Early-stage Risk Diabetes Prediction: Project Report

Carlsean Claricia, Sierra Harris, Sean Iida, Abhay Solanki,

Brian Tran, Kenneth Valero, and Brian Vu

College of Engineering, California State University, Long Beach

CECS 456: Machine Learning

Professor Mahshid Fardadi

April 26, 2023

1 Introduction

The detection of early stage risk of diabetes can be closely predicted using a machine learning model that identifies if a patient is newly diabetic or would be. In order to do so first, the dataset needs to be checked for any missing or malformed data and then perform a preliminary test to get a better understanding of the data such as the number of features, the number of patients, along with those diagnosed with diabetes and those not. This information will further help determine the correct model of classification that can best represent the data and give them the most accurate predictions. Classifications such as logistic regression, random forest classifier, and support vector machines will be compared to determine the best for the dataset**.** Next, we can analyze the relationships between variables and the patient's diabetes classification. Lastly we will plot ROC curve of each classification model and calculate the area under the curve (AUC) which will lead us to choose the best classification model to use for a dataset.

2 Data Analysis

Before creating the machine learning classification model, analysis has to be performed to gain a better understanding of the dataset before making any assumptions through a common process known as Exploratory Data Analysis (EDA). By performing EDA, we can check for missing or malformed data in the dataset, compare and contrast relationships and patterns within the data, and deal with outliers and other anomalies [1]. In this experiment, we use Python to leverage powerful libraries for data manipulation, data visualization, and machine learning.

2.1 Dataset Overview

First, we performed preliminary tests on our dataset to gain a better understanding. In this project, we use an early stage diabetes dataset sourced from the UCI Machine Learning Repository; the dataset contains various features that may be a precursor to diabetes. Using the Pandas Dataframe methods, we find that the dataset has 17 features from a collection of 520 patients. Looking closer at the data in Fig. 1, all columns of the dataset besides the Age, are binary classifications with the last column being the diabetes classification.

Fig 1.

```
# First 5 rows of dataset
df.head()
```

[4]

Visualize

	Age int64	Gender object	Polyuria object	Polydipsia object	sudden weight l...	weakness object	Polyphagia obje
0	40	Male	No	Yes	No	Yes	No
1	58	Male	No	No	No	Yes	No
2	41	Male	Yes	No	No	Yes	Yes
3	45	Male	No	No	Yes	Yes	Yes
4	60	Male	Yes	Yes	Yes	Yes	Yes

2.2 Data Cleaning

Second of all, the dataset needs to be cleaned before any work can be done with it. The dataset did not have any missing or null values to clean up. Additionally, since Age was the only numerical feature, we quickly found we did not have any outliers. However, the binary classifications needed to be converted into 1's and 0's to be used in Python. In Fig. 2, we successfully map "Yes", "Positive", and "Female" to 1 and ''No", "Negative, and "Male" to 0.

Fig. 2

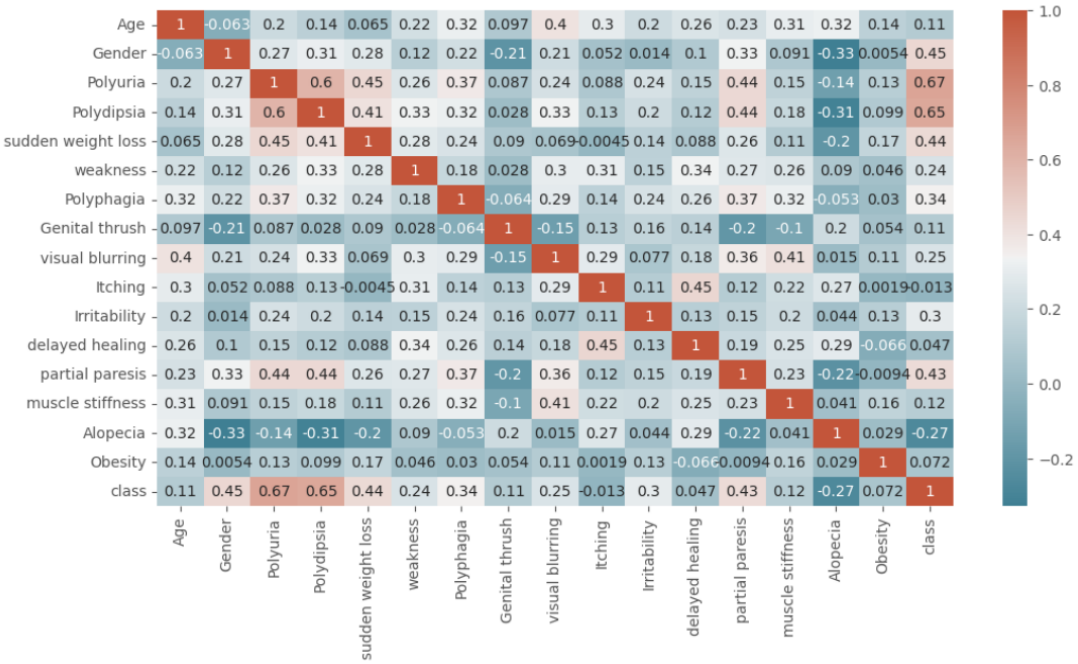
	Age int64 16 - 90	Gender int64 0 - 1	Polyuria int64 0 - 1	Polydipsia int64 0 - 1	sudden weight l... 0 - 1	weakness int64 0 - 1	Polyphagia int6 0 - 1
0	40	0	0	1	0	1	
1	58	0	0	0	0	1	
2	41	0	1	0	0	1	
3	45	0	0	0	1	1	
4	60	0	1	1	1	1	

