Diabetes: A Machine Learning Approach

Oluwademilade Adisa
Faculty of Science
University of Prince Edward Island
Charlottetown, Canada
odadisa@upei.ca

Shivangi Sharma
Faculty of Science
University of Prince Edward Island
Charlottetown, Canada
ssharma3@upei.ca

***Abstract*—Diabetes is a leading cause of death in the world and the author Shivangi is quite familiar with this treacherous disease, her family (who is from India) has a history of diabetes. So this presents us with an issue, can we predict the possibility of diabetes developing in an individual? This is what our research is aimed at. We set out to see if we can use the prediction tools we learned from our Machine Learning course to try and predict the appearance of diabetes in women at least 21 years old of Pima Indian heritage. Although there are a multitude of tools and techniques at a doctors disposal to identify the development of diabetes in an individual, we set out to try and create a cheaper and more affordable option for the people who cannot afford expensive doctor checkups. Please note that all models were implemented using Python and a couple of its data libraries. Since this is a classification problem, we have set out to use three models, namely, K-Nearest Neighbour, Logistic Regression, and Decision Tree.**
Please note, a copy of the research done and the data-set can be accessed using this link: https://github.com/DemiAdisa/Diabetes-ML

I. Introduction

As mentioned earlier in the Abstract, one of the authors is quite familiar with diabetes. Our goal in the previous assignment was to see if we can predict the appearance of diabetes in women of Indian descent later on in their lifespans when they got older. To accomplish this, we needed to get information from women at least 21 years old of Pima Indian heritage and When computing our predictions, we took into account these details from the women:

- Pregnancies (Preg): Number of times a woman has been pregnant before the survey was carried out.
- Glucose (Gluc): Plasma glucose concentration a 2 hours in an oral glucose tolerance test
- Blood Pressure (B.P): Diastolic blood pressure (mm Hg)
- Skin Thickness(S.T): Triceps skin-fold thickness (mm).
- Insulin: 2-Hour serum insulin (mu U/ml).
- BMI: Body mass index (weight in kg/(height in m-square).
- Diabetes Pedigree Function(D.P.F): Diabetes pedigree function
- Age: Age in years
- Outcome (O.C): Whether they have diabetes (1) or not (0)

This data was collected from a total of 768 women. The survey was represented on a table and each feature observation from each woman was displayed as a row on this table. Below is a representation of the said table and its information:

Preg	Gluc	B.P	S.T	Insu	BMI	D.P.F	Age	O.C
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	0	23.3	0.167	21	0
8	183	64	0	0	23.3	0.672	32	1

FIG.1: Concise Representation of Data-set

II. Background

Now in FIG.1 lies a concise tabular representation of our data-set. In this data-set, we can see an outcome column, this is the column/data we are interested in and are trying to predict. It is a class variable (0 or 1) where 0 means they do not have diabetes and 1 means they have diabetes. We made use K-Nearest Neighbour, Logistic Regression and Decision Tree classification models to help us make our predictions. Due to the nature of the data we are trying to predict (it is discrete data), we are making use of classification supervised machine learning algorithms. These algorithms are:

- K Nearest Neighbor Classification Model
- Decision Tree Classifier
- Logistic Regression

The K-Nearest Neighbor classifier is one of the basic supervised classifiers that any data scientist should be familiar with. Fix Hodges utilised this approach for the first time in 1951 for a pattern categorization task. Pattern recognition is the goal of KNN. (Saji, 2021)

KNN, or K-Nearest Neighbor, is a supervised learning technique that may be used for both regression and classification tasks. It's most commonly used in machine learning for categorization challenges. (Saji, 2021)

KNN is based on the idea that every data point that is close to another belongs to the same class. In other words, it uses similarity to classify a new data point. In simplier terms, the KNN is a supervised classification algorithm that classifies new data points based on the nearest data points (i.e. K). (Saji, 2021)

