Diabetes Prediction

Final project for Data Mining course

ABSTRACT

The leading cause of death in the modern world is diabetes.
In many situations the patient does not even know that he is at risk until he has a seizure, so we would like to predict If the patient is at increased risk for diabetes and thus help prevent a heart attack or stroke. We would like to investigate which patients may suffer from diabetes in the near future by using the medical data of patients and analysis of these data. Our goal is to prevent deterioration in patients' health and even save lives.
In this project, we will achieve the predicted results by applying various methods on our raw data, starting from cleaning the data, through its classification into subgroups, and activating data mining prediction algorithms.

INTRODUCTION

We would like to predict which patients will be at high risk of developing diabetes based on various medical data. Diabetics may suffer irreversible damage to the brain and heart muscle, which requires a change in lifestyle, as they are at increased risk for heart attacks and other strokes. However, this is a reversible process if it is treated in time. Early detection of diabetes will prevent surgeries and hospitalizations and thus will also save budgets from public funding.
In addition, during the corona period the demand for hospital beds increased and the prevention of heart attacks and strokes would save unnecessary hospitalizations.
In this project we intend to analyze and create a model on our dataset to predict if a particular observation is at a risk of developing diabetes, given the independent factors.
This prediction contains the methods followed to create a suitable model, including EDA along with the model.

BACKGROUND

Diabetes is a disease that occurs when your blood glucose, also called blood sugar, is too high. Blood glucose is your main source of energy and comes from the food you eat. Insulin, a hormone made by the pancreas, helps glucose from food get into your cells to be used for energy. Sometimes your body doesn’t make enough — or any — insulin or doesn’t use insulin well. Glucose then stays in your blood and doesn’t reach your cells.
Diabetes affect many people worldwide and is normally divided into Type 1 and Type 2 diabetes. Both have different characteristics.
According to WHO about 422 million people worldwide have diabetes. Since diabetes affects a large population across the globe and the collection of these datasets is a continuous process and it comprises of various patient related attributes such as age, gender, symptoms, insulin levels, blood pressure, blood glucose levels, weight etc.

MATERIAL AND METHODS

The Data

This dataset is originally from the hospital Frankfurt, Germany, and were taken from Kaggle.
The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old.
The dataset consists of several medical parameters and one dependent (outcome) parameter of binary values. This dataset description is as following
9 columns with 8 independent parameters and one outcome parameter with uniquely identified 2000 observations having 644 positives for diabetes (1) and 1264 negative for diabetes (0).

VARIABLE	TYPE	DESCRIPTION
Pregnancies	Integer	Number of times pregnant
Glucose	Integer	Plasma glucose concentration
BloodPressure	Integer	Diastolic blood pressure (mm Hg)
SkinThickness	Decimal	Triceps skin fold thickness (mm)
Insulin	Decimal	2-Hour serum insulin (mu U/ml)
BMI	Integer	Body mass index (weight in kg/(height in m)^2)
DiabetesPedigreeFunction	Integer	Diabetes pedigree function
Age	Integer	(years)
Outcome	Binary	Class variable (0 or 1)

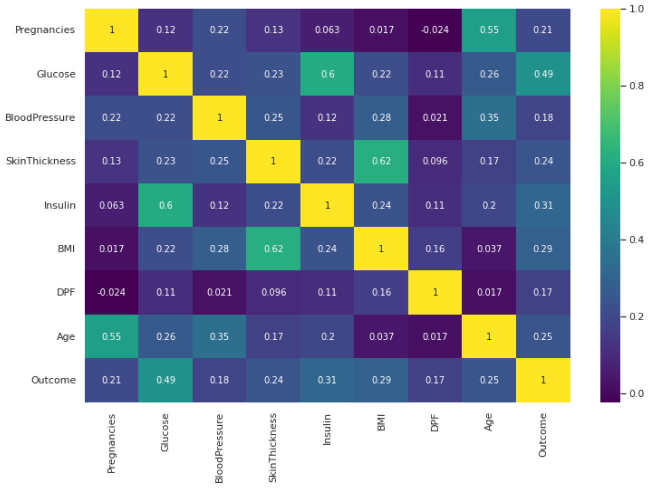
Preprocessing

Data Cleaning

irrelevant columns - The first part of the data cleaning was understanding the meaning and datatypes of the columns. All columns were found to be relevant for our prediction.