2.3 Data Visualization

Finally, we analyzed the dataset to find relationships between variables and the patient's diabetes classification. We use a heatmap to quickly visualize the correlation between features and the classification, which is represented as a value between -1 and 1. The greater the number, the higher the correlation between the two variables. For example, the correlation between Polydipsia, excessive thirst, and Polyuria, excessive urination, is 0.6, meaning there is a

positive correlation between those features. In our experiment, we decide the best features to use for our learning model based on the correlation between the feature and the patient's classification, which is the bottom row or right column of the heatmap found in Fig. 3.

Fig. 3



As a result of the binary categories for nearly all of the features, data visualizations such as the scatterplot or boxplot are not reliable because all the data points sit at either 1 or 0. Consequently, we plotted the features with the highest correlation from the heatmap onto grouped bar charts that display the number of counts for each feature. Fig. 4 shows a grouped bar chart with all major features with a correlation on the heatmap vs. count. The red bar represents the number of "No" or "Male" classifications and the blue represents the number of "Yes" or "Female" classifications. Fig. 4 tells us that out of all of these features, men are significantly more likely than women to have positive diabetes classification.

Fig. 4

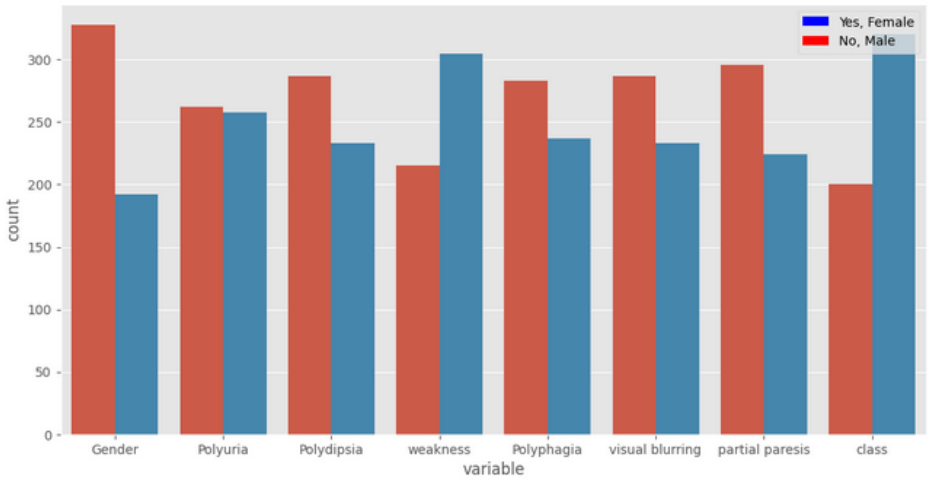
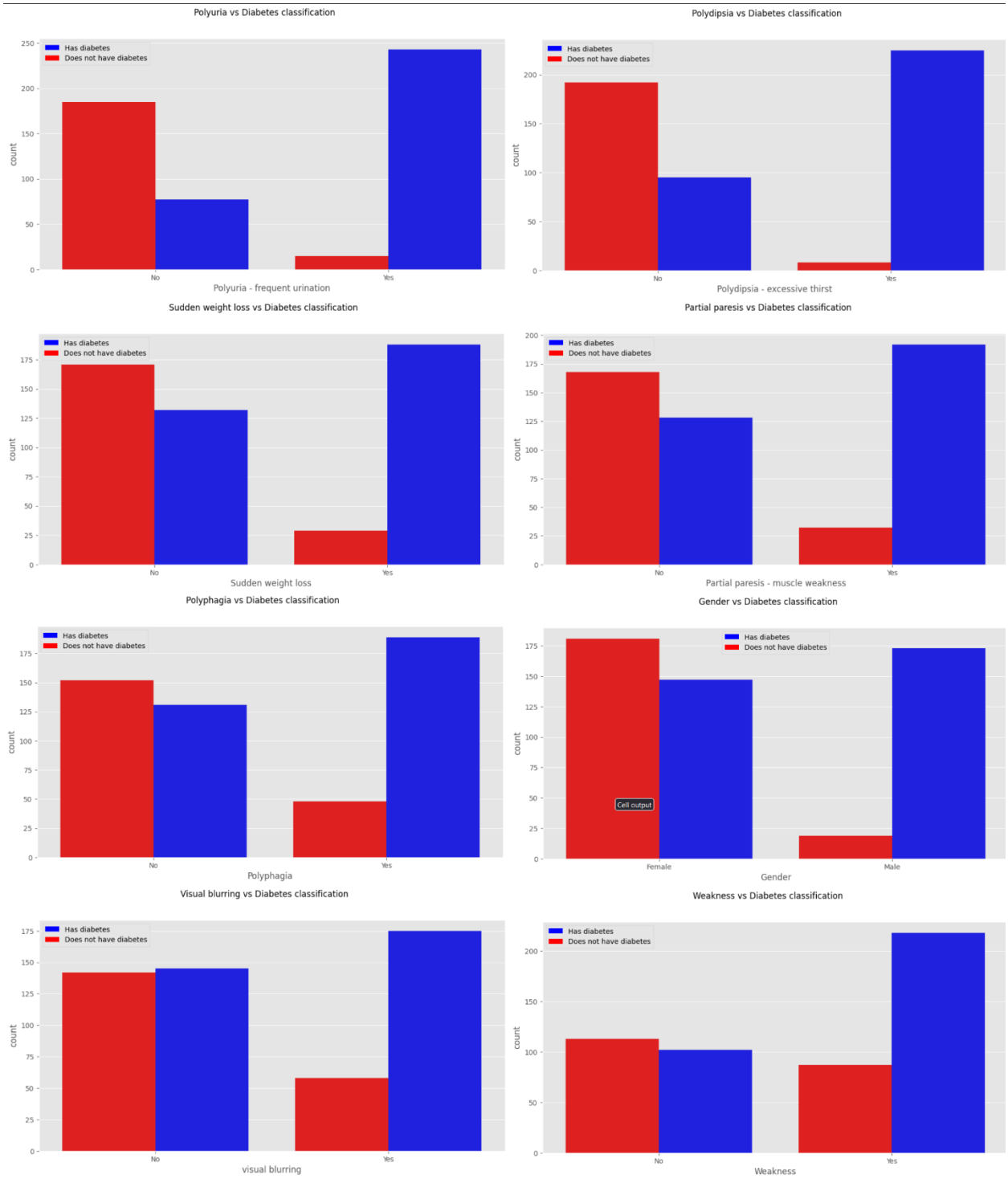


Fig. 5 displays group bar charts between a single feature and the patient's diabetes classification. The left pair of bars represents the count of "No" or "False" classifications for a condition, whereas the right pair represents the count of "Yes" or "Positive" classifications. Additionally, the red bar represents a false classification of diabetes, while the blue bar represents a positive classification. We found that within the dataset, the features found in Fig. 5 had a high correlation with diabetes when the condition was present. Furthermore, features with a low diabetes count, when the condition was not present, have a stronger correlation to a positive diabetes classification than features with a higher diabetes count. In other words, some features, such as Polyuria and Polydipsia, had a greater correlation to a patient having diabetes than features, like visual blurring or weakness. Based on the analysis, the following features in the dataset may be suitable for classifying diabetes: Polyuria, Polydipsia, sudden weight loss, partial paresis, polyphagia, gender, visual blurring, and weakness.

