Project Report for Prediction of onset of Diabetes

Authors: Faith Nassiwa, Jiahui Zeng

Khoury College of Computer Science, Northeastern University

Abstract

Diabetes is a leading disease in the world. With the seriousness of diabetes and its complexity in diagnosis, we aimed to produce a model to help with prediction of onset of diabetes. Three models, Logistic Regression, Gradient Boosting and Random Forest were performed and evaluated to predict the onset of diabetes. A dataset of size 781 that includes information about some Indian Population were used. The population are specific to Indian women that are at least 21 years old and of Pima Indian Heritage. Methods of standardizing including SMOTE and hyperparameter tuning are performed. Random Forest performed the best with an accuracy score of 81.8%, followed by Gradient Boosting (78%), and followed by Logistic Regression (76%). Glucose, BMI and Age are the top predictors for Diabetes according to random forest feature importance. Because of the limited dataset we used in this dataset, more future available datasets are hoped to improve the accuracy of the models and give more information about the onset of diabetes. Moreover, this dataset is very specific to some group, future datasets with information about broader groups (including more age, gender and race) might gave more insights about this issue.

Introduction

Diabetics is a leading disease in the world. According to CDC, 37.3 million people (that is 11.3% of the US population) have diabetes and 96 million people that aged 18 years or older have prediabetes. Moreover, Diabetes is a major cause of blindness, kidney failure, heart attack, stroke and lower limb amputation.

The normal tests for type 1 and type 2 diabetes and prediabetes are AIC (Glycated hemoglobin) test, random blood sugar test, fasting blood sugar test, and oral glucose tolerance test, according to one online survey, about 25% of all participants were misdiagnosed with diabetes, and the misdiagnosis was associated with risk for Diabetic ketoacidosis.

Therefore, due to this above information, this problem is interesting because it might be helpful for the diagnosis of diabetes. It provides a way for machines to diagnose diabetes. This problem and dataset provide a way for us to gain some perspective of diabetes. It looks at some features, like blood glucose, BMI and other factors to try to predict if a patient has diabetes or not. It can help with the prevalence of misdiagnosis of diabetes. Moreover, this problem can also be examined to help us find out which features are mostly related to the diagnosis of diabetes. Therefore, this problem is interesting because it provides us with a way to machine predicts the diagnosis of diabetes and a way to look at the features related to diabetes.

What is the approach you propose to tackle the problem?

We will use selected supervised machine learning models to predict the onset of diabetes in the given dataset. Below is a list of models that we will use and compare their performance based on defined evaluation metrics (Accuracy, F1-score, Recall, Precision, RocAUC).

- Logistic regression
- Boosting methods, we shall use Gradient Boosting

- Bagging methods, we shall use Random Forests

We will also explore feature selection approaches (forward selection) to select the best features for the prediction.

Why is the approach a good approach compared with other competing methods?

We shall use the above-selected models because the dataset is small, numerical, tabularly structured, and work on classification problems. We will not use any deep learning methods because the number of observations is few (768) and the data is already tabularly structured thus no need to find a good vector representation.

Logistic regression makes no assumptions about the underlying distribution of the data and should be the first method to try for classification problems. It does not require a linear relationship between the target and predictors. Gradient boosting iteratively combines multiple decision trees removing the randomness seen in Random Forest and it is often more scalable. K Nearest Neighbors is sensitive to noise and missing data, does not work well with high dimensionality as it complicates the distance calculating process to calculate the distance for each dimension

Random Forests applies random sampling of predictors while performing bagging which addresses the problem that Bagging often produces similar decision trees. It requires little to no data preprocessing and automatically handles overfitting. It is suitable for nonlinear problems and gives better results than decision trees. accuracy is usually high, it does not overfit with more features like other methods.

What are the key components of my approach and results? Any specific limitations?

Data Preprocessing and Cleaning, Data Analysis/Exploration, Modeling, and Evaluation of results.

Limitations of the selected models include; lack of data, most of the models perform best with large datasets, and lack of interpretability most especially with Gradient boosting.

Preliminaries

The Pima Indians Diabetes dataset is made up of eight independent variables which include Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function and Age that can be used to predict onset of diabetes and one dependent variable, Outcome that has binary data values with 0 meaning no diabetes and 1 meaning the woman is diabetic. The dataset contains 768 observations describing female patients, and all the eight independent variables are of a numeric (int or float) data type.

