Dataset Name: Pima Diabetes Set

Group No.: 5 On Campus/cloud: CLOUD

STUDENT ID	STUDENT FULL NAME	Individual contribution *
218403172	GEORGE DAVID NICHOLSON	5
219352756	ANTHONY TRINH	5

* 5 – Contributed significantly, attended all meetings
4 – Partial contribution, attended all meetings
3 – Partial contribution, attended few meetings
1 – No contribution, attended few meetings
0 – No contribution, did not attended any meetings

NOTE: IF ANY OF THE CELLS IN INDIVIDUAL CONTRIBUTION MARK IS EMPTY ALL STUDENTS WOULD GET 3 MARK BY DEFAULT

(Page count starts here & Minimum font size – 11; Maximum font size - 12)

Section 1: Brief Summary & ML Problem Formulation (max 2 pages)

Introduction

This report explores the findings of the Pima Indian people and diabetes within the community. The main issue in exploring this dataset is diabetes and its causes presented within the data processed in the dataset being explored. From the first report, as seen in Figure 13[1], after calculating the sum of people within the dataset to be 768, we found only 268 presented to be diabetic and the remaining 500 people to be non-diabetic. Already from the dataset, we can deduce a clear bias toward data of those without diabetes, leaving less than 35% of the dataset compared to finding the critical links to diabetes within the Pima population. From this data, we also had to account for the anomalies within the raw data that needed to be processed and cleaned, which shows an already biased and anomalous dataset on which we are required to base our findings. As mentioned in the previous report, we used various data cleansing methods to combat the bias to give us a clean dataset with properly calibrated central tendencies depending on the skewness[2]. This is highlighted in Figure 14[3] and Figure 15[4] when comparing the visual change present in the Insulin and Skin Thickness independent variables data columns. Despite the apparent bias and issues with the data set, The dataset was salvageable and ready to proceed with the exploration phase, in addition to enabling bias-free findings and the ability to create a machine learning model to predict and find the leading causes of diabetes within the Pima Population with more accuracy due to the data cleansing.

The model types selected to be run alongside the dataset were Classification based models. This is due to the fact that "Classification methods aim at identifying the category of a new observation among a set of categories on the basis of a labeled training set. Depending on the task, anatomical structure, … and features the classification accuracy varies."[5]. As mentioned, the Classification based models were to be best suited for our dataset as this category of machine learning algorithms "perform a common task of recognizing objects and demonstrate the ability to separate them into categories successfully."[6], thus since "Classification is the task of predicting a discrete class label"[7] and "Regression is the task of predicting a continuous quantity."[8] then it is clear that going with Classification based models was the better choice for our given dataset featuring the discrete Outcome. The results required being prediction-based. It was decided that from this scientific point of view, the best type of machine learning model to develop that our system is to be based on would be the classification type machine learning algorithms when building the machine learning model. This is simply due to the data types of the variables within the dataset with the Outcome. Thus independent variable being of a discrete nature, classification type algorithms were best suited towards processing our dataset into reliable and helpful information.

Below can be seen the machine learning flowchart depicting the step by step process taken depending on what kind of data is held within the dataset and what kind of outcomes we are looking for from processing the data.

---

[1] (George Nicholson and Anthony Trinh, 2022e)
[2] (*Central tendency: Mean, median and mode*, 2020)
[3] (George Nicholson and Anthony Trinh, 2022i)
[4] (George Nicholson and Anthony Trinh, 2022d)
[5] (Jimenez-del-Toro *et al.*, 2017)
[6] (*5 Best Machine Learning Classification Algorithms + Real-World Projects*, 2022)
[7] (Brownlee, 2017)
[8] (Brownlee, 2017)

Section 2: Results and Discussion (max 7 pages)