Fig. 5


Polyuria vs Diabetes classification
Has diabetes
Does not have diabetes
count
No
Yes
Polyuria - frequent urination
Polydipsia vs Diabetes classification
Has diabetes
Does not have diabetes
count
No
Yes
Polydipsia - excessive thirst
Sudden weight loss vs Diabetes classification
Has diabetes
Does not have diabetes
count
No
Yes
Sudden weight loss
Partial paresis vs Diabetes classification
Has diabetes
Does not have diabetes
count
No
Yes
Partial paresis - muscle weakness
Polyphagia vs Diabetes classification
Has diabetes
Does not have diabetes
count
No
Yes
Polyphagia
Gender vs Diabetes classification
Has diabetes
Does not have diabetes
count
Cell output
Female
Male
Gender
Visual blurring vs Diabetes classification
Has diabetes
Does not have diabetes
count
No
Yes
visual blurring
Weakness vs Diabetes classification
Has diabetes
Does not have diabetes
count
No
Yes
Weakness

3 Experiment

3.1 Machine Learning Model

When choosing a classification model to use for a dataset, it is important to pick a classification that is the most accurate for the conditions of the provided dataset. To verify the effectiveness of each classification model for the dataset, we will be plotting the ROC curve of each classification model and calculating the area under the curve (AUC).

Before we are able to set up the ROC curve, we will need to use the sk_learn function predict_proba for each model to calculate the probability that the model will classify whether each patient is diabetic or not. Additionally, we will also be calculating the probability for an entirely randomized model as a reference point. (Fig 6)

Fig. 6

```
r_probs = [0 for _ in range(len(y_test))]
lr_probs = log_reg.predict_proba(X_test)[:, 1]
rf_probs = rf.predict_proba(X_test)[:,1]
svm_probs = svm.predict_proba(X_test)[:,1]
```

Once we have calculated the prediction probabilities for each model, we will use the data collected to create the ROC curve for each model, plotting the false positive rate versus the true positive rate. Additionally, we will use the prediction probabilities to calculate the area under the ROC curve for each model. (Fig. 7)

Fig. 7

```
#Calculating the ROC Curve (false positive rate vs. true positive rate)
random_fpr, random_tpr, threshold1 = roc_curve(y_test, r_probs)
lr_fpr, lr_tpr, threshold2 = roc_curve(y_test, lr_probs)
rf_fpr, rf_tpr, threshold3 = roc_curve(y_test, rf_probs)
svm_fpr, svm_tpr, threshold4 = roc_curve(y_test, svm_probs)
```

```
#calculating AUC
random_auc = roc_auc_score(y_test, r_probs)
lr_auc = roc_auc_score(y_test, lr_probs)
rf_auc = roc_auc_score(y_test, rf_probs)
svm_auc = roc_auc_score(y_test, svm_probs)
```

3.2 Feature Ranking

In order to determine a potential ranking of features, it would be necessary to decide on a specific feature selection algorithm. In this case, the Recursive Feature Elimination(RFE) algorithm was chosen due to multiple different tools that it provides through the usage of its parameters. The first parameter is the estimator which will consist of the learning algorithm that will be used as a basis for the RFE algorithm, which in this case is the Logistical Regression model. Next is the feature to select parameters where users may dictate the number of features to maintain at the end of the algorithm's pruning process. This parameter was set to one in order to determine the more prominent feature that may lead to diabetes. The last parameter is called step, which will dictate the number of features, one, that will be pruned at each iteration of the RFE algorithm until the feature to select parameter is met. These parameters would lead to building a rankings list of the most prominent to least prominent features(Fig. 8), showing how the most prominent feature leading to diabetes is polydipsia which has a 80% accuracy rating to predicting the correct results.