Variable	Description
Pregnancies	Number of times pregnant
Glucose	Plasma glucose concentration 2 hours in an oral glucose tolerance test
Blood Pressure	Diastolic blood pressure (mm Hg)
Skin Thickness	Triceps skin fold thickness (mm)
Insulin	2-Hour serum insulin (mu U/ml)
BMI	Body mass index (weight in kg/(height in m)^2)
Diabetes Pedigree Function	Diabetes pedigree function
Age	Age(Years)
Outcome	Class variable (0 or 1) 268 of 768 are 1(diabetic), the others are 0(non-daibetic)

Below is a summary of the basic statistic of the variables in the dataset.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Figure 1: Summary stats for Diabetes dataset

While analyzing the data, we observed that there were no nulls however some of the independent variables contained ‘0’ data entries that are not expected / feasible for the variables.

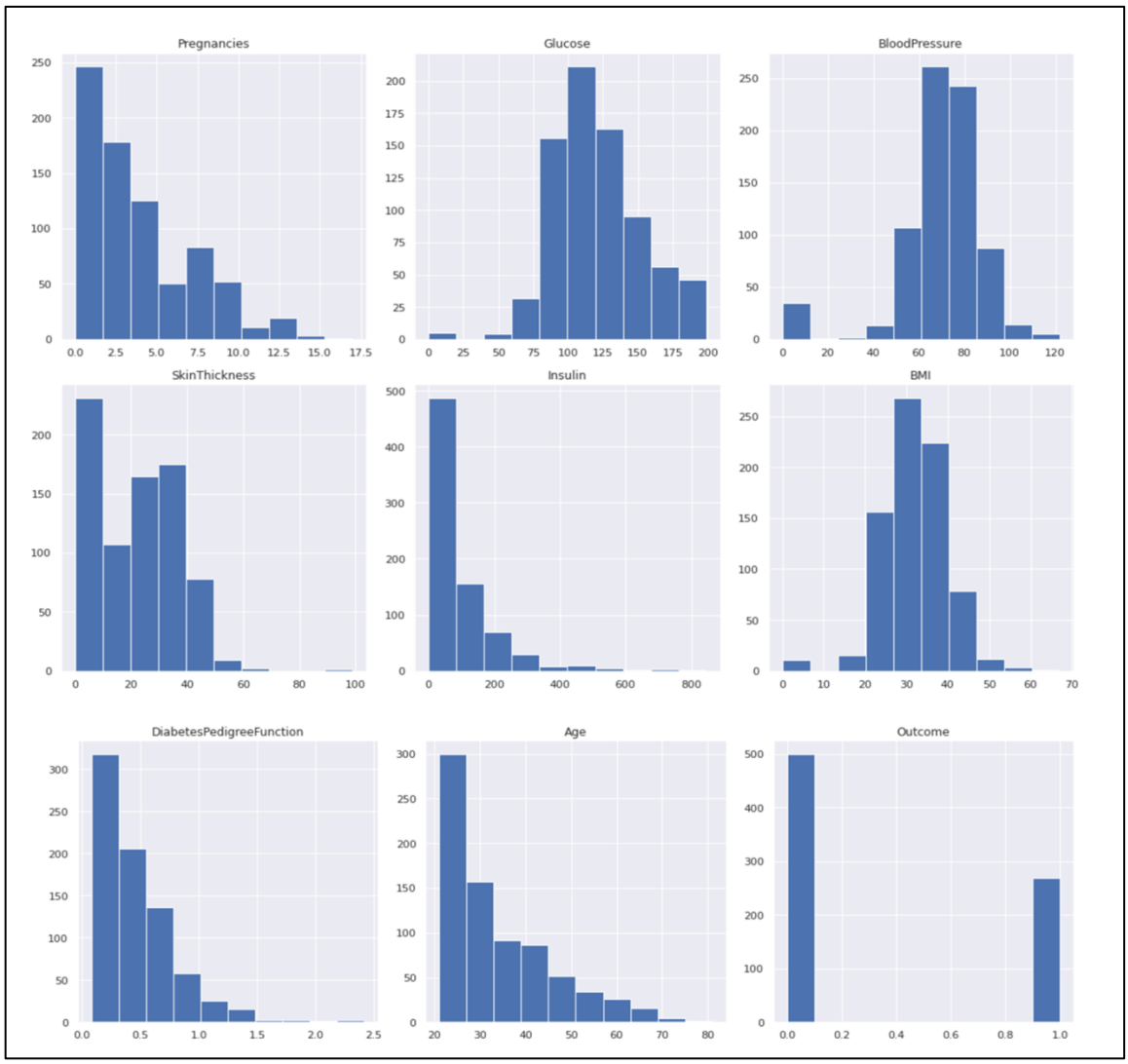


Figure 2: Histograms for all features in the Diabetes dataset

Below are the variables and their corresponding analysis findings:-

- Glucose - 5 observations with 0 reading (0.7% of the observations)
- Blood Pressure - 35 observations with 0 reading (5% of the observations)
- Skin Thickness - 227 observations with 0 reading (30% of the observations)
- Insulin - 374 observations with 0 reading, (49% of the observations)
- BMI - 11 observations with 0 reading (1.4 % of the observations)

We decided to impute all the zero readings in the above variables with the column means as removing this data would have greatly reduce the number of observations in the dataset that is already small.

We then used a correlation matrix to check for multicollinearity and how the independent variables relate to the Outcome and observed that Glucose has the highest correlation to the Outcome followed by BMI, Age and Pregnancies and that all the features are positively correlated to the Outcome / target feature.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1.000000	0.129459	0.141282	-0.081672	-0.073535	0.017683	-0.033523	0.544341	0.221898
Glucose	0.129459	1.000000	0.152590	0.057328	0.331357	0.221071	0.137337	0.263514	0.466581
