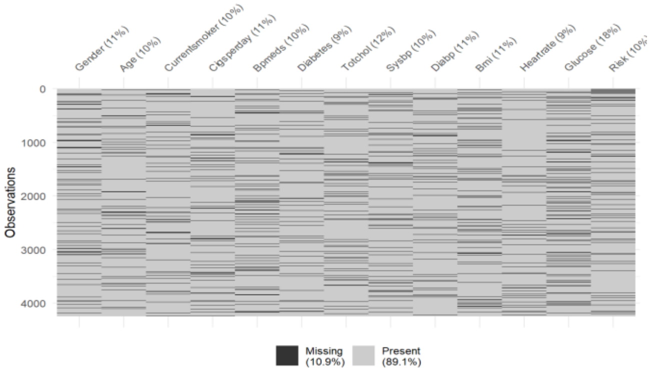
Fisher's Exact Test for Count Data
data: Bpmeds_risk
p-value < 2.2e-16
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
54.33934 11326.06387
sample estimates:
odds ratio
309.01

Table 3 Fisher's Exact Test for C.2

As is seen in Fisher’s Exact Test, since the p-value is less than alpha, there is a significant relation between bp meds and risk. The odds ratio is relatively high, approximately 310. This shows that people who use blood pressure medication are 310 times more likely to be in high-risk status than those who do not use blood pressure medication

D. Missingness

NA values increased by 10% to cover the entire dataset



Missing data occur randomly, independent of other observed or unobserved variables. Truly missed data may not have been recorded due to carelessness or other reasons, but this loss has no direct relationship to the examined person. In this case, the missing data is completely randomly distributed, called Missing Completely At Random (MCAR). The classification and regression trees (CART) method was applied for imputation for a mice package because the cart method imputes missing values based on the values of other variables. also, the cart is easily applicable and can handle categorical variables.

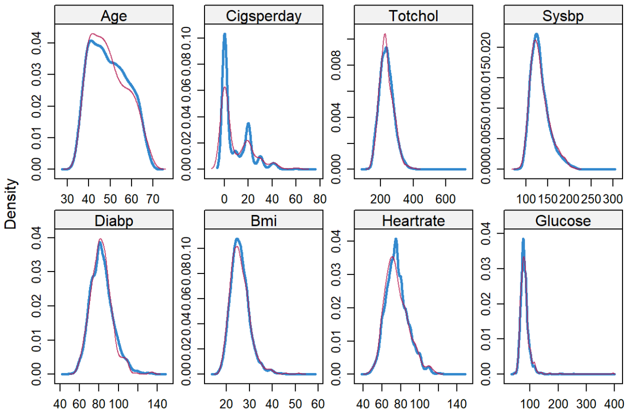


Figure 3 Density Plot of Before-After Imputation

As seen from the density plots, the imputation for variables does not cause a considerable change in the distribution of the variable;

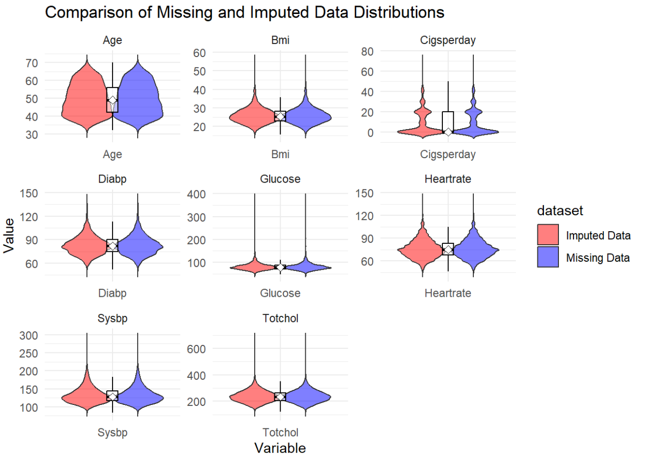


Figure 4 Violin Plots of Before-After Imputation

as seen in the violin plot, the distributions, medians, and interquartile ranges are very similar to each other. We can draw the following conclusion from these two plots: The values filled in with the cart method show that they do not disrupt the original data structure

E. Cross-Validation Addressing Imbalance Issues

The k-repeated cross-validation method was used to evaluate the model's performance and test its generalizability. The data set was divided into different subsets by repeated cross-validation. The model was trained and tested on each subset. This process used oversampling, undersampling, rose, and smote to combat imbalance problems. Additionally, cross-validation was performed to evaluate how these techniques affected model performance on each subset. Cross-validation increased the generalizability of the model by reducing the risk of overfitting.

