Diabetes Prediction

*Abstract*—

Diabetes, a common disease, was investigated in this project.

The main purpose of this project is to understand the relationship between the outcome variable (dependent variable) that show a person diabetes or not, and independent variables. These are pregnancies, gender, blood pressure, skin thickness, BMI, age, daily calorie intake, frequency of doing exercise in a day, sleep duration in a day, amount of insulin and glucose in the blood. In other words, purpose is to investigate the independent variables that may or may not affect diabetes, and also, is to investigate the effect levels of independent factors. For this purpose, some statistical tests and machine learning method was used such as Logistic regression, Neural Network, Support Vector Machine, Random Forest and Xgboost. Before starting the modeling, data was made convenient. Then, the data was better understood with some research questions .After that, by handling missingness in data set, models are established and model performances are compared.

Keywords— Artificial Neural Network, Logistic Regression, XGBoost, SVM, Random Forest

# INTRODUCTION

Diabetes mellitus is a disease that occurs as a result of a deficiency in the amount of insulin in a person's body. There are some symptoms to watch out for in order to understand whether a person has diabetes or not. Some of the symptoms are the change in skin thickness, the amount of sugar in the blood.. Etc. Variables such as pregnancies, gender, blood pressure, skin thickness, BMI, age, daily calorie intake, daily exercise frequency, daily sleep duration, insulin and glucose amount in the blood were taken, and which variables were more important in the diagnosis of diabetes were investigated. Using the machine learning algorithms specified in the abstract, prediction models were created and the performances of these models were compared using appropriate techniques, such as accuracy, sensitivity and specificity.

# LITERATURE REVIEW

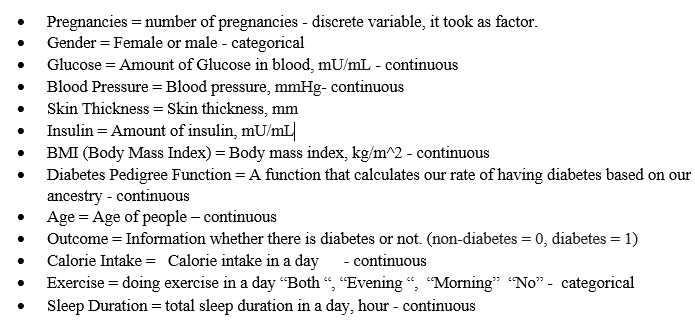
According to the estimates of International Diabetes Federation (IDF) 2021, approximately 10% of adults have diabetes. (International Diabetes Federation, 2021). Logistic regression and Neural Networks are the most commonly used methods for making predictions. Prediction models with high accuracy rate are created in these models.

# METHODOLOGY

1. *Dataset*

The data set taken from Kaggle website:

<https://www.kaggle.com/datasets/dslearner0406/diabetes-dataset> . In this data set, expressed the status of diabetes as binary outcome (diabetes and non-diabetes). In this dataset, there exists 13 columns and 768 observations. In this project, outcome (Information whether there is diabetes or not) variable is used as the response variable. Overall, there exists 3 categorical variables, also Pregnancies coded as categorical, since it has small range. Rest of the variables are continuous.



1. *Descriptive Statistics*

Descriptive statistics numerical of variables:

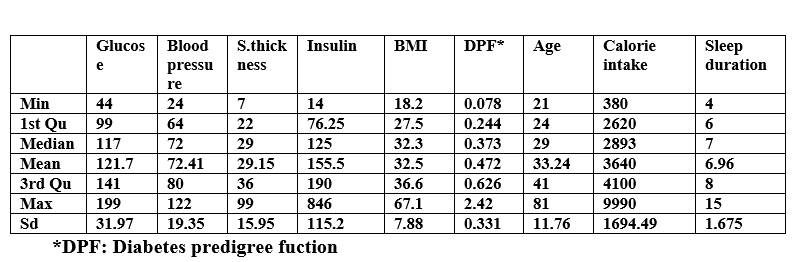


Table 1: Summary statistics of Numerical Variables

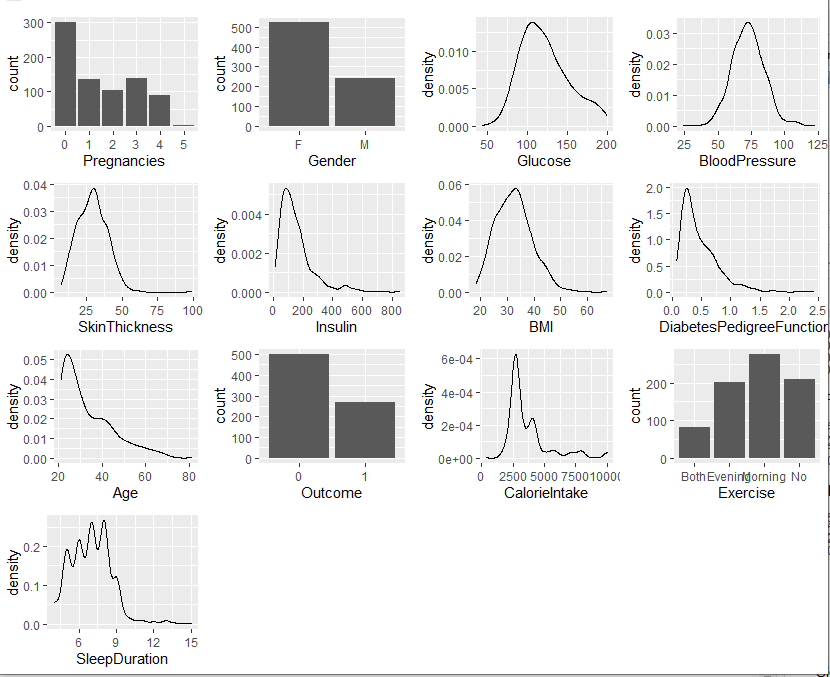
We can say that he distribution of Glucose, Blood Pressure, BIM almost normal, while Skin Thickness, Insulin and Diabetes Pedigree Function show right skewed distribution. Also, Age variable chance between 21 and 81. Mean and median of age almost equal. While Insulin and Calorie intake have high standard deviations, DPF and Sleep duration have small standard deviations.

|  |  |  |  |
| --- | --- | --- | --- |
| **Pregnancies** | **Gender** | **Outcome** | **Exercise** |
| **0:301** | **F: 525** | **0:500** | **Both:82** |
| **1:135** | **M:243** | **1:268** | **Evening:201** |
| **2:103** |  |  | **Morning:276** |
| **3:138** |  |  | **No:209** |
| **4:90** |  |  |  |
| **5:1** |  |  |  |

Table 2: Summary Statistical of Categorical Variables

According to summary of the categorical variable,

The number of women is more than twice the number of men; hence it is not a balanced dataset for both genders. The outcome variable is also not balanced by whether or not they have diabetes since the number of people without diabetes is almost twice as high as those with diabetes. Moreover, in this data set most of the people doing sport in the morning while very few people do sports both in the morning and in the evening.



As can be seen from the above graph that Glucose, Blood pressure almost normal, while Skin thickness, Insulin, BMI, DPF, Age and Calorie intake almost right skewed.

1. *Exploratory and Confirmatory Analysis*

*C.1* What is the average amount of glucose of genders?

The amount of glucose in the blood is an important factor that helps us to understand whether a person has diabetes. In this research question, it has been investigated whether this factor varies according to gender.

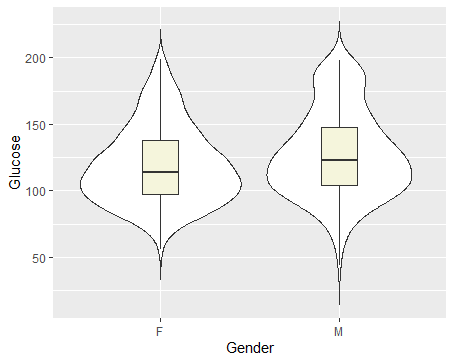


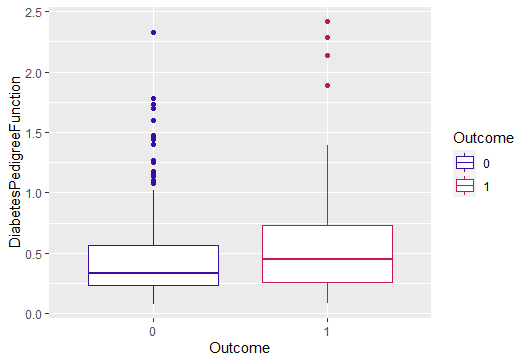
Figure 1 Violin Plot Gender vs Glucose

Violin plots shows that median of glucose differ for genders, although there is no big differences.