Machine Learning model

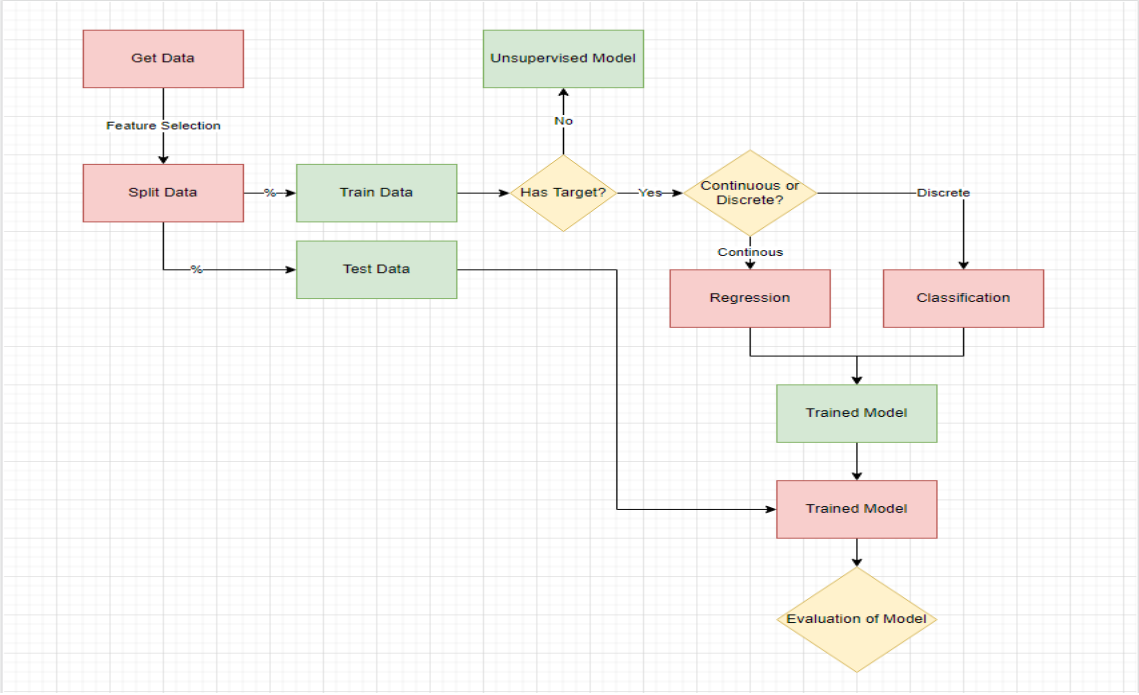


Figure 1 - Generic machine learning model flowchart

Since the given dataset at hand, as described, includes the use of continuous dependant variables and a discrete independent outcome variable, and thus we know we will be using a supervised machine learning model to predict the Outcome of diabetes, the flowchart below is more centred around our dataset and machine learning model.