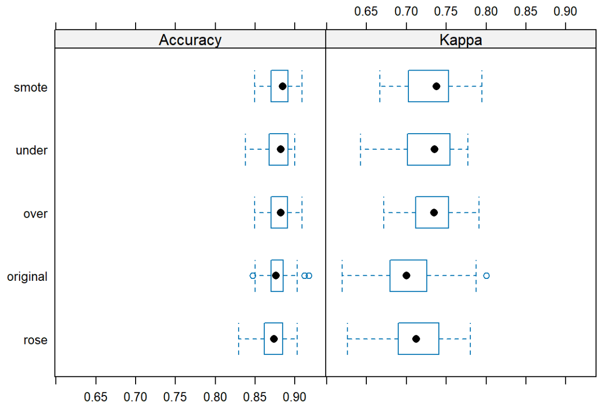


Figure 5 Comparison of Imbalanced Methods-1

According to the table, the smote method is the most accurate.

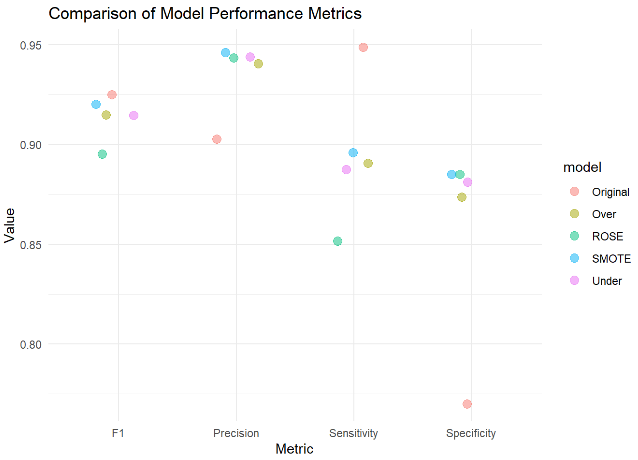


Figure 6 Comparison of Imbalanced Methods-2

When all metrics are evaluated, the smote technique is the best solution for the imbalanced problem.

F. Modelling

After the imputation of data, new data were created, but since the data had an imbalance problem, we applied some techniques, and Smote gave us the best result. SMOTE (Synthetic et al.) is an oversampling process that enables synthetic data to be produced. Before building the models, the data set was divided into 80% train and 20% test for risk prediction. All models were built by repeated cross-validation using the caret package and applying the SMOTE sampling technique.

1. Logistic Regression

Logistic regression is a classification algorithm, although it includes "regression" in its name. It was often used in binary classification problems. This study used the caret package for logistic regression, and the model was created.

Significant variables were determined by applying the ANOVA test.

Coefficients:				
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.67188	0.11119	6.043	1.51e-09
Age	0.22934	0.05393	4.253	2.11e-05
BpmedsYes	1.76917	0.41205	4.294	1.76e-05
Sysbp	2.68510	0.11718	22.915	< 2e-16
Diabp	1.09745	0.09086	12.079	< 2e-16
Bmi	0.26250	0.05481	4.789	1.67e-06
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
(Dispersion parameter for binomial family taken to be 1)				
Null deviance: 6501.7 on 4689 degrees of freedom				
Residual deviance: 2790.8 on 4684 degrees of freedom				
AIC: 2802.8				
Number of Fisher Scoring iterations: 9				

Table 4 Result of Logistic Regression

It is seen that all variables were significant with the ANOVA test. For the final model, we calculated deviation with the with() function and then calculated degrees of freedom for the chi-square test with the same function. Since we have a large deviation statistic(3710.918), we can conclude that all covariates significantly contribute to the model and that the model fit is good. Also, A multicollinearity test was performed, and all vif values < 10, so there is no multicollinearity problem.