Although transformation was attempted, the normality assumption could not be met, hence Wilcoxon rank sum test was used to test the research question. It was concluded that glucose amounts differ for both genders (the p-value of Wilcoxon rank sum test less than alpha).

*C.2* What is the average result of Diabetes Pedigree Function of outcome?

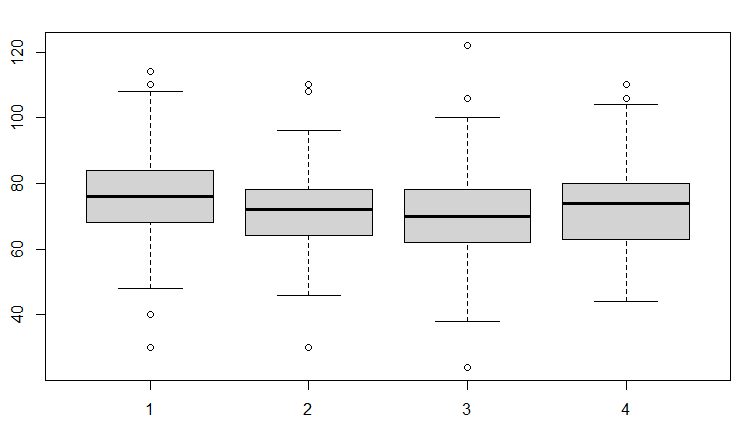
|  |  |  |  |
| --- | --- | --- | --- |
| **Outcome** | **count** | **mean** | **sd** |
| **0** | **500** | **0.430** | **0.299** |
| **1** | **268** | **0.550** | **0.372** |



*Figure 2 Box Plot Table3: Groups information*

Diabetes Pedigree Function is a function produced by using different variables used when diagnosing diabetes. Therefore, we can say that Diabetes Pedigree Function rate is directly related to the state of being diabetic. In order to be sure Wilcoxon rank sum test was used, since the assumption did not met. It was concluded that diabetic patients had a higher Diabetes Pedigree Function rate.

*C.3* Is the average blood pressure of people under each exercise categories ("no", "morning", "evening", "both") is equal to each other?



*Figure 3 Box Plot of Exercise Categories vs Blood pressure*

metin içeren bir resim

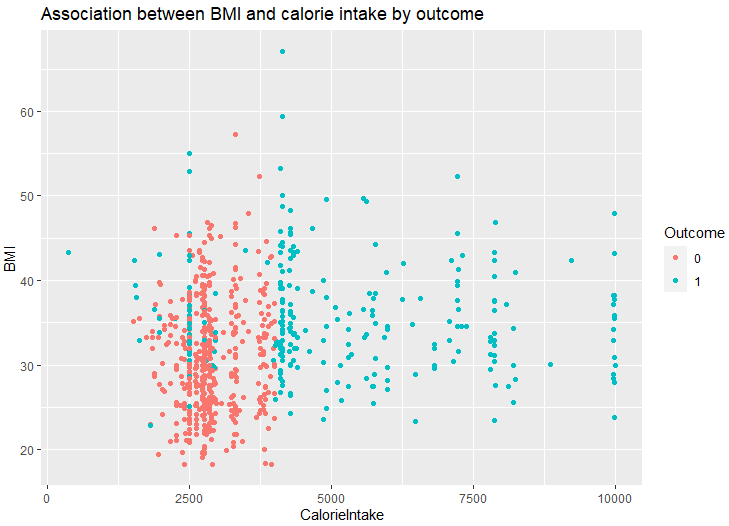
Açıklama otomatik olarak oluşturuldu

*Figure 4: Pairwise Comparison*

For the figure3, group1 to group4 represent "No" , "Evening" ,"Morning", "Both" respectively. ANOVA test was used for this research question.

