Diabetes Prediction Using Machine Learning

DSCI3415 Project Phase II

Mona Ibrahim
Mathematics and Actuarial Science
American University in Cairo
900212749
monamahmoud@aucegypt.edu

Maya Elshweikhy
Mathematics and Actuarial Science
American University in Cairo
900204233
mayahanyy1@aucegypt.edu

ABSTRACT

This paper addresses the global health crisis of diabetes, a chronic illness affecting millions worldwide. If accurate early prediction of the illness is achieved, the risk factor and severity of diabetes can be considerably decreased. This project aims to create a system capable of early diabetes prediction for patients, with increased accuracy through the integration of outcomes from diverse machine learning techniques. Employing algorithms such as K-nearest neighbour, Logistic Regression, Random Forest, Support Vector Machine, and Decision Tree, the model's accuracy is assessed for each algorithm. Subsequently, the algorithm demonstrating notable accuracy is selected as the predictive model for type 2 diabetes. Through cleaning and pre-processing the Pima Indians Diabetes Dataset, this paper explores the current landscape of employing machine learning techniques in predicting type 2 diabetes.

INTRODUCTION

Diabetes is a persistent condition marked by insufficient insulin production or ineffective utilization, posing serious risks. Uncontrolled diabetes often results in hyperglycemia, causing significant harm to the nerves and blood vessels. As of 2014, 8.5% of adults had diabetes, contributing to 1.5 million deaths in 2019, with 48% occurring before age 70. Our project aims to produce accurate early diabetes predicition to avoid its complications.

DATA DESCRIPTION

We chose the Pima Indian Diabetes Dataset to create a machine-learning model that predicts diabetes based on certain diagnostic measurements included in the dataset. This dataset is donated by the National Institute of Diabetes and Digestive and Kidney diseases [1], and collected from the Pima Indian population near Phoenix, Arizona. It contains 768 observations. It has 9 main features.

VARIABLE DESCRIPTION:

Variable	Type	Description
Pregnancies	Quantitative (Numeric)	Number of Pregnancies
Glucose	Quantitative (Numeric)	Plasma Glucose Concentration in 2 hours in an oral glucose tolerance test
BloodPressure	Quantitative (Numeric)	Diastolic blood pressure (mm Hg)
SkinThickness	Quantitative (Numeric)	Triceps skin fold thickness (mm)
Insulin	Quantitative (Numeric)	Insulin level in blood 2-Hour serum insulin (mu U/ml)
BMI	Quantitative	Body Mass Index

	(Numeric)	(weight in kg/(height in m)^2)
DiabetesPedigreeFunction	Quantitative (Numeric)	Diabetes percentage (measures the patient's diabetic family history)
Age	Quantitative (Numeric)	Age in years
Outcome (target)	Categorical (binary)	Class variable (0 or 1) (1: Yes, the individual has diabetes; 0: No, the individual does not have diabetes)

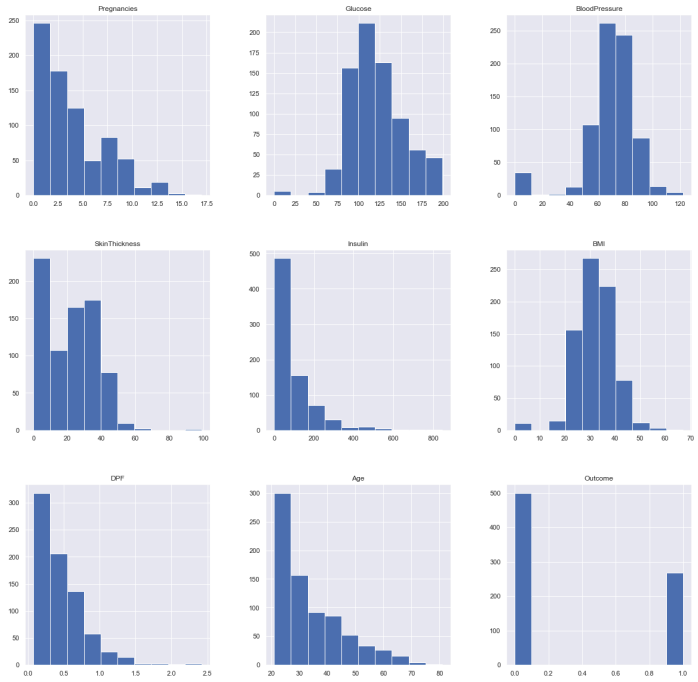


Figure 1. Box Plots of all Variables

We choose this dataset as it is relevant as it shows multiple features that determine if a person develops diabetes, and it has a label variable aiding us in applying machine learning models. In particular, all patients in the dataset are females, at least 21 years old of Pima Indian heritage. We chose it as it contains features that, according to the literature review, have been linked/ related to diabetes. Such features are the the glucose levels, insulin levels, genetic factors (described in the diabetes pedigree function) and the BMI.