Analysis of Deviance Table					
Model: binomial, link: logit					
Response: .outcome					
Terms added sequentially (first to last)					
	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			4689	6501.7	
Age	1	578.00	4688	5923.7	< 2.2e-16
BpmedsYes	1	333.70	4687	5590.0	< 2.2e-16
Sysbp	1	2576.84	4686	3013.2	< 2.2e-16
Diabp	1	198.81	4685	2814.4	< 2.2e-16
Bmi	1	23.56	4684	2790.8	1.21e-06
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

Table 5 Result of ANOVA

When we look at the model, all variables positively affect high risk hypertension. When one unit increases in age, log odds increase by 0.223. if Bpmed is yes, log odds increase by 1.76917. Also, one unit of increase of Sysbp causes a 14.664 change in the odds ratio of high risk, and one unit of increase of Diabp causes a 2.997 change in the odds ratio of high risk. Lastly, when one unit increases of Bmi, log-odds increase by 0.2625

2. *Support Vector Machine*

Support Vector Machines (SVM) are the separation and classification of points on the plane by a line or hyperplane. Before building the model, one hot encoding was applied to categorical variables, and the smote technique was used for the imbalance problem. svmradial has been selected as the SVM type. Moreover, sigma and cost are tuned as 0.014 and 4.921, respectively. These are the parameters that give the best performance.

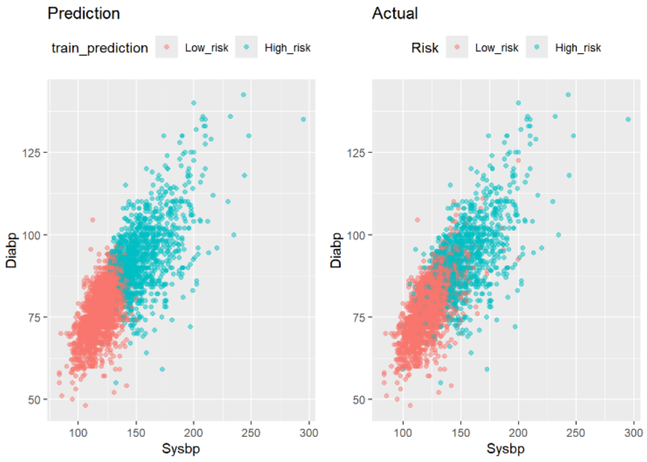


Figure 7 Plot of Sysbp vs Diabp

When you look at both graphs, individuals with high Sysbp and Diabp are in the high-risk group, and individuals with low Sysbp and Diabp are in the low-risk group. The similarity of these two graphs shows that the model works with high accuracy.

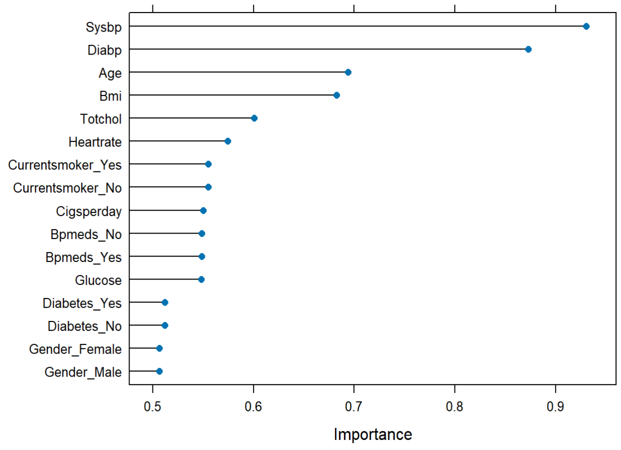


Figure 8 Variable Importance of SVM

Encoding before building the model made significant contributions to the model's performance for some categorical variables that seemed unimportant in other models. When we look at the graph, encoding is essential for the model in the variables applied, although not as much as sysbp and diabp

3. *Decision Tree*

A *decision tree* is a diagram that identifies solutions to a particular problem and visualizes these solutions. It divides large amounts of records into small record groups by following them step by step, and the graph looks like a tree. While building the model, rpart2 was chosen as the method, and the smote technique was used for the imbalance problem. The model is built, and the "tunelength" and "cp"(complexity parameter) have been tuned.