Before the conducting ANOVA test assumption was checked. According to result of ANOVA test, the null hypothesis can be rejected since, the p value is less than 0.05. It indicates that some of the group means are different. In order to understand comparison, the multiple pairwise (TukeyHSD ) test was used. There is a difference between No-Evening and No-Mornings. It can be seen that the confidence interval for their mean difference does not include 0.

*C.4* What is the association between BMI and calorie intake by outcome?



*Figure 5: Bubble Plot of Calorie Intake vs BMI*

According to figure5, there is no significant association between calorie intake and BMI. In addition, the daily calorie intake of non-diabetic patients(Outcome=0) is less than those with diabetes(Outcome=1). For this research question Hotelling's T2 test was used.

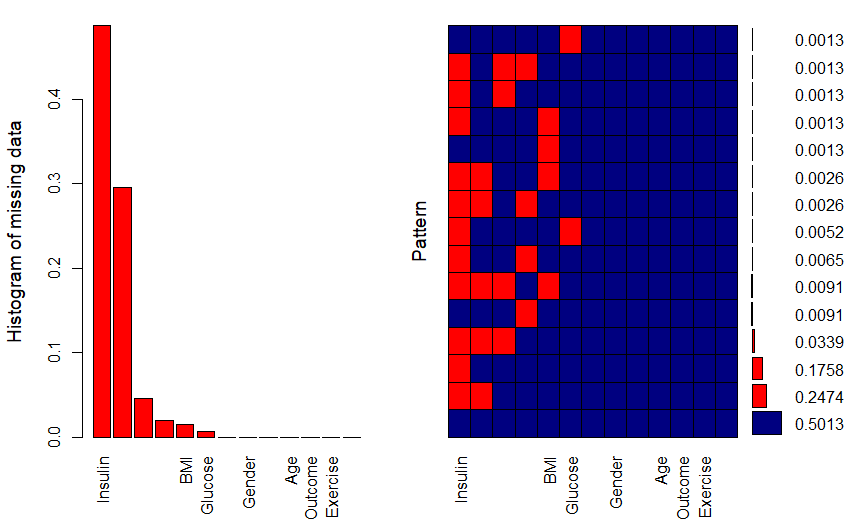
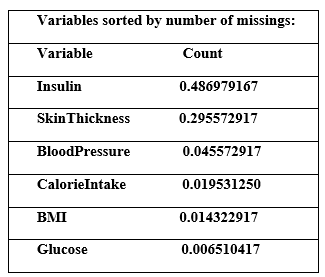
|  |  |  |  |
| --- | --- | --- | --- |
| Hotelling's two sample T2-test | | | |
| T.2 = 3461 | df1 = 1 | df2 = 1534 | p-value < 0.05 |

*Table4: Result of Hotelling’s T2 test*

The results indicated that T2 = 3461, with 2 and 1534 degree of freedom and p = 2.2e-16 which is less than alpha. The null hypothesis, which is no group mean difference, is rejected. Hence the result is that the alternative hypothesis is true location difference is not equal to c (0,0).

1. *Missingness*

In the diabetes dataset, there are many completely at random missing values. In order to handle this missingness imputation methods (“mice” packages in R) was used.

*Figure 6: Aggregation Graph of Missing Values*

The table show that almost 0.49 % of Insulin and 0.30% of SkinThickness are missing. In order to impute missing values, mice package was used. Descriptive statistics were compared to see if there were any changes in the characteristics of the data after the imputation. After comparison, it is concluded that the descriptive statistics of imputed dataset and original dataset are so close to each other.

## Modelling

Before applying the statistical models, the data set was divided into two groups as test and train. 80% were used in the train set and 20% in the test set. In addition, the continuous variables used was scaled.