ANALYSIS OF DIFFERENT FEATURES

Descriptive statistics were performed to have an initial look at the variables. Glucose, Blood Pressure and BMI graphs are almost symmetric. While Pregnancies, SkinThickness, Insulin, Age, and DPF are positively skewed

The histograms in figure 1 show diverse distributions among the variables in the dataset. Variables such as Glucose, BloodPressure, and BMI exhibit bell-shaped distributions, suggesting a symmetrical spread of values around the mean. On the other hand, variables like Pregnancies, SkinThickness, Insulin, DPF (Diabetes Pedigree Function), and Age display left-skewed distributions, indicating that the majority of values are concentrated towards the lower end. These differences in distribution shapes hint at varying data characteristics and potential insights into the underlying factors influencing these variables.

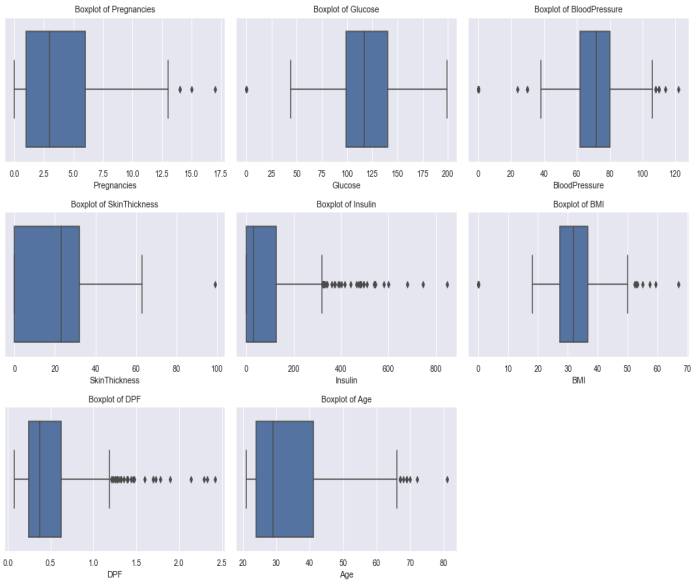
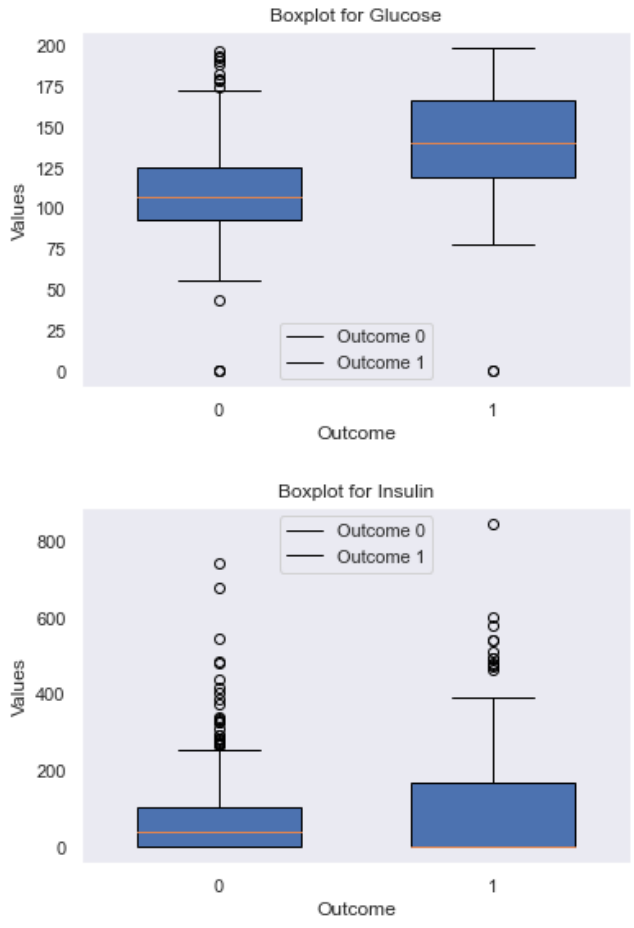