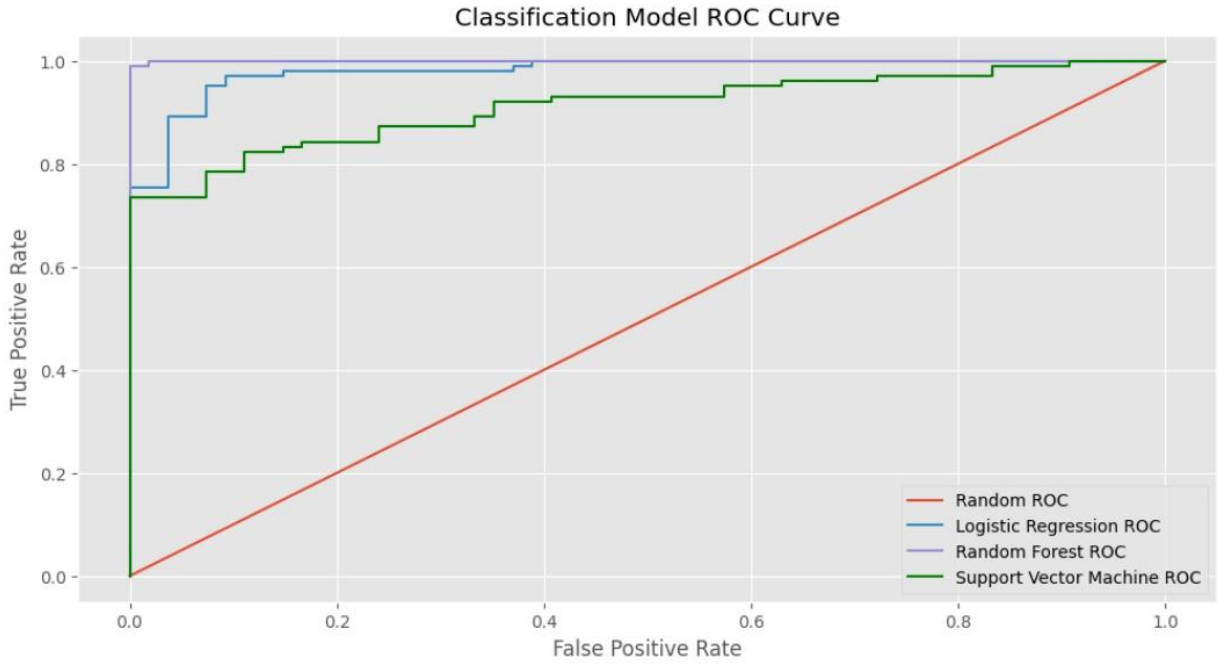
Fig. 8

```
Accuracy:  0.8076923076923077
[1, 'Polydipsia']
[2, 'Polyuria']
[3, 'Gender']
[4, 'Irritability']
[5, 'Itching']
[6, 'sudden weight loss']
[7, 'Genital thrush']
[8, 'partial paresis']
[9, 'visual blurring']
[10, 'delayed healing']
[11, 'weakness']
[12, 'muscle stiffness']
[13, 'Alopecia']
[14, 'Polyphagia']
[15, 'Obesity']
[16, 'Age']
```

4 Results

The graph displayed in Fig. 9 shows the ROC curves for each classification method as well as their AUC scores. Out of the three classification methods focused on for this study, we found that the Random Forest classifier is most effective for classifying patients as diabetic or not while the Support Vector Machine classifier is the least effective.

Fig. 9




```
Random AUC = 0.500
Logistic Regression AUC = 0.980
Random Forest AUC = 1.000
Support Vector Machine AUC = 0.911
```

5 Discussion

Out of the three classification models used for this study, Random Forest had the highest AUC score. This is likely due to the fact that Random Forest is typically most effective with large datasets with many features that affect classification. In this case, there are many features that can relate to diabetes to account for. Logistic regression is likely not as effective for this dataset due to logistic regression's reliance on strict linear relationships between a feature and a classification, which is not entirely present in this dataset. Lastly, contrary to Random Forest, the Support Vector Machine model tends to struggle with large datasets with many important features to take into account for classification, which leads to it having the worst performance of the three models in this dataset.

6 Conclusion

In conclusion, plotting the ROC curve for each classification model and calculating their AUC scores we have determined that the best classification model to best represent the dataset is the random forest classification. The random forest classification was calculated to have the highest AUC out of the ROC curves plotted. This means that this classification has the highest rate of positively classifying patients as either diabetic or not. This classification has also been determined to work best on larger datasets that contain many important features to consider such as polyuria, polydipsia, sudden weight loss, partial paresis, polyphagia, gender, visual blurring, and weakness. On the other hand, we found that the support vector machine model was the least effective with this dataset at determining a patient to be diabetic or not.

7 References

[1] IBM, 2023. What is exploratory data analysis? Retrieved April 19, 2023 from https://www.ibm.com/topics/exploratory-data-analysis#:~:text=The%20main%20purpose%20of%20EDA,interesting%20relations%20among%20the%20variables.