Dropping registries - Our data set originally consisted of 2000 registries.
Although in this dataset none of the columns contain missing values, some of the measurements (Glucose, Blood Pressure, Skin Thickness, Insulin and BMI) have values of 0, which is not possible for a living human organism.
Here come to approaches either completely remove missing values or normalize them with mean or median. By using approach one we will lose about 50% of our dataset then our training model will not have much data to be train. By using approach two there are some very crucial parameters as Glucose and Blood Pressure that will affect the most on the outcome. By having these consideration, we used a hybrid approach - remove missing values of the parameters that will affect the most on the outcome and normalize others using mean or median. Now our dataset contains 1908 registries.

EDA & Data reduction

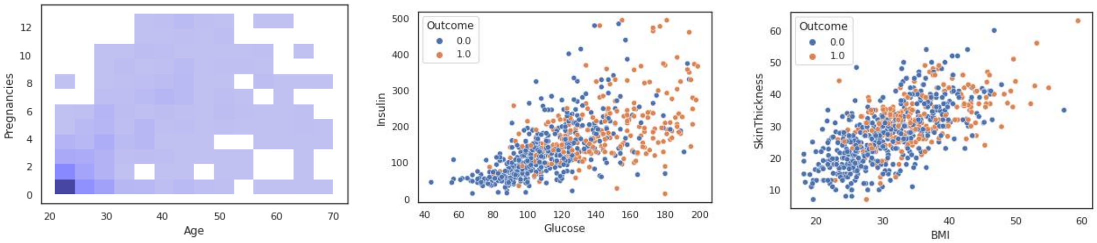


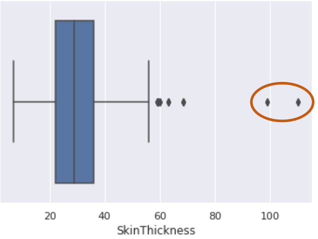
In the data set there are columns that represent tests, so our goal is to check if there is a connection between certain tests in order to reduce costs without compromising the quality of the analysis.

Therefore, we examined column dependence to identify whether there is a relationship between the variables.

The following conclusions can be drawn from the heatmap:

- There is not much correlation among different predictors.
- Age and Pregnancy have a positive corr indicating that adults have more children.
- There is high corr between BMI and SkinThickness, generally people with very high BMI are considered obese thus explaining thicker skin.
- There is positive correlation among Insulin and Glucose as well which could be explained by the fact that perhaps the type 1 diabetic patients who generally have high Glucose , were given Insulin injections.
- Glucose also has corr with our Outcome - diabetic patients have higher level of glucose in their blood.





Although there are columns with a relatively high correlation, in order to waive certain medical tests, we will require a correlation of at least 0.75, so we will not remove any medical tests.

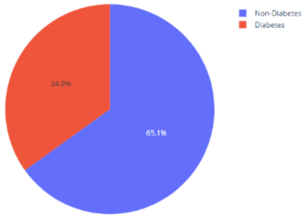
We dropped all data that was beyond the whiskers, meaning the outliers.

For example, SkinThickness above 99 was removed as outlier because it was too far from the norm. we removed 18 outliers and we were left with registries.

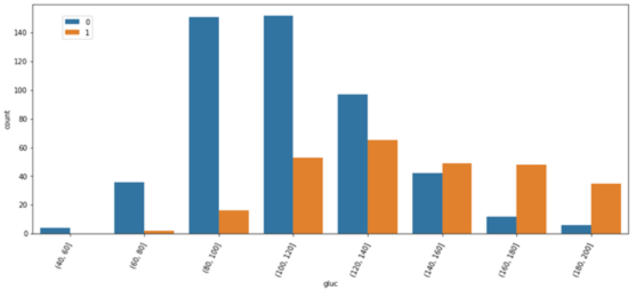
EDA & Data transformation

To get general insights on our dataset we made some comparisons.

The dataset seems to be imbalanced with 65% of input as non-diabetic, we will use SMOTE to deal with this later as we train our model.