Figure 2. Box Plots of all variables



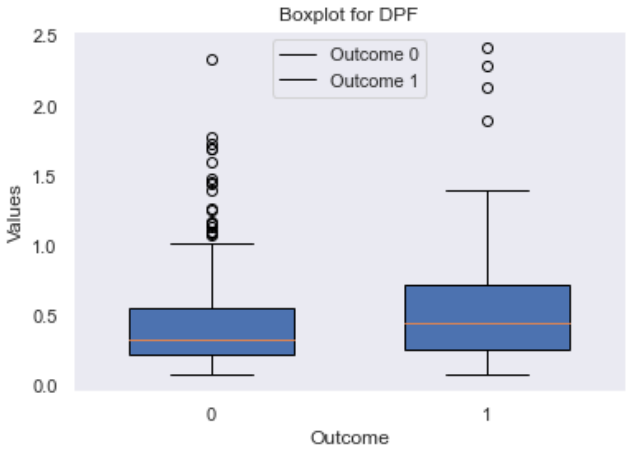


Figure 3. Box Plots of Some Variables in Both Groups (Outcome = 1 and 0).

Boxplots are a standardized way of showing the distribution of data based on a five number summary ("minimum", first quartile (Q1), median, third quartile (Q3), and "maximum")[2]. As shown from the Boxplots there are some outliers in each variable as any points below the minimum or above the maximum are considered outliers as illustrated in Figure 4 below. The minimum is calculated using the following formula (first quartile (Q1) - 1.5* Interquartile Range) and the maximum is calculated using the following formula (third quartile (Q3) + 1.5* Interquartile Range). As shown from the Boxplots grouped by the outcome, there are some outliers for both classes in each variable, but we can see that class with label zero has more outliers and this could be due to that this class has more number of observations than the other class. For example, as shown above, we can see that in the following variables: Glucose, Insulin, and DPF, there are more outliers in the Outcome class = 0.

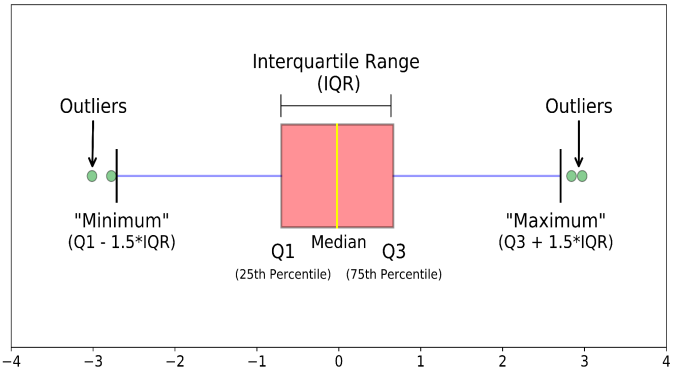


Figure 4. Box plot Interpretation For Outliers Detection [2]

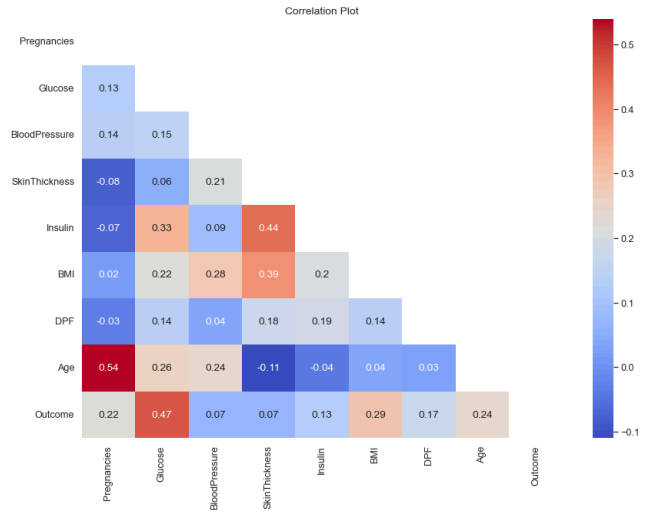


Figure 5. Correlation matrix illustrating the correlation between variables.