BloodPressure	0.141282	0.152590	1.000000	0.207371	0.088933	0.281805	0.041265	0.239528	0.065068
SkinThickness	-0.081672	0.057328	0.207371	1.000000	0.436783	0.392573	0.183928	-0.113970	0.074752
Insulin	-0.073535	0.331357	0.088933	0.436783	1.000000	0.197859	0.185071	-0.042163	0.130548
BMI	0.017683	0.221071	0.281805	0.392573	0.197859	1.000000	0.140647	0.036242	0.292695
DiabetesPedigreeFunction	-0.033523	0.137337	0.041265	0.183928	0.185071	0.140647	1.000000	0.033561	0.173844
Age	0.544341	0.263514	0.239528	-0.113970	-0.042163	0.036242	0.033561	1.000000	0.238356
Outcome	0.221898	0.466581	0.065068	0.074752	0.130548	0.292695	0.173844	0.238356	1.000000

Figure 3: Correlation Matrix for the Diabetes dataset

We created a scatter plot using the top two correlating features to the Outcome and observed that the women with a high Glucose and high BMI reading tend to be diabetic compared to their counterparts.

Figure 4: Scatter plot showing relationship between Glucose, BMI and the Outcome

There were no outliers detected in the dataset per the box plots that we ran for very feature.

We then split our dataset to have 80% training data and test on 20% of the data and we standardized the data using StandardScaler() to cater for any independent variables that had values that were greatly higher or lower than other values in the dataset. StandardScaler transforms the feature by subtracting the mean and dividing with the standard deviation. This way the feature also gets close to standard normal distribution with mean 0.

We then noticed that the dataset was imbalanced with 2:1 ratio on 0 and 1 Outcome values respectively. Usually for a dataset to be called imbalance a ratio of 10:1 is required however with the size of the data the 2:1 ratio is significant.

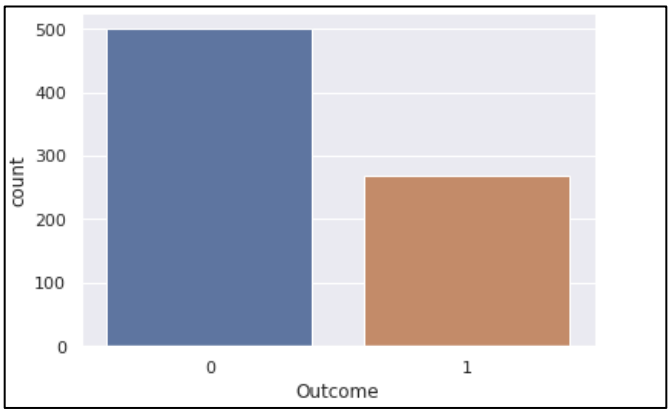


Figure 5: Outcome data distribution

To deal with the data imbalance, we used Synthetic Minority Oversampling (SMOTE) to create synthetic data points for the minority class since the dataset is small and does not suffer from high dimensionality. This helped increase the size of our training dataset and our models' performances across the board.