The Decision Tree algorithm is part of the supervised learning algorithms family. The decision tree approach, unlike

other supervised learning algorithms, may also be utilised to solve regression and classification issues. (Chakure, 2022)

By learning simple decision rules inferred from past data, the purpose of employing a Decision Tree is to develop a training model that can be used to predict the class or value of the target variable (training data). (Chakure, 2022)

We start from the root of the tree when using Decision Trees to forecast a class label for a record. The values of the root attribute and the record's attribute are compared. We follow the branch that corresponds to that value and jump to the next node based on the comparison. (Chakure, 2022)

Logistic regression is a "supervised machine learning" approach that can be used to model the likelihood of a specific class or occurrence. It is a predictive analysis algorithm and based on the concept of probability. (Pant, 2021)

There are basically two types of Logistic Regression:

- Binary
- Multi-linear

Predicting an output variable that is discrete in two classes is referred to as binary classification. (Pant, 2021)

Yes/No, Pass/Fail, Win/Lose, Cancerous/Non-cancerous, and other binary classifications are examples. (Pant, 2021)

Since what we want to predict is a Yes/No classification (whether they will have diabetes or not), the binary logistic regression will work like a charm. (Pant, 2021)

III. Results Analysis

A. Analysis

We check to see if there are any correlations between the features of our cleaned data-set. We were able to produce the following heat map to help visualize this correlation:

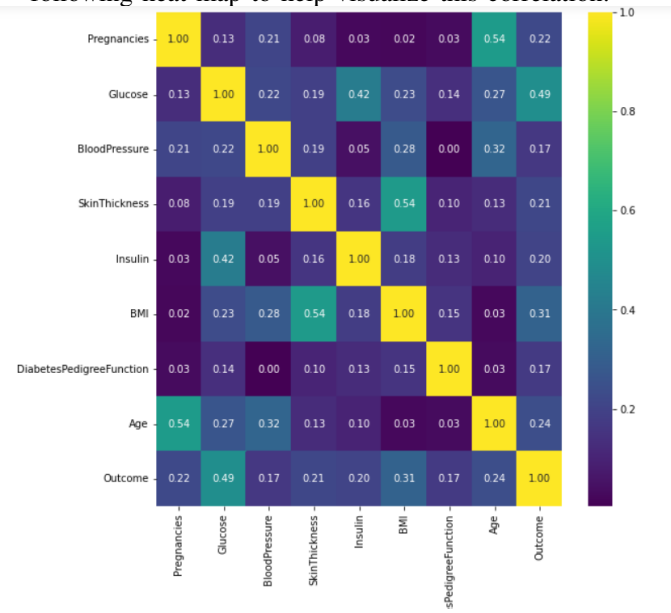

FIG.2: Heat Map of Correlation

As you can see in FIG.2, some features are more correlated than others. For example: BMI and Skin Thickness correlate with a value of 0.54(out of 1) and Skin Thickness and Blood Pressure do not correlate as much (0.19). We also check our data for biases and we found this out:

```
0    500
1    268
Name: Outcome, dtype: int64
```

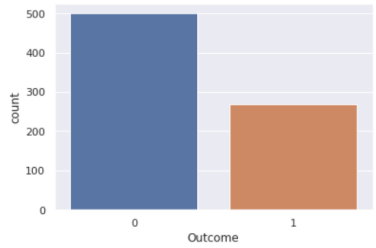

FIG.3: Biases of Data

We can see that there is a bias in the data where the count of patients with no diabetes is almost double the count of patients with diabetes.

B. Pre-Processing

We split our data-set into train and test split our data-set and use stratify to split our data-set.

To make use of this data, some cleaning and tuning and feature selection would need to be done to the data-set we are making use of. To achieve this, we took some of the following steps:

First, we check the data-set for any missing values, and in our first run through python, we did not find any. However, the histogram distribution for the below mentioned features had columns of 0 values, which was an indicative of missing values. We replaced our missing values with the mean or median of the distribution in the test and train data-set. The nulls in the trained dataset before amputation were 475 and in the test dataset were 177, and after our cleaning procedure, the null values were 0 after the amputation.