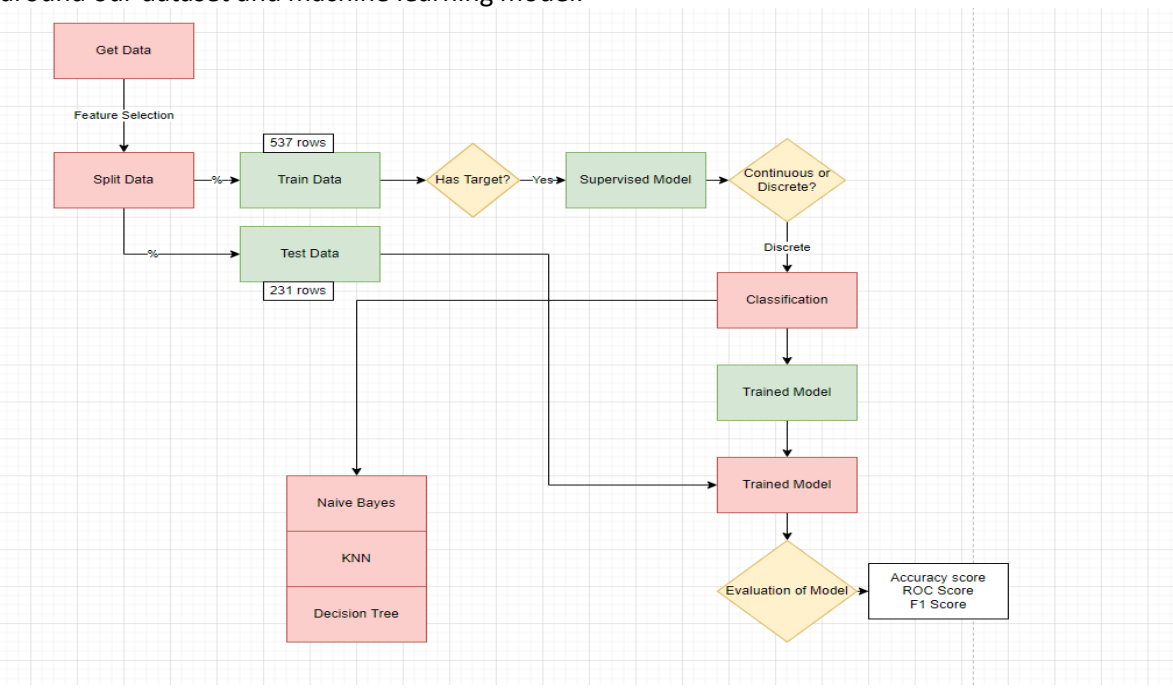


Figure 2 - Discrete data machine learning process

This process above shows the data splitting into testing and training data to be processed through the classification algorithms Naïve Bayes, KNN and Decision Trees. We chose these algorithms because "The Naïve Bayes algorithm quickly predicts the class of the test data set. Moreover, it also performs accurately in a multi-class prediction scenario."[9] And because using "a Naive Bayes classifier performs better than other models like logistic regression. It works with lesser training data too."[10] Our limited dataset of the Pima females population we were given and needed to process proved practical. On top of this, the decision to use K-Nearest Neighbours as also "The k-nearest neighbours (KNN) algorithm is a data classification method"[11] which is best suited towards our dataset since "K-nearest neighbours (k-NN) is a pattern recognition algorithm that uses training datasets to find the k closest relatives"[12] which is useful when making predictions on a dependent outcome using independent variables. And lastly, we have the implementation of decision tree-based classification algorithms as a "decision tree is a supervised learning algorithm that is perfect for classification problems, as it's able to order classes on a precise level."[13]

Feature Selection

We needed a supervised feature selection method for our dataset because the dataset has a dependent target variable, Outcome. There are three different methods for deciding the best features for our dataset, Filter, Wrapper and Embedded. We have chosen the Filter method as it is a statistical-based approach that measures the components by their correlation with the dependent variable. That would be the correlation strength between the two best selected independent variables concerning the causation towards the dependant 'outcome' variable. Another reason why we chose the Filter method is because it has a low computational cost and has a low risk of overfitting compared to other methods. We use ANOVA, Chi-Square Test, and Information Gain to find out the best feature within our dataset. We gained from exploratory analysis (first report) that glucose and BMI had the highest correlation to a person having diabetes. The ANOVA test showed that 'Glucose' and 'BMI' were the two best features decided upon in this feature selection/filter method of features.

	Glucose	BMI
0	148.0	33.6
1	85.0	26.6
2	183.0	23.3
3	89.0	28.1
4	137.0	43.1

Figure 3- First 5 rows of the best feature[14]

While on the other hand, the Chi-square Test and Information Gain displayed 'Glucose' and 'Insulin' to be the best two features.

[9] (*5 Best Machine Learning Classification Algorithms + Real-World Projects*, 2022)
[10] (*5 Best Machine Learning Classification Algorithms + Real-World Projects*, 2022)
[11] (*5 Best Machine Learning Classification Algorithms + Real-World Projects*, 2022)
[12] (*5 Types of Classification Algorithms in Machine Learning*, 2020)
[13] (*5 Types of Classification Algorithms in Machine Learning*, 2020)
[14] (George Nicholson, 2022)

	Glucose	Insulin
0	148.0	120.0
1	85.0	120.0
2	183.0	120.0
3	89.0	94.0
4	137.0	168.0

Figure 4 - first 5 rows of best feature selection for chi-square and information gain[15]

Even though we have different best features related to 'BMI' and 'Glucose' depending on the filter method used, it was decided that the implementation of the two best features presented by the Anova feature selection method was to be used within our machine learning model. This is simply due to the fact that "Feature selection improves the machine learning process and increases the predictive power of machine learning algorithms by selecting the most important variables and eliminating redundant and irrelevant features."[16] And considering the dataset, input variables are numerical. At the same time, the output is categorical, and ANOVA's linear correlation coefficient is stated to be the best feature selection method for our given dataset conditions and datatypes by Kartik Menon, as mentioned in his tutorial on 'Everything you need to know about feature selection in machine learning.'[17]

So, it was decided to stick to 'Glucose' and 'BMI' to be our best features as matched by the analysis from the first report and our research into the best feature selection algorithm to use, which proved to be ANOVA. Our findings from the ANOVA feature selection gave a different perception of the links related to diabetes, as diabetes should have a direct link to Insulin or glucose levels. However, since the ANOVA method is the best suited for our dataset, the decision to use the 'Glucose' and 'BMI' independent variables as our best features stems from the fact that since ANOVA is the best-suited feature selection method for our dataset, the ANOVA algorithm results may lead us to more accurate findings as opposed to going with the apparent Insulin and glucose levels indicated from the Chi-square Test and Information Gain algorithms used to process the dataset. Thus, the best top features presented were 'BMI' and 'Glucose'.