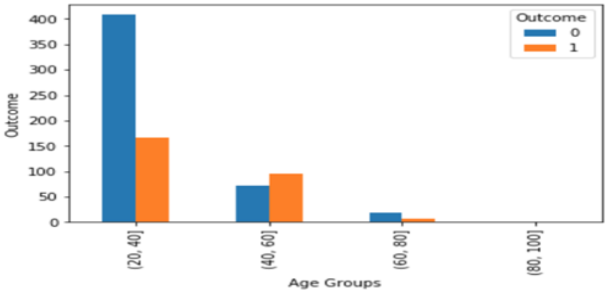


We made a comparison between outcome 0 (non-diabetic) and 1 (diabetic) over the number of patients and their different levels of glucose (range from 40 to 200):



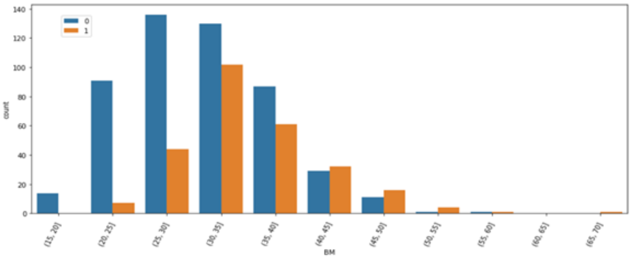
the highest number of patients with the outcome 0 is displayed between the range of 80 to 120, whereas 120 to 140 range of glucose has the highest count of patients with the outcome of 1.

We made a comparison between outcome 0 and 1 over the number of patients and their age after binning (equal-width, range of age groups from 20 to 100 with an interval of 20):



The highest number of outcomes are in the age group 20 to 40 for outcome resulting 1 and 0.

This bar chart illustrates the BMI range varying from 15 to 70 with an interval of 5:



The 25 to 30 BMI group has a maximum count of patients without diabetes whereas the BMI range with the highest count of diabetes is 30 to 35.

Normalization: In order for the variance to be uniform among all the columns, we chose to standardize the data (Z-score normalization).

Algorithms

In order to make a decision about the best algorithm, we used the performance metrics:
Accuracy - failure to detect the disease in patients can cause serious health complications and even death, so in order to reduce the chance of harming the patient, as reliable prediction as possible is required.
Recall - Of all the positives for diabetes, how many of them are the algorithm really classified as such? This index is important because it is a life-saver, it is necessary for the model to predict as many of the positives as such.
We have selected two algorithms for comparison that are suitable for building a classification: K-Nearest Neighbors and Decision Tree.

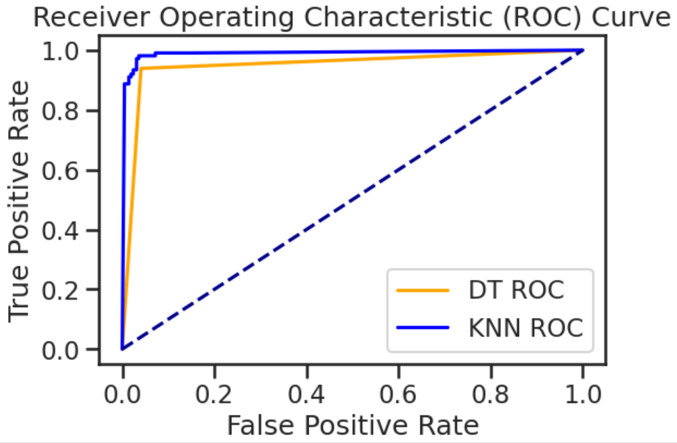
K-Nearest Neighbors - Given the input of a new sample, the algorithm associates it with the final analysis group. This algorithm is suitable for classification problems when they are nominal and numeric. The final classification column is whether the patient has diabetes / no diabetes. From both the recall and the Accuracy score we obtained that k = 23 can be considered as an optimal value.
Decision Tree - A supervised classification model that uses 70% of the data as training set and tests itself on the remaining 30%.

Numerical Analysis and Results

To compare both models results:

Algorithm	Accuracy	Recall	Precision
K-Nearest Neighbors	0.97	0.98	0.94
Decision Tree	0.95	0.94	0.92

Lastly, we performed an ROC comparison between the two models:



Discussion

- As we can see, both of accuracy and Recall scores in KNN algorithm (0.97,0.98) is higher than in Decision Tree algorithm (0.95, 0.94).
- Also, we can see that the ROC curve of KNN algorithm is better than the Decision Tree ROC curve.
- Therefore, we chose the KNN algorithm as the best predictor for our model.

Conclusions

In this project, we wanted to predict which patients will be at high risk of developing diabetes based on various medical data.
we applied various methods on our raw data, starting from cleaning the data, through its classification into subgroups, and activating data mining prediction algorithms – K-Nearest Neighbors and Decision Tree.
According to the metrics that were found the most important in predicting diabetes, we selected the K-Nearest Neighbors as the right model due to high Accuracy and Recall score.
For future work, the same method could be considered and many other machine learning classifiers algorithms could be considered to compare the most accurate one. This method can also be implemented on various other disease and medical datasets.