We now see the correlation in our train dataset can be found in Fig 2.1

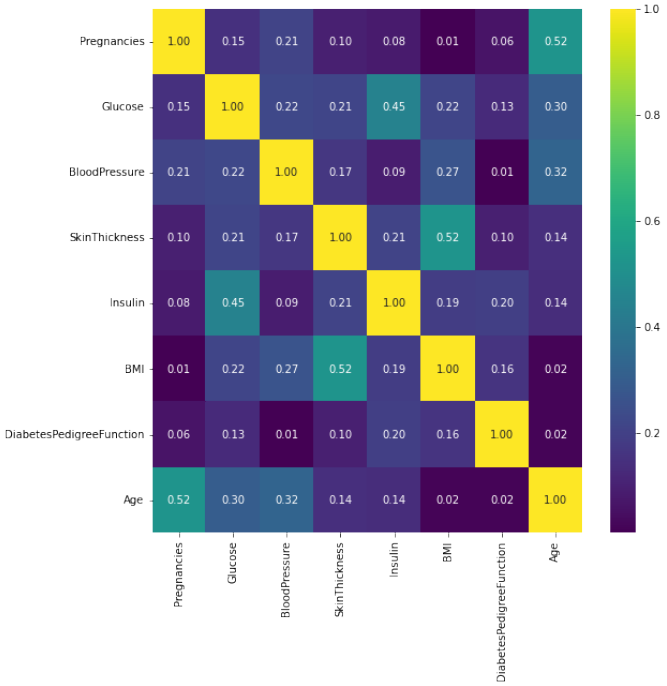

FIG.2.1: Correlation in our train set

The bias was reduced in our dataset with the count of 0 was 375 and count of 1 was 201, as seen in Fig 3.1

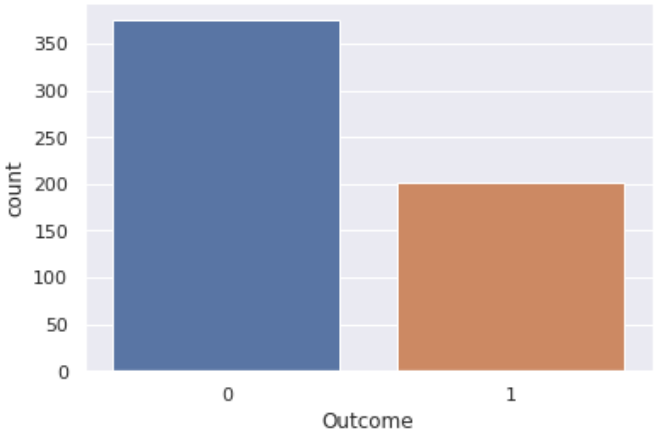

FIG.3.1:Bias in our train set

C. Feature Selection

We made use of a Decision Tree Classifier to select our important features. With the help of this classifier, we were able to produce this bar chart (On the chart, the features are labelled 0-7, 0 for Pregnancies, 1 for Glucose and so on):

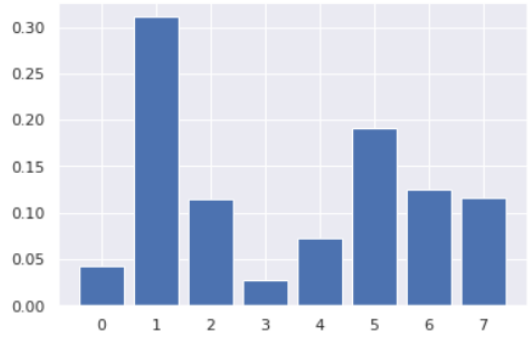

FIG.4: Feature importance through Decision Tree Classifier

From this chart, we can determine that feature Glucose and BMI hold the most importance in our classifier, while Skin Thickness was the least important feature.