Correlation between variables was also calculated to identify if there are dependent variables. As we can see from the below heatmap, some variables correlate positively, such as Age and Pregnancies, with a correlation coefficient of 0.54. Also, other variables correlates positively with the target such as Glucose and BMI with a correlation coefficients of 0.47 and 0.29 respectively. While some variables correlate negatively,such as Age and SkinThickness with correlation coefficient of -0.11, and Pregnacies and SkinThickness with correlation coefficient of -0.08 and both are considered weak negative correlations.

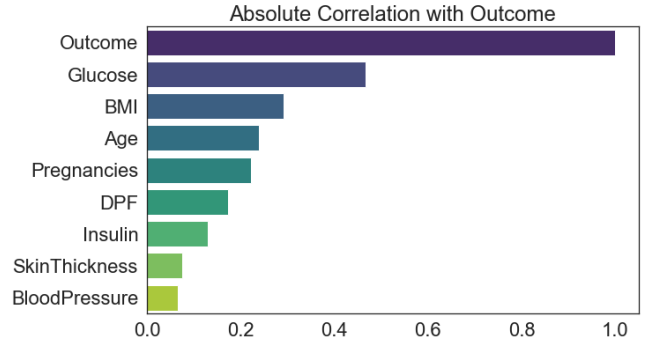


Figure 6. Correlation Between Each Variable and the Outcome.

The plot in figure 6 of the absolute correlation with outcome shows the correlation between each variable with outcome in descending order. Therefore, glucose has the highest correlation followed by BMI, Age, Pregnancies, DPF, Insulin, SkinThickness, and BloodPressure.

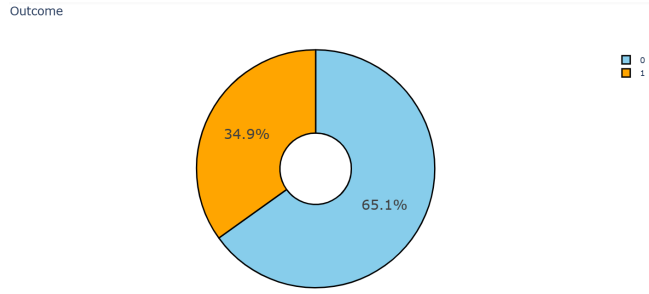


Figure 7 . Pie chart of Outcome Variable

The target variable is the class variable Outcome showing the patient's status: 0 if the patient doesn't have diabetes and 1 if the patient has diabetes. As shown in Figure 5, there are 268 patients who have diabetes, which represents 34.9% of the data, and 500 patients who do not have diabetes, which represents 65.1% of the data.

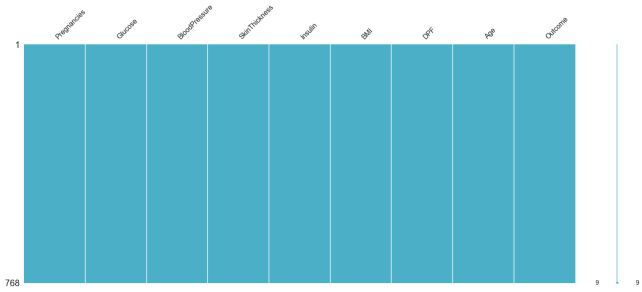


Figure 8 . msno.matrix Plot for all Variables

As shown in figure 8, the msno.matrix plot shows that there are no missing values in the dataset, indicating a complete dataset. This is beneficial for analysis as it ensures that all variables have values recorded, allowing for a comprehensive examination of the data.

CLEANING AND PRE-PROCESSING THE DATA

To clean and preprocess the dataset, we firstly checked the outliers.