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| Train data | 0.6504 | 0.349 |
| Test data | 0.6534 | 0.346 |

*Table5: Proportion of outcome*

According to table5, the proportion of the levels of the response are close to each other

## Logistic Regression

Let’s look at the correlation of continuous variables:

metin, makbuz içeren bir resim

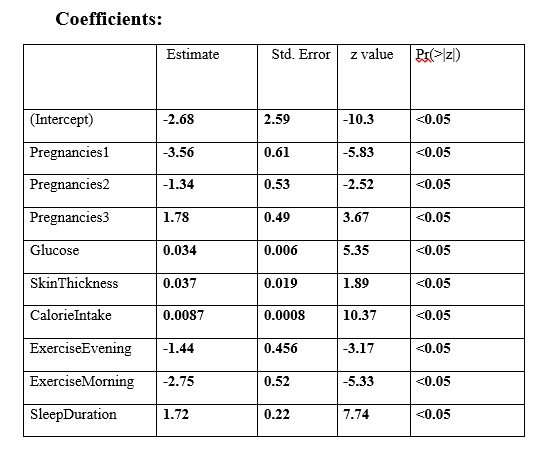
Açıklama otomatik olarak oluşturuldu

*Table6: Correlation Table*

|  |  |
| --- | --- |
| 0 | 1 |
| Before | 400 | 215 |
| After | 860 | 645 |

The output shows that Insulin- Glucose and BMI-Skin Thickness are correlated. In order to avoid the existence of the multicollinearity, one of the related variables are included in the model. Since it is an imbalanced data, a balanced data was obtained with the smote function.

First, it started with the full model, then a new model was established with all the variables then, using the significant ones. In addition, it was compared between the full model and the reduced model by performing a chi-square test, so the result is that reduced model can be used instead full model.



*Table7: Coefficient Table of Model*

The coefficient estimate of glucose variable is 0.34. Hence, increase in glucose is related with rise in the probability of being diabetes. Also, when we look at the coefficient, it can be understood that doing sport has negative relation with being diabetes, while no exercise has positive association with being diabetes. Also, in the exploratory and confirmatory analysis part, I get the same result.

To better interpretation, we can look at odds ratios.

|  |  |  |  |
| --- | --- | --- | --- |
| **(Intercept)** | **Pregnancies1** | **Pregnancies2** | **Pregnancies3** |
| **2.4e-12** | **2.83e-02** | **2.63e-01** | **5.98** |
| **Glucose** | **Skin**  **Thickness** | **Calorie**  **Intake** | **Exercise**  **Evening** |
| **1.03** | **1.038** | **1.003** | **0.236** |

*Table8: Odds ratios*

When one unit increase in glucose, the odds of being diabetes (outcome =1) increases by 1.03.

When one unit increase in CalorieIntake, the odds of being diabetes (outcome =1) increases by 1.003.

When one unit increase in SleepDuration, the odds of being diabetes (outcome =1) increases by 5.54.

Lastly, model performance was checked after obtaining optimal cut of point (0.05).To find best cut of point optimal cutoff function is used in R.

1. *Support Vector Machine (SVM)*

Support Vector Machine (SVM) can be used both classification and regression. In this project, I used SVM for creating prediction model for outcome variable (Diabetes-non diabetes). Tuning the parameters, I found the best performance from the model. Then, "svmRadial" was used as method, and model constructed. Lastly, predictions were made.