References

Diabetes Dataset, hospital Frankfurt, Germany
https://www.kaggle.com/johndasilva/diabetes/code?datasetId=23663

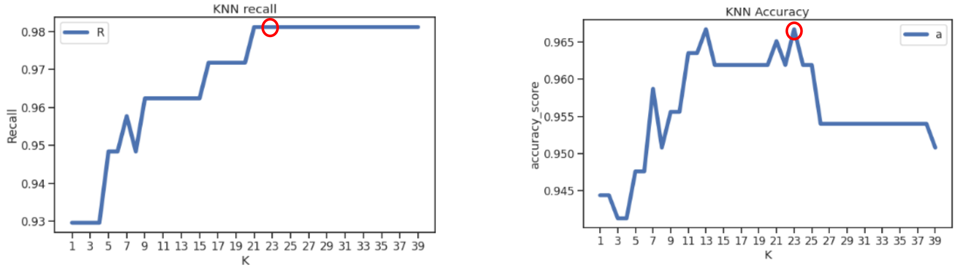
1 Appendix

Group number – **47**

First name	Family name	ID num
Shoval	Cohen	208986489
Yossef	Cohen	308107432

Email address for correspondence:
shoval2465@gmail.com

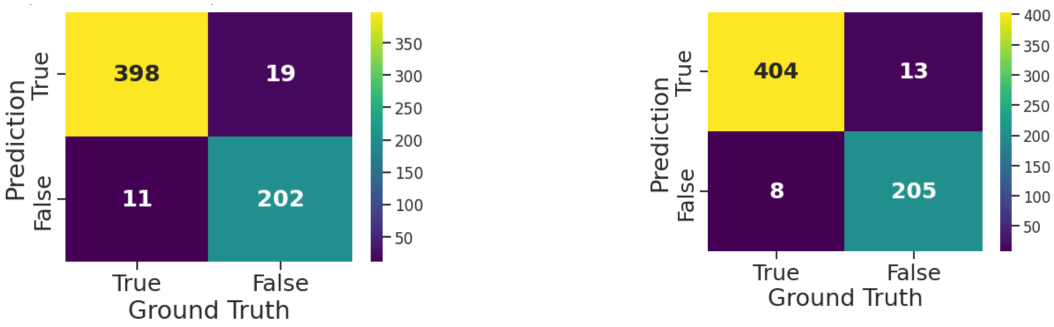
Optimal K value (KNN)



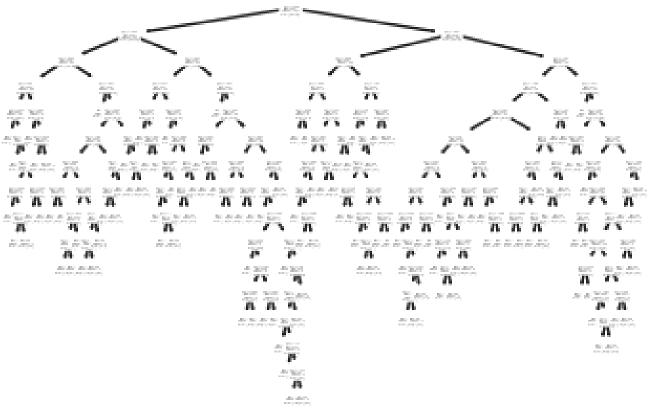
Confusion matrix

Decision tree

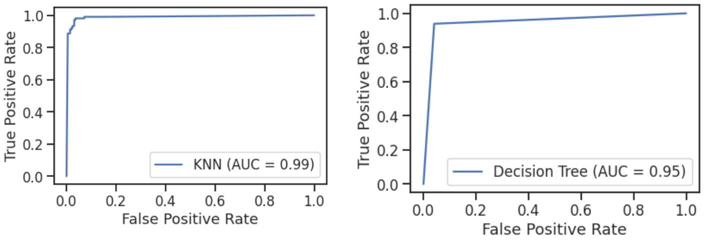
KNN



Decision Tree visualization



Receiver Operating Characteristic (ROC) Curve


True Positive Rate
KNN (AUC = 0.99)
False Positive Rate
True Positive Rate
Decision Tree (AUC = 0.95)
False Positive Rate