D. Feature Scaling

We scaled our train and test set accordingly, using the standard scaler. And the shape of our scaled trained set was (576, 8) and test set was (192, 8).

E. Principal Component Analysis - PCA

Using our scaled dataset, we reduced our dimensions to 2 using Principal Component Analysis. Our variance ration for the first component was 0.29 and the variance ratio of the second component was 0.19, and together they gave us 0.47 of information.

F. Applying Models

A quick overview of the learning models used and the processes used on them. We plan to use the following models:

- KNN classifier
- Decision Tree
- Logistic regression

We first perform the test/train split on our models using a scaled version of our data-set. After training our models, we are provided with these accuracy:

- KNN - With this model our accuracy was 0.73 (73 percent), this means that for every 768 women, our current KNN Model can properly predict the onset of diabetes in 560 of them.
- Decision Tree - With this model our accuracy was 0.72 (72 percent), this means that for every 768 women, our current Decision Tree Model can properly predict the onset of diabetes in 552 of them.
- Logistic regression - With this model our accuracy was 0.70 (70 percent), this means that for every 768 women, our current LR Model can properly predict the onset of diabetes in 360 of them.

These results can be better seen the diagram below:

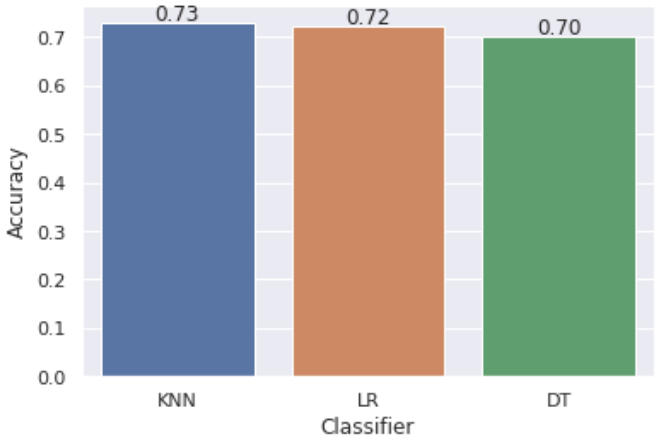

FIG.5: Bar Chart of Initial Model Accuracy

K-fold cross validation

To further try and improve the accuracy of our models, we applied K-fold cross validation. The accuracy of all models changed as follows:

- KNN - 0.73 to 0.76
- Decision tree - 0.72 to 0.78
- LR - 0.70 to 0.67

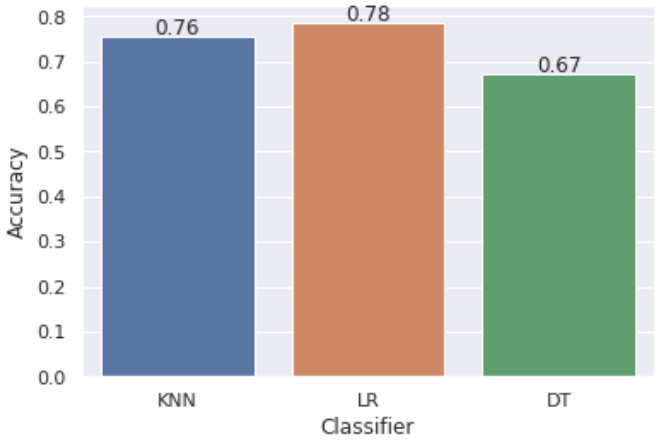

FIG.6: K-Fold cross validation accuracy scores

G. Hyper parameter tuning with Grid Search CV

Decision Classifier

We used GridSearchCV to tune our hyperparameters, namely, max-depth, criterion, and min-sample-leaf for the decision tree classifier. And the results showed the following best parameters max-depth=4, min-samples leaf=25, random state=42.