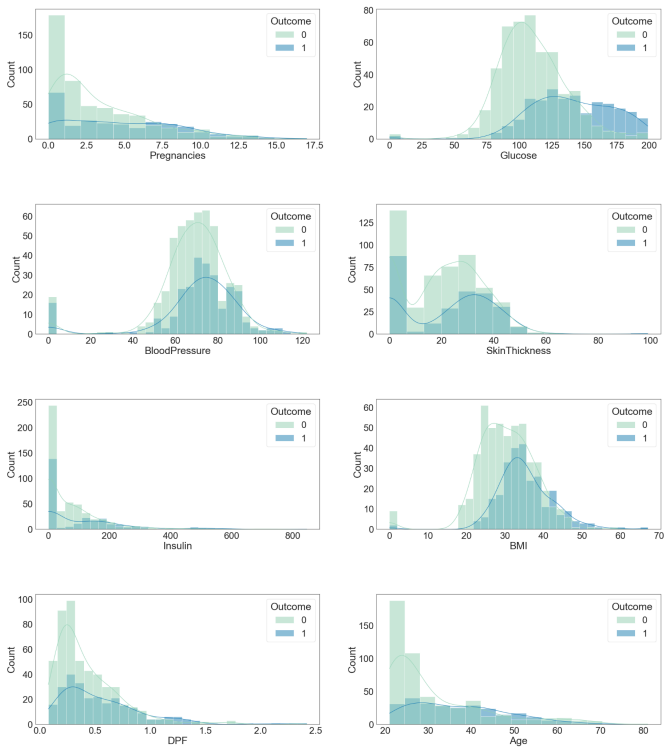


Figure 9. Histograms for each feature to show distribution and outliers.

It seemed that there are a lot of features that had 0 values and these should be treated as outliers. Age & DiabetesPedigreeFunction do not have to have minimum 0 value so no need to replace , also number of pregnancies as 0 is possible as observed. However it doesn't make sense for BMI to be equal to zero nor glucose levels. We did not want to delete the observations with zero values as our dataset contains only 768 observations, so we could not afford to lose observations.

Outliers Treatment

We counted the number of zeroes in each column that needed to be treated. We replaced the zero values in Glucose, BloodPressure, SkinThickness, Insulin, and BMI with NaN to be able to treat them. Then we replaced those NaN values with the median of the feature according to the target variable "Outcome".

```
median_target('Insulin')
```

	Outcome	Insulin
0	0	102.5
1	1	169.5

Figure 10. Median of Insulin according to target feature.

For example, Insulin's medians by the target are really different, 102.5 for a healthy person and 169.5 for a diabetic person. Therefore we needed to replace the NaN values accordingly, and we repeated the same steps for the rest of the variables until we had no more zero values.

Adding new features

We added new features as part of our feature engineering to experiment with them and see if they improve the models accuracy. We added the features:

- BMI*Thickness
- Preg/Age
- Age/Insulin
- BMI*Age

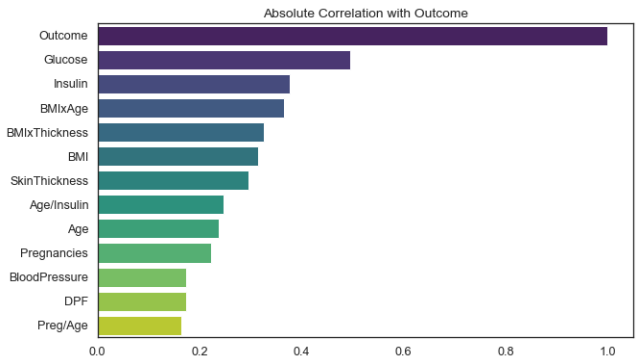


Figure 11. Sorted correlation with target feature.

Then we calculated the correlation with the target feature "Outcome" where we found some of the new features having stronger correlation with the target variables and so could be used for predictions. We needed a more reliable way to accurately choose the best features among the existing features and the new features we created. And therefore we used Extra Trees and K Best Features to estimate the importance of features.