1. *Artificial Neural Networks (ANN)*

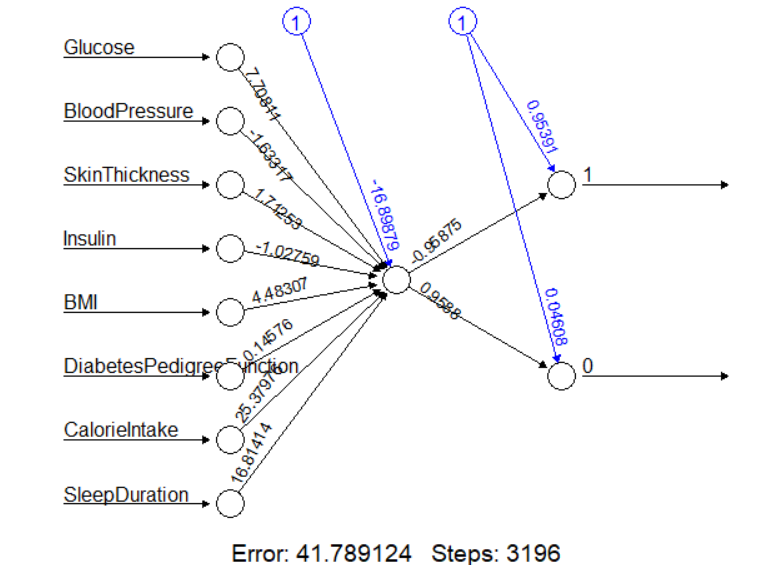
**

Figure 7: Neural Network

Again, for this machine learning algorithm, the main goal is to discover a model which predicts. Since the response is binary, the model is set as act.fct="logistic".

I made a classification, hence I sated the linear. output as FALSE and the err.fct to “ce” ,which for classification. In ANN model, AIC is equal to 16.899, and BIC is equal to 7.708.Also, cross-entropy error is 41.79 which can be seen from the figure7.

1. *Random Forests (RH)*

The number of trees is 500 in the model and no. of variable tried at each split are 2(It can be seen from the Figure8). While Classification error in nondiabetic (0) is 0.004, which is approximately 0.4%, diabetes (1) is 0.09 which is almost 9%.

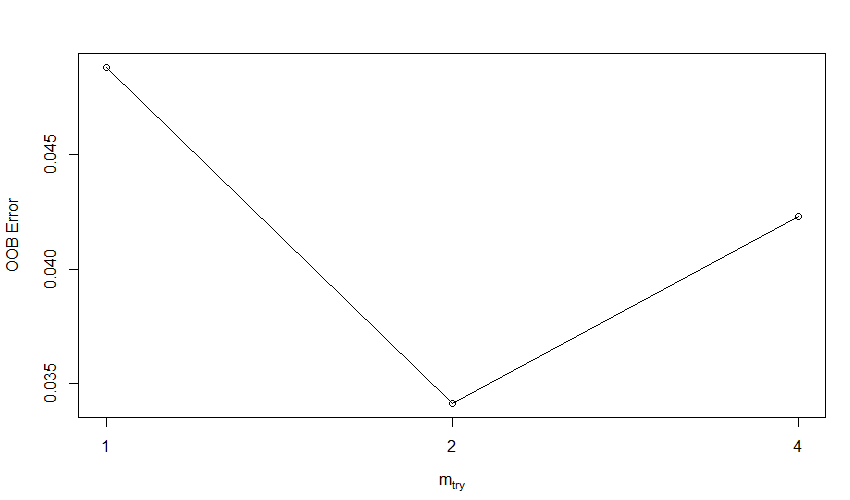


Figure 8: Neural Network

1. *XgBoost*

As in other algorithms, the most suitable model parameters are determined in the XGBoost classification. Also, optimal through is found to with simple cross-validation.

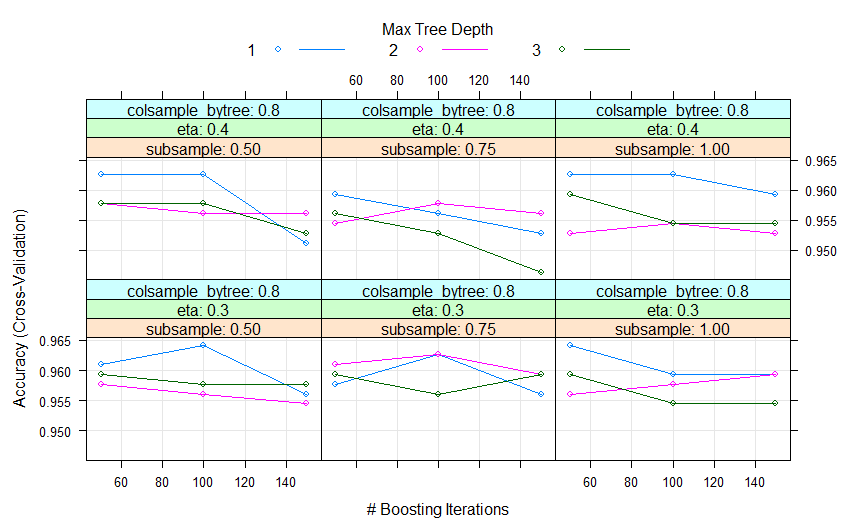


Figure 9: XGBoost Plot

For this classification method used to the following parameters, max\_depth = 1, eta = 0.3, gamma = 0, colsample\_bytree = 0.6, min\_child\_weight = 1 and subsample = 0.75.