We run our model with the test set, and we attained an accuracy of 0.74 and F1 score of 0.81 for 0 and 0.60 for 1. We can see the confusion matrix in figure 7.

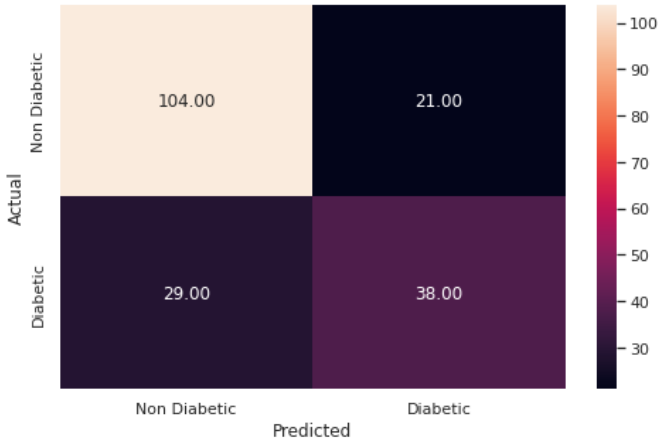

FIG.7: Confusion matrix on test set for Decision Tree

K-Nearest Neighbour

The hyperparameter that was tuned for K-Nearest Neighbour is the K values, that is, the nearest neighbour. As shown in the figure 8, the best K value was 5.

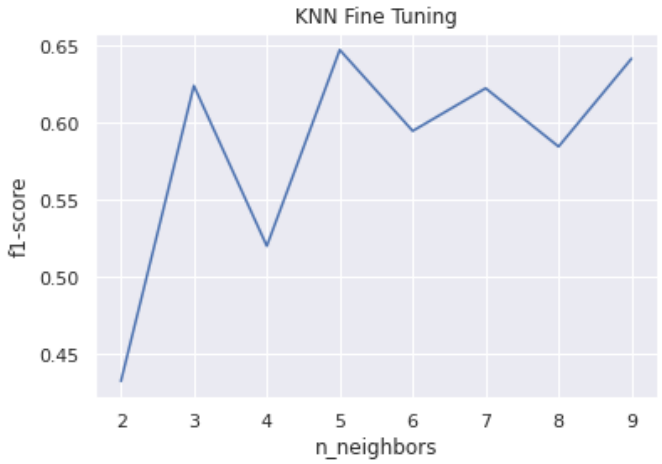

FIG.8: K value optimisation result

With this hyperparamter optimisation, we ran our model on the train and test set and the results of the test set had 0.73 accuracy with F1 score for 0 or non-diabetic to be 0.80 and 1 or diabetic to be 0.73. The confusion matrix for the test set is seen in figure 9.

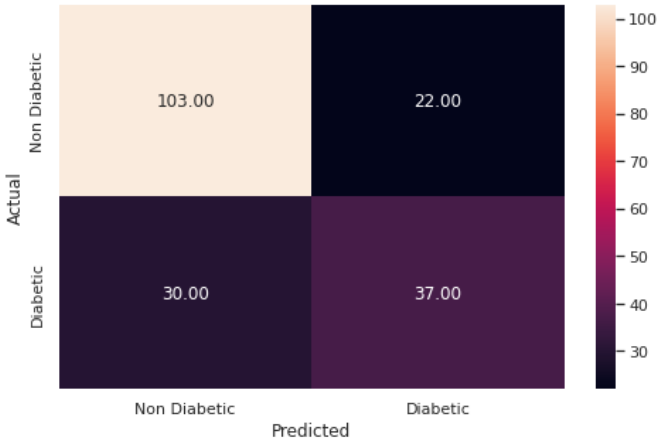

FIG.9: Confusion Matrix of test set for KNN

FIG.11: Confusion Matrix Test set Logistic Regression

Logistic Regression Classifier

We used again the Grid Search CV to optimise C or the regularisation hyperparameter and in our results, as is seen in figure 10, best C value is 0.1