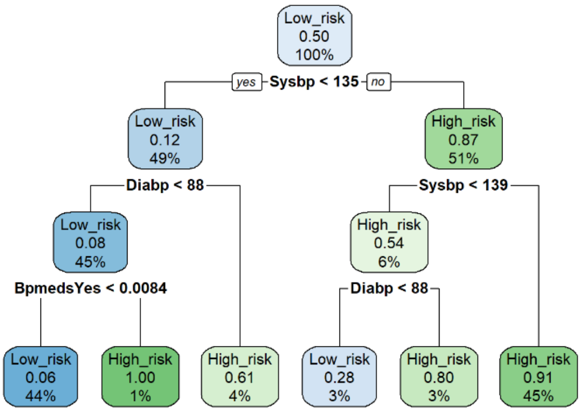


Figure 9 Interaction Plot of Decision Tree

Firstly, if Sysbp is less than 135, these individuals constitute 49% of the total population, and the probability of being in low-risk status is 0.12. When we move on to the lower branch if the Diabp is less than 88, the individuals here constitute 45% of the total population, and the probability of being in low-risk status is 0.08. Let us consider the opposite in the same branch; if the Diabp is greater than 88, the individuals here constitute 4% of the total population, and the probability of being in high-risk status is 0.61. Secondly, if Sysbp is greater than 135, the individuals here constitute 51% of all individuals, and the probability of being in the high-risk category is 0.87. When we move to a lower branch if the Sysbp is greater than 139, the individuals here constitute 45% of all individuals, and the probability of being in the high-risk category is 0.91. Let us consider the opposite in the same branch; if Sysbp is less than 139, the individuals here constitute 6% of the total population, and the probability of being in the high-risk category is 0.54.

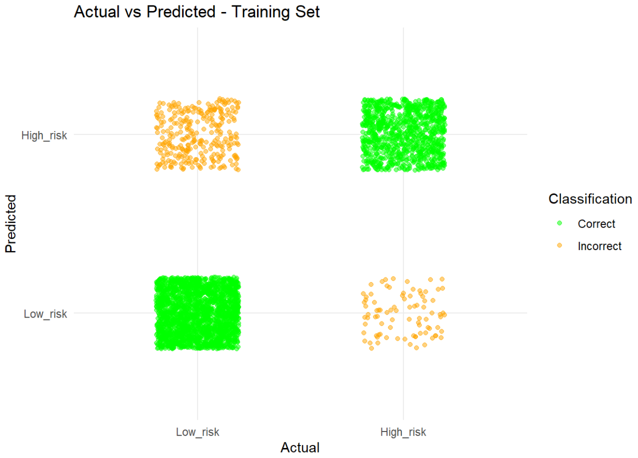


Figure 10 Train Fitting Plot for Decision Tree

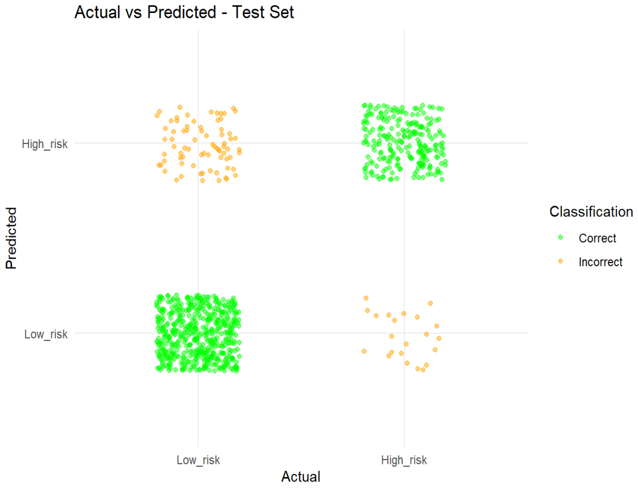


Figure 11 Test Fitting Plot for Decision Tree

While orange dots in the graphs represent misclassification, green dots represent correct classification. In both models, there are more green dots than orange dots. Furthermore, the distribution of classifications is very similar in both graphs. This indicates that there is no imbalance in the model. As a result, the train and the tests are consistent with each other in both correct and incorrect classification, and the correct classification rate is high in both. So, the model is reliable.