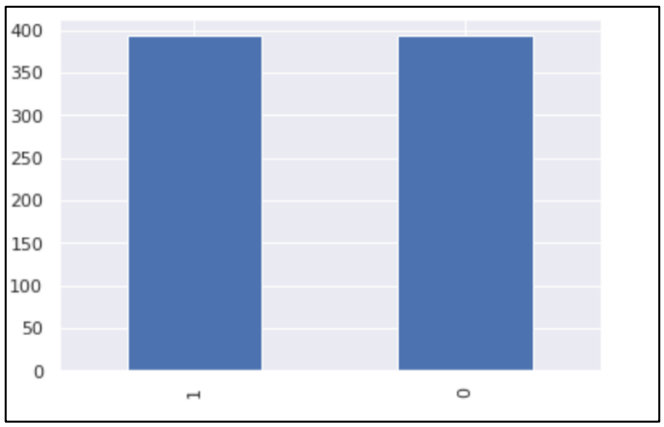


Figure 6: Outcome training data distribution after applying SMOTE

We used three models; Logistic Regression, Gradient Boosting and Random Forest for modeling our dataset due to reasons stated in the Introduction. We completed cross validations to tune our models to identify the best hyperparameters for the model and evaluated the models on specific metrics including Accuracy, F1-score, Recall, Precision, RocAUC. We executed our project on Google Colab using Python 3.8.16.

Results

Logistic Regression

We first used Logistic Regression to model our dataset and we got an accuracy of 75% on the test data. Out of the predicted positives(Precision), 57% were correct and out of the actual positives(Recall), 70% were correct. With a ROC AUC score above 0.5, the model performed better than random.

```
Logistic Regression Model Accuracy Score is: 0.747
Logistic Regression Model Confusion Matrix:  [[82 25]
 [14 33]]
Logistic Regression Model Precision Score: 0.569
Logistic Regression Model Recall Score: 0.702
Logistic Regression Model F1 Score: 0.629
Logistic Regression Model ROC AUC Score: 0.712
```

We then used Grid Search to tune the model and observed a slight accuracy score improvement to 76%.

```
Fitting 10 folds for each of 40 candidates, totalling 400 fits
Logistic Regression Grid Search Best parameters: {'C': 0.03359818286283781, 'penalty': 'l2', 'solver': 'liblinear'}
Logistic Regression Grid Search Best Training Score: 0.749
Logistic Regression Grid Search Best Test score: 0.760
```

Gradient Boosting

Secondly, we tested using Gradient Boosting and we got an accuracy score of 78% on the test data before any hyperparameter tuning. Out of the predicted positives(Precision), 60% were correct and out of the actual positives(Recall), 81% were correct. With a ROC AUC score above 0.5, the model performed better than random.

```
Gradient Boosting Classifier Model Accuracy Score is: 0.779
Gradient Boosting Classifier Model Confusion Matrix:  [[82 25]
 [ 9 38]]
Gradient Boosting Classifier Model Precision Score: 0.603
Gradient Boosting Classifier Model Recall Score: 0.809
Gradient Boosting Classifier Model F1 Score: 0.691
Gradient Boosting Classifier Model ROC AUC Score: 0.752
```

We then used Grid Search to tune the model but observed no accuracy improvement.

```
Fitting 10 folds for each of 144 candidates, totalling 1440 fits
Gradient Boosting Classifier Grid Search Best parameters: {'learning_rate': 0.04, 'max_depth': 4, 'n_estimators': 90}
Gradient Boosting Classifier Grid Search Best Training Score: 0.795
Gradient Boosting Classifier Grid Search Best Test score: 0.773
```