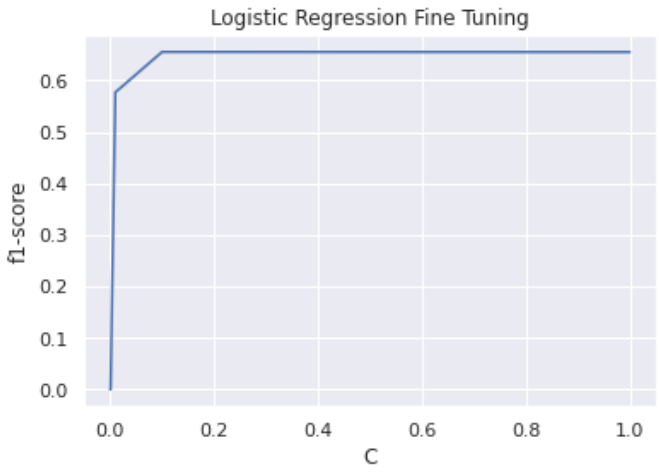

FIG.10: Hyperparameter C optimisation

So, now we use our hyperparameter tuned model on the train and test set and the results were an accuracy score of 0.73 and the with F1 score for 0 or non-diabetic to be 0.80 and 1 or diabetic to be 0.57. The confusion matrix for the test set is seen in figure 11.

H. PCA Dataset

Logistic Regression

We run our hyperparameter tuning for logistic regression for our PCA data-set and the results our C = 1.

So our test accuracy was 0.72 and F1 score for 0 or non-diabetic to be 0.80 and 1 or diabetic to be 0.58. The confusion matrix is seen in Fig 13.

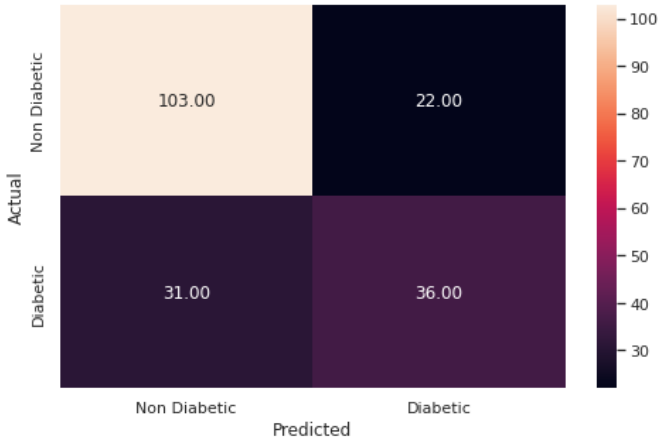

FIG.13: Confusion Matrix for PCA dataset with Logistic Regression

K-Nearest Neighbour

Our hyperparamter tuning for KNN model gave use K = 7 as the best value. Our results on the PCA dataset on the test set has accuracy of 0.67 and the F1 score for 0 or non-diabetic to be 0.75 and 1 or diabetic to be 0.50. The confusion matrix is seen in Fig 14.

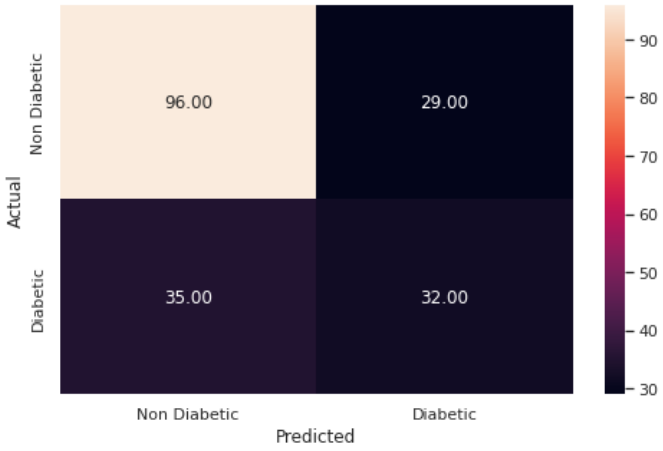

FIG.14: Confusion Matrix for PCA dataset with K-Nearest Neighbour

.