4. *Random Forests*

Random forest is created by combining more than one decision tree. The purpose of this is to obtain more accurate results. However, the biggest problem of the random forest model is overfitting. While building the model, the smote technique was used to solve the imbalance problem under the caret package. ntree, mtry, maxnodes, nodesize have been tuned in the final model to solve the overfitting problem.

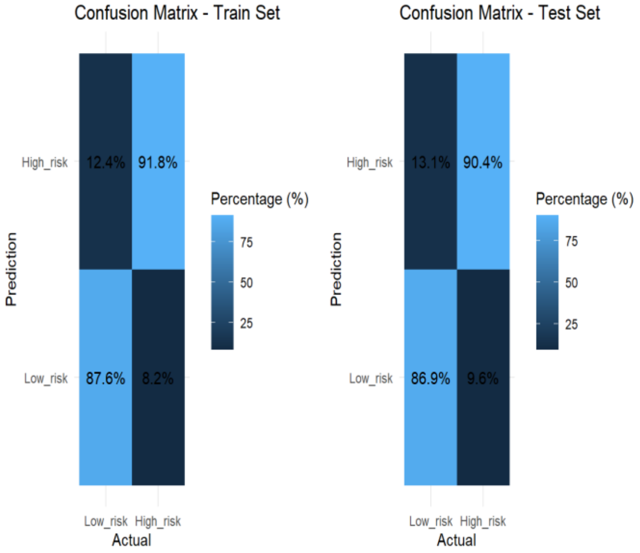
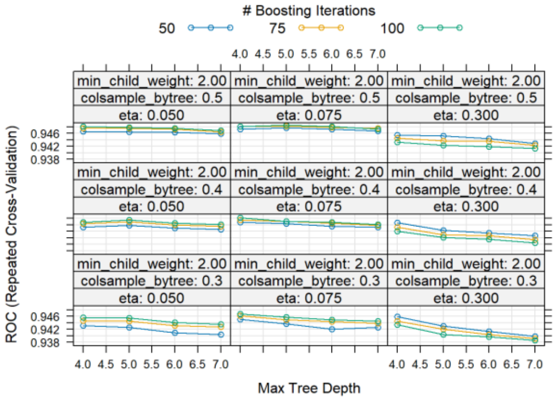


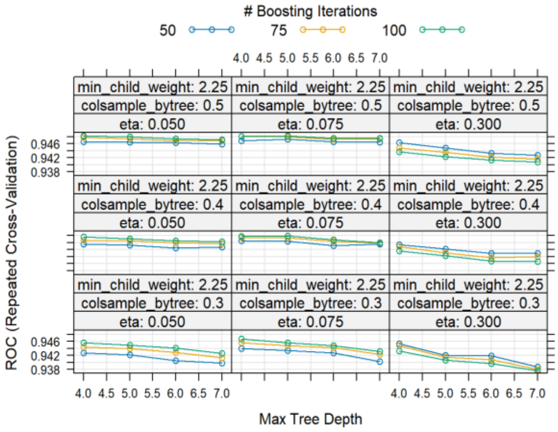
Figure 12 Train and Test Fitting Plots for Random Forest

The percentages in both confusion matrices are very close to each other, so it can be seen that there is no overfitting problem. The accuracy rate is high in both train data and test data. This shows that the model is successful in classification.

5. XgBoost

The XGBoost algorithm adds a new model by reducing the errors of the existing model. It is frequently used in machine learning with its high performance and flexibility. Before building the model, one-hot encoding was applied to categorical variables, and the smote technique was used to solve the imbalance problem in the model. In addition, gamma and subsample were kept constant. The max_depth, eta and gamma parameters are tuned as 5, 0.075, and 0, respectively.





The graphs show the steps followed in the hyperparameter tuning process and the effects of these steps on model performance.

6. Artificial Neural Networks

Artificial neural network imitates the structure of biological neural networks in the brain and their learning, remembering, and generalization abilities. It can be used in classification problems.