Features Selection

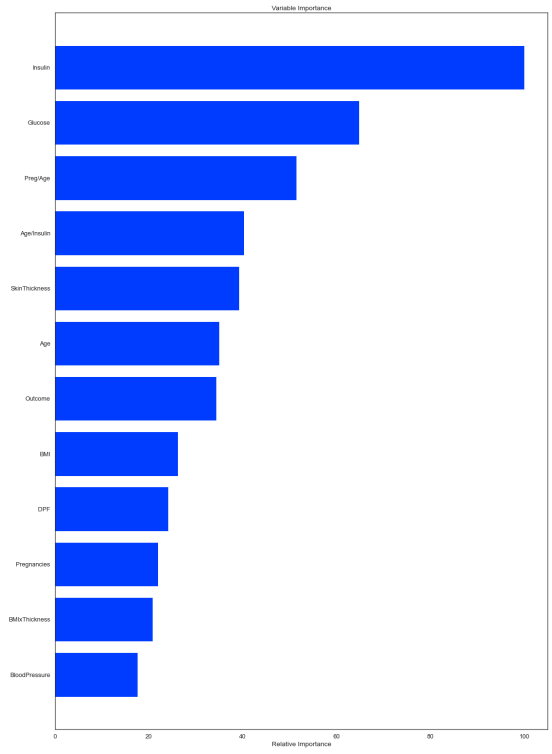


Figure 12. Sorted Variable Importance according to Extra Trees.

```
Selected features: Index(['Pregnancies', 'Glucose', 'SkinThickness', 'Insulin', 'Age',
       'BMIxThickness', 'BMIxAge'],
      dtype='object')
```

Figure 13. K Best Features output.

```
Selected Features: Index(['Glucose', 'BloodPressure', 'SkinThickness', 'BMI', 'DPF', 'Age',
       'Preg/Age', 'Age/Insulin', 'BMIxAge'],
      dtype='object')
```

Figure 14. Logistic Regression output.

Extra Trees, K Best Features, and Regularized Logistics Regression produced slightly different results, and so we chose the common features between them as our final selection of features. We chose the features:

- Pregnancies
- Glucose
- Insulin
- SkinThickness
- Age
- BMIxThickness
- BMIxAge

as they had the highest relative importance and therefore were selected to our final clean dataset.

Normalization

Finally, we had to normalize our final dataset using the MinMax scaler. It linearly scales the data down into a fixed range, where the largest occurring data point corresponds to the maximum value (1) and the smallest one corresponds to the minimum value (0). We wanted to check the accuracy of the predictions after normalizing the dataset, and so we calculated the accuracy score for some of the classifiers.

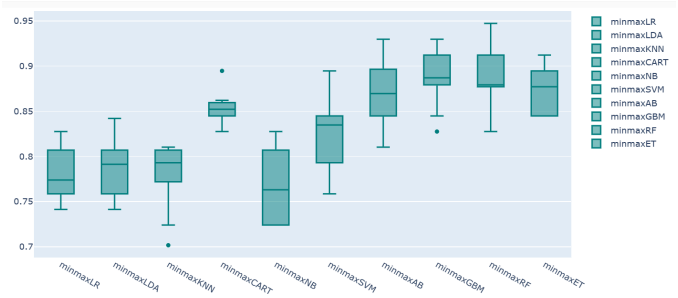


Figure 15. Accuracy of classifiers after dataset normalization.

After dataset normalization, All classifiers proved a score above 0.7 which is relatively good. Therefore we decided to proceed with this dataset.

FINAL LIST OF CHOSEN FEATURES

	Pregnancies	Glucose	SkinThickness	Insulin	Age	BMIxThickness	BMIxAge	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	0.226180	0.501142	0.240107	0.153550	0.204015	0.233260	0.301627	0.348958
std	0.198210	0.196543	0.096639	0.107092	0.196004	0.125907	0.189209	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.058824	0.359677	0.195652	0.106370	0.050000	0.146290	0.156007	0.000000
50%	0.176471	0.470968	0.228261	0.106370	0.133333	0.218169	0.262290	0.000000
75%	0.352941	0.620968	0.271739	0.186899	0.333333	0.295569	0.421203	1.000000
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Figure 16. Description of the final clean dataset.

Our final clean dataset contains 8 main features. The original features Glucose, Insulin, SkinThickness, Age and Outcome, and the engineered features BMIxThickness and BMIxAge. We chose these as the final list of features as accordng to Extra Trees, K-Best Features, and Logistic Regression they had the highest relative importance/scores.

```
final_df.shape
```

```
(768, 8)
```

Figure 17. Size of the final clean dataset.

Our new dataset now has 768 observations, as we did not lose any of our observations. It has 8 features, including the target feature “Outcome”. It does not have any missing values nor incorrect zero values. Our final dataset is also normalized in order to produce the most accurate results.

REFERENCES

[1] Dua, D., & Taniskidou, E. K.,2018,Kaggle Machine Learning Repository, “Pima Indians Diabetes Dataset”, [Online]: Retrieved from: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database/data

[2] *Understanding Boxplots - KDnuggets*. (n.d.). KDnuggets. https://www.kdnuggets.com/2019/11/understanding-boxplots.html