Decision Tree

The Grid Search CV hyperparameter tuning gave us the best parameters as minimum sample leaf to be 20, random value as 42, and minimum depth as 3. Wee run our hyperparamter optimised model on the test PCA data-set and the accuracy score of 0.70 and the F1 score for 0 or non-diabetic to be 0.75 and 1 or diabetic to be 0.61. The confusion matrix is seen in Fig 15.

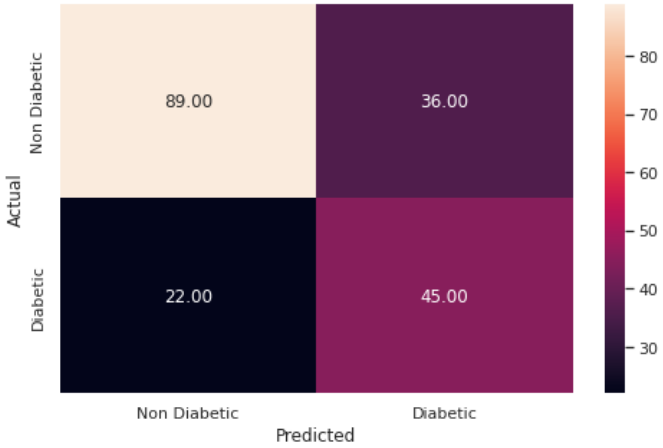

FIG.15: Confusion Matrix for PCA dataset with Decision Tree

.

Model	Accuracy	Precision	Recall	F1 Score
KNN	0.729167	0.627119	0.552239	0.724768
LR	0.729167	0.636364	0.52238	0.722049
DT	0.739583	0.644068	0.567164	0.735354