Random Forest

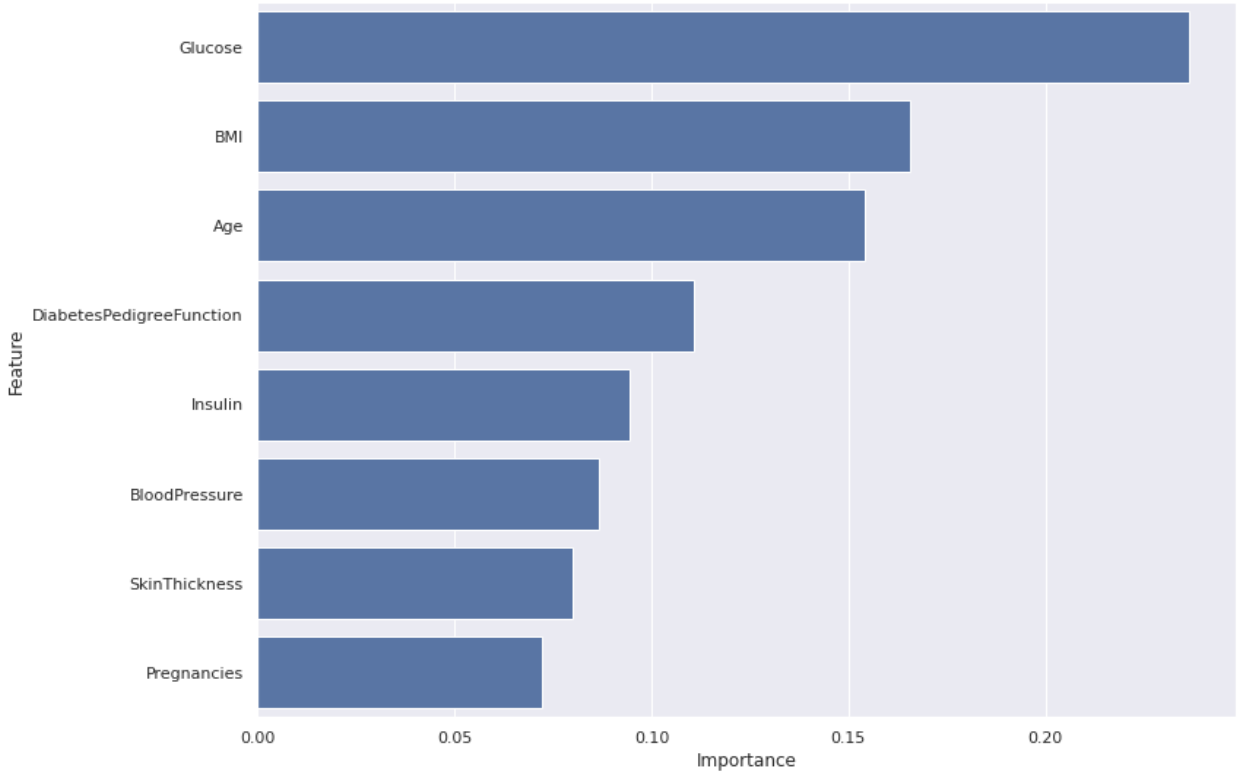
Finally, we tested using Random Forest and we got an accuracy score of 80% on the test data before any hyperparameter tuning. Out of the predicted positives(Precision), 65% were correct and out of the actual positives(Recall), 72% were correct. With a ROC AUC score above 0.5, the model performed better than random.

```
Random Forest Classifier Model Accuracy Score is:0.799
Random Forest Classifier Model Confusion Matrix:  [[89 18]
 [13 34]]
Random Forest Classifier Model Precision Score: 0.654
Random Forest Classifier Model Recall Score: 0.723
Random Forest Classifier Model F1 Score: 0.687
Random Forest Classifier Model ROC AUC Score: 0.763
```

```
Random Forest Classifier Model Accuracy Score is:0.818
Random Forest Classifier Model Confusion Matrix:  [[90 17]
 [11 36]]
Random Forest Classifier Model Precision Score: 0.679
Random Forest Classifier Model Recall Score: 0.766
Random Forest Classifier Model F1 Score: 0.720
Random Forest Classifier Model ROC AUC Score: 0.785
```

Then we tuned the random forest model with cross validation on the parameters n_estimators, max_depth and max_features. We got a model with accuracy 0.818 with n_estimators = 300, max_depth = 9 amd max_features equal to the square root of number of features, we got an accuracy score of 0.818. Random forest seems to perform the best out of all the models we tested.