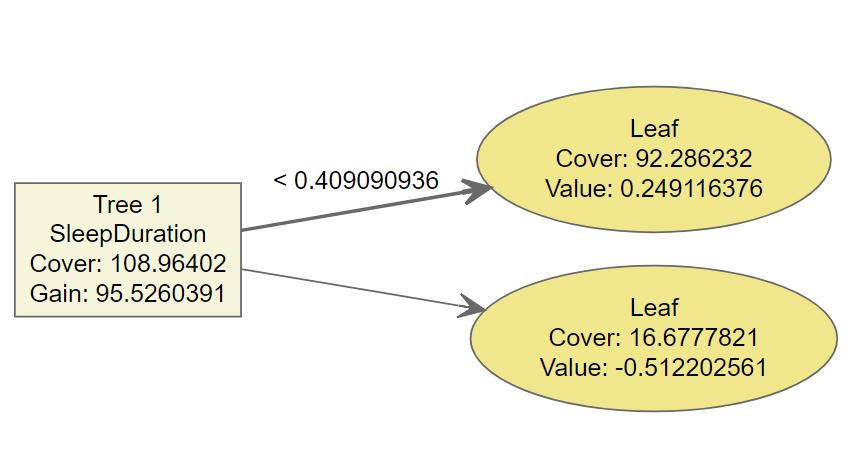


Figure 10: XGBoost Classification Plot

1. *Performances Comparison Formulas:*
2. *Classification Error Rate*

|  |  |
| --- | --- |
|  | (1) |

1. *Accuracy, Sensitivity and Specificity*

|  |  |
| --- | --- |
|  | (2) |
|  | (3) |

|  |  |
| --- | --- |
|  | (4) |

# RESULTS

We have set up the models with Logistic Regression Model, Support Vector Machine (SVM), Artificial Neural Network (ANN), Random Forests (RF) and XgBoost methods respectively, in this part model performances will be compared.

|  |  |  |  |
| --- | --- | --- | --- |
|  | ACC | SN | SP |
| LR | 0.96 | 0.98 | 0.99 |
| SVM | 0.92 | 0.89 | 0.94 |
| ANN | 0.8767 | 0.81 | 0.94 |
| RF | 0.79 | 0.74 | 0.85 |
| XGBoost | 0.78 | 0.74 | 0.83 |

*Table8: Performance Comparison of Models*

According to model results, LG is the model with the highest values. Also, as can be seen from the table, XGBoost has the lowest values.

# CONCLUSION

In this study, firstly, the structure of the data was checked and the exploratory data analysis (EDA) part was performed. Then the results obtained with statistical tests (EDA) were also tested. Then, within the model lot, Logistic Regression Model, Support Vector Machine (SVM), Artificial Neural Network (ANN), Random Forests (RF) and XgBoost algorithms were used, respectively. The most suitable parameters were found for each model in the machine learning models used. Finally, by comparing the performance measures, it was understood which model made the best prediction.

# REFERENCES

|  |  |
| --- | --- |
| S.YOLOĞLU ,Tip2 Diyabet Mellitus ile Risk Faktörleri Saptamada çok Değişkenli Yöntemlerin Karşılaştırılması  http://abakus.inonu.edu.tr/xmlui/bitstream/handle/11616/9668/10225376.pdf?sequence=1&isAllowed=y, 2018. |  |
| Y. ÖZKAN, B. SARER YÜREKLİ, A.SUNER,Diyabet tanısının tahminlenmesinde denetimli makine öğrenme algoritmalarının  performans karşılaştırması, https://dergipark.org.tr/tr/download/article-file/1379253 |  |
| <https://www.kaggle.com/datasets/dslearner0406/diabetes-dataset> |  |
|  |  |
|  |  |
|  |  |