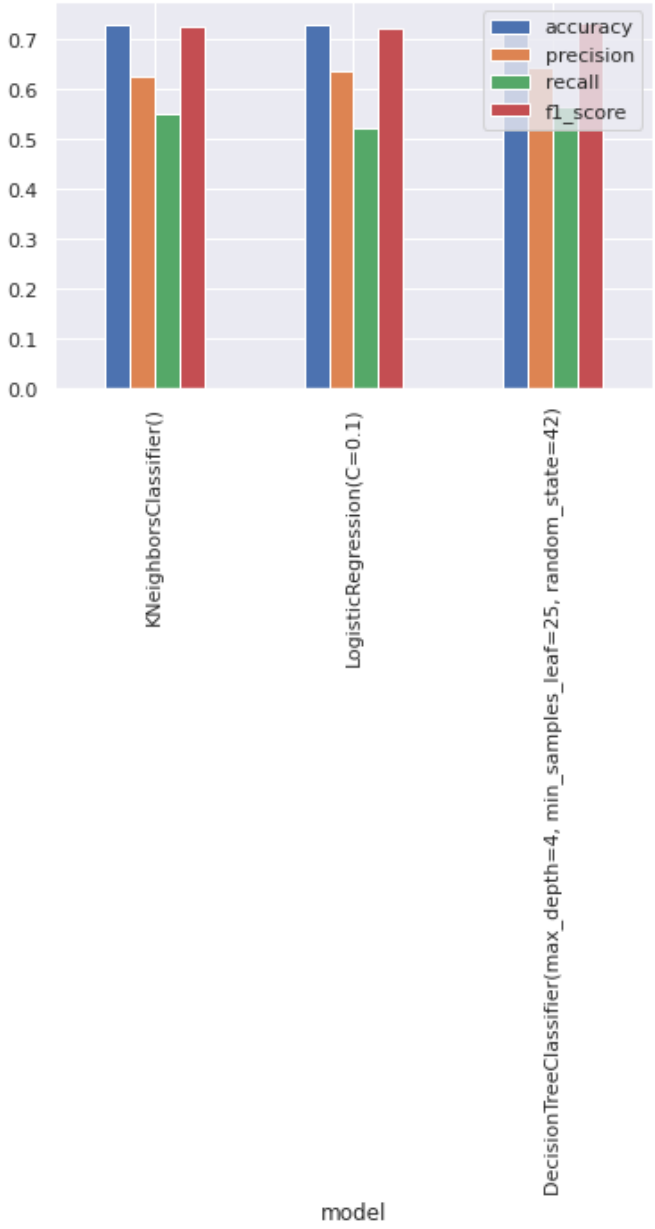

FIG.12: Classification Report Comparison

IV. Conclusion

Because our result is a discrete value, this was a classification problem, as we are predicting whether or not the patient has diabetes. We chose the following structure for our models: KNN, Decision tree, and Logistic Regression. We applied these models and the final result after hyperparamter optimisation on the test dataset can be compared as follows:

Using all this information and graphs, we can see that the decision tree classifier learning model is the best, with a 0.74 accuracy (74 percent).

Later in the future, we intend to research the use of the power of Neural Networks to further improve our accuracy in predicting the onset of Diabetes in individuals. Although this might require more time and funding, we are still pushing forward to making this affordable for anyone to access this technology.

V. Importance and Use of this Study

Diabetes has taken the lives of so many individuals, both young and old. A lot of these people had no idea they had diabetes. If we are able to create a machine that can accurately predict the onset of diabetes in any individual, we might just be able to give them a fighting chance before the disease becomes too much of a hassle.

In 2021, over 10.5 percent of the global adult population had diabetes; by 2045, this percentage is predicted to grow to over 12 percent. Diabetes, often known as diabetes mellitus, is a series of metabolic illnesses characterised by persistently elevated blood sugar levels. Diabetes is currently one of the top ten leading causes of death worldwide, with major health problems such as cardiovascular disease, chronic renal disease, and stroke. (Statista, 2021)

Diabetes is a worldwide illness that affects a wide range of countries. China presently has the most diabetics in the world, with 141 million people suffering from the condition. French Polynesia, Mauritius, and Kuwait, on the other hand, have the highest diabetes prevalence. In many nations, diabetes rates have risen in recent years, as has obesity, one of the primary risk factors for the condition. (Statista, 2021)

Diabetes is expected to continue to be a problem in the future. Between 2021 and 2045, Africa is anticipated to witness a 134 percent increase in diabetes, while North America and the Caribbean are expected to see a 24 percent increase. China is expected to have the biggest number of diabetics in the world by 2045, with the United States accounting for the fourth highest number. (Statista, 2021)

As you can see, Diabetes is a growing world problem, although this study is no cure to it, it can serve as a tool to help curb its spread and inform and warn individuals who have a high risk of getting diabetes.

VI. References

- Saji, B. (2021, January 22). K Nearest Neighbor Classification Algorithm — KNN in Python. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/01/a-quick-introduction-to-k-nearest-neighbor-knn-classification-using-python/#:%7E:text=This%20article%20concerns%20one%20of,This%20gives%20a%20competitive%20result.
- Chakure, A. (2022, February 10). Decision Tree Classification - The Startup. Medium. https://medium.com/swlh/decision-tree-classification-de64fc4d5aac
- Statista. (2021, December 15). Diabetics prevalence worldwide 2021, 2030. and 2045. https://www.statista.com/statistics/271464/percentage-of-diabetics-worldwide/#:%7E:text=Around%2010.5%20percent%20of%20the,chronic%20high%20blood%20sugar%20levels.
- Pant, A. (2021, December 7). Introduction to Logistic Regression - Towards Data Science. Medium. https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148