We also did feature importance with the random forests and find out that Glucose, BMI and Age are the top three predictors for determining the onset of diabetes.



Discussion

As shown above, random forest performed the best out of all the three models we tested (random forest, Gradient Boosting and Logistic Regression), and that are suitable for our data.

Random forest gave an accuracy score of 81.8%, it works better than Gradient boosting since it's an updated version of gradient boosting with random selected subset of features when performing bagging each time. It gave randomness to the features that we selected to build each decision tree and, in a way, helps with preventing overfitting. Random forest also performs better than logistic regression because it uses sampling and subset selection, while logistic regression depends more on the features and overfits more easily.

Moreover, Random Forest also gave us some important information about which predictors are best for predicting the onset of diabetes. As we can see, Glucose is the best predictor for onset of diabetes, followed by BMI, age and Diabetes Pedigree Functions. Glucose is the level of blood sugar, which are directly to the Diabetes since diabetes are an inability to produce insulin, which are for processing blood sugar. BMI is the second predictor because overweight individuals are less sensitive to insulin and thus an important predictor for diabetes. People who are older have more insulin resistance, thus producing less insulin, which is also related to diabetes. Other features also help with prediction of diabetes.

Our results gave around 82% accuracy with random forest. Random forest also seems to perform the best out of all models in the current trend for classification problem in the machine learning field. Our accuracy score was only 77% before we performed the SMOTE method, which helps with the data imbalance in the original data. It improved from 77% to 82% after performing SMOTE method. However, our data were not as good at predicting because of the limited dataset we had. Such model can probably perform better with more data provided in the future and can be a useful tool in the medical field.

Conclusion

Three machine learning models including Logistic Regression, Gradient Boosting and Random Forest are selected and performed to determine the best one for predicting the onset of diabetes using the Pima Indian Diabetes dataset. We used Python 3.7 and Google Colab to perform basic exploratory data analysis, preprocessed the data (splitting train/test sets, applying SMOTE to help with data imbalance and standardizing the data using scalar method), ran and trained the models using hyperparameters tuning, and use them to predict target variable in the test data set, and evaluated the results for each model.

All the three models performed better after standardizing and applying the SMOTE method to the training data. Random Forest emerges as the best model with an accuracy score of 82%, followed by Gradient Boosting with accuracy score 78% , and Logistic Regression with accuracy score 76% came in last. We were shooting for an accuracy score of 85% but we faced limitation reaching this target due to the few observations in the dataset.

References

1. Learning, UCI Machine. “Pima Indians Diabetes Database” Kaggle, October 6, 2016. Accessed at https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database on November 10, 2022
2. Centers for Disease Control and Prevention. “National Diabetes Statistics Report” CDC, June 29, 2022. Accessed at https://www.cdc.gov/diabetes/data/statistics-report/index.html on November 10, 2022
3. World Health Organization. “Diabetes” Accessed at https://www.who.int/news-room/fact-sheets/detail/diabetes on November 10, 2022
4. Mayo Foundation for Medical Education and Research. “Diabetes” Mayo Clinic. Accessed at https://www.mayoclinic.org/diseases-conditions/diabetes/diagnosis-treatment/drc-20371451 on November 10, 2022.
5. U.S. National Library of Medicine. “Misdiagnosis and Diabetic Ketoacidosis at Diagnosis of Type 1 Diabetes: Patient and Caregiver Perspectives” Clinical diabetes : a publication of the American Diabetes Association, July 2019. Accessed at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6640891/ on November 11, 2022
6. JDRF. “The Complexity of Diagnosing Type 1 Diabetes” Accessed at https://www.jdrf.org/t1d-resources/about/diagnosis/ on November 11, 2022.
7. Dibble, Megan. “A Guide to EDA in Python” Medium. Level Up Coding, June 13 2020. Accessed at https://levelup.gitconnected.com/cozy-up-with-your-data-6aedfb651172 on December 01, 2022.
8. Choudhary, Ishan. “Pima Indian Diabetes Prediction.” Medium. Towards Data Science, March 19, 2020. Accessed at https://towardsdatascience.com/pima-indian-diabetes-prediction-7573698bd5fe on December 10, 2022.