One-hot encoding was applied to categorical variables before building the model. Data are scaled in the range of 0-1 with the max-min method. The Smote technique was used for the imbalance problem, the maxit parameter was used to prevent overfitting, and the sigmoid function was used as the activation function of the model. The NN model has two layers, the first of which has 16 neurons, and the second one has 5 neurons. The network uses 91 weights to produce the final output.

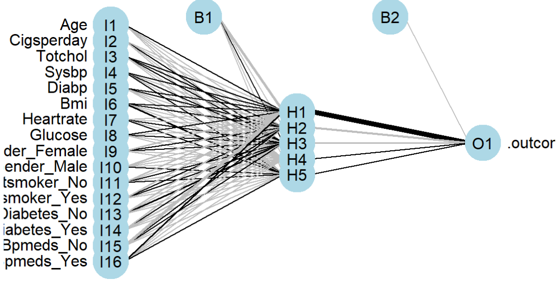


Figure 13 Artificial Neurak Networks Plot

G. Performance Comparison

Logistic regression, support machine vector, decision tree, random forest, xgboost, and artificial neural network models were compared based on their accuracy, sensitivity, specificity, and F1 score.

	Accuracy	**Sensitivity**	**Specificity**	**F Score**
LR	0.8806366	0.884435	0.8721374	0.8186296
SVM	0.8900678	0.8938166	0.8816794	0.8320576
DT	0.8868258	0.8733475	0.9169847	0.8334779
RF	0.8891836	0.8763326	0.9179389	0.8365217
XGB	**0.9257294**	**0.9296375**	**0.9169847**	**0.8840846**
NN	0.8900678	0.8878465	0.8950382	0.8341485

Table 6 Performance Measures of Models on the Train Set

	Accuracy	**Sensitivity**	**Specificity**	**F Score**
LR	0.8913813	0.9010239	0.8697318	0.8315018
SVM	0.8819362	0.8890785	0.8659004	0.8188406
DT	0.8854782	0.8720137	0.9157088	0.8313043
RF	0.879575	0.8686007	0.9042146	0.8222997
XGB	**0.8984652**	**0.9044369**	**0.8850575**	**0.8430657**
NN	0.8913813	0.8959044	0.8812261	0.8333333

Table 7 Performance Measures of Models on the Test Set

When all performance models are evaluated, it is seen that there is no overfitting or underfitting problem in any of them. Also, the performance metrics are all very close to each other, both in the train and test sets. However, the xgboost model is the most successful model in predicting risk status, although there is little difference.

IV. CONCLUSION

The primary purpose of this study was to estimate hypertension risk status, and this prediction was made based on clinical variables. First, exploratory data analysis was performed to understand the data better and to discover the medical variables used, and these were tested with various statistical methods. Meaningful inferences were made from these tests. These inferences were very useful to us when building the models. In addition, since there was an imbalance problem in our data, we tried various techniques and decided that the most suitable technique was smote, and the models started to be built. Various optimizations were made while building the models, and each model showed us the results without any problems. Finally, by comparing the models according to various performance metrics, it was shown that xgboost was the best model for predicting the risk of hypertension.

V. REFERENCES

[1] Zhao, H., Zhang, X., Xu, Y., Gao, L., Ma, Z., Sun, Y., & Wang, W. (2021). Predicting the risk of hypertension based on several Easy-to-Collect risk factors: a machine learning method. *Frontiers in Public Health*, *9*. https://doi.org/10.3389/fpubh.2021.619429

[2] Du, J., Chang, X., Ye, C., Zeng, Y., Yang, S., Wu, S., & Li, L. (2023). Developing a hypertension visualization risk prediction system utilizing machine learning and health check-up data. *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-023-46281-y

[3] Islam, M. M., Rahman, M. J., Roy, D. C., Tawabunnahar, M., Jahan, R., Ahmed, N., & Maniruzzaman, M. (2021). Machine learning algorithm for characterizing risks of hypertension, at an early stage in Bangladesh. *Diabetes & Metabolic Syndrome*, *15*(3), 877–884. https://doi.org/10.1016/j.dsx.2021.03.035