Classification

Within the realm of Data Classification concerning machine learning, "The process starts with predicting the class of given data points. The classes are often referred to as target, label or categories."[18] With this in mind, it was clear that data classification was the next phase in our dataset exploration. We are to predict our independent outcome category using the best features selected to be imputed into the machine learning process.

We used the following code to split the dataset into training and testing data:

[15] (George Nicholson and Anthony Trinh, 2022g)
[16] (*What is Feature Selection? Definition and FAQs | HEAVY.AI*, no date)
[17] (Kartik Menon, no date)
[18] ('Classification In Machine Learning | Classification Algorithms', 2019)

We will be using classification models as we are trying to predict someone has diabete or not (True or False)

```
In [131]: bestFeature1 = data_clean_copy[['Glucose', 'BMI']] #First Best Feature
          bestFeature2 = data_clean_copy[['Glucose', 'Insulin']] #Second Best Feature

In [132]: Xtrain, Xtest, ytrain, ytest = train_test_split(bestFeature1, y, test_size=0.3, random_state=42)
          print (Xtrain.shape, ytrain.shape)
          print (Xtest.shape, ytest.shape)

          (537, 2) (537,)
          (231, 2) (231,)
```

Figure 5 - training and testing data[19]

Concerning the classification methods used within the machine learning model. The classification models are KNN, Naïve Bayes, and Decision Trees. KNN is used as "a simple, supervised machine learning algorithm that can solve classification and regression problems. It's easy to implement and understand""Naive Bayes uses a similar method to predict the probability of different classes based on various attributes. This algorithm is mostly used in text classification and multiple class problems."[20] And "Decision trees tend to be the method of choice for predictive modelling because they are relatively easy to understand and very effective."[21] This left the process to determine which classification algorithms were the most precise after predicting within the machine learning model.

Evaluation Metrics

To determine which model performance is the best on our dataset, we must perform different metrics on the models. We used to use more than one evaluation metric as a model that can perform well using one measurement from one evaluation metric but may perform poorly using another measure from another evaluation metric. The main evaluation metrics we will use are accuracy, AUC/ROC, confusion matrix, and classification report (Precision, recall and f1-score).

Accuracy

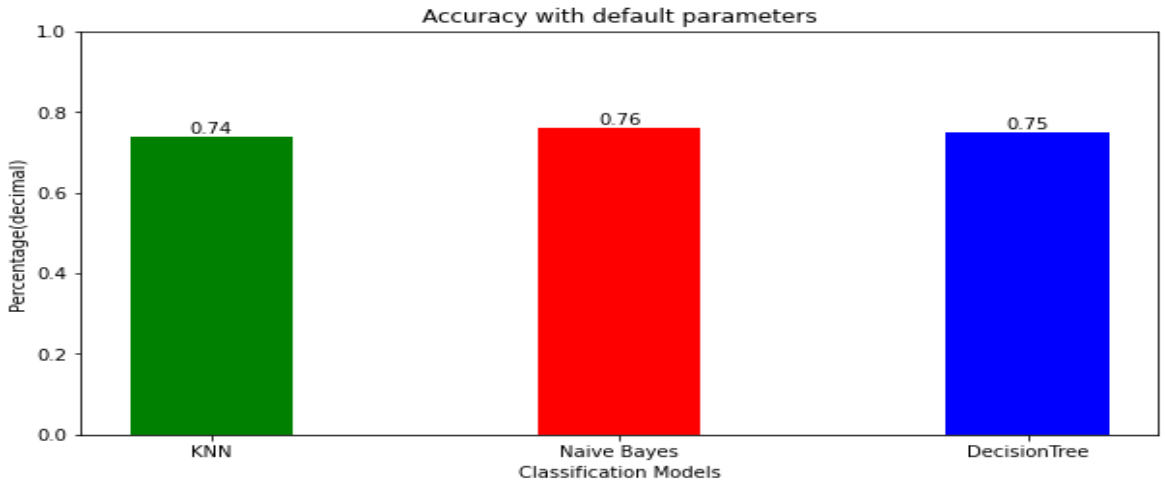


Figure 6 - Accuracy with default parameters[22]

Accuracy is measured by $\frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predicitions}$. This graph above shows that all three classification models have similar accuracy, demonstrating that using only accuracy as our evaluation metric isn't enough and that more evaluation metrics are needed.

[19] (George Nicholson and Anthony Trinh, 2022l)
[20] (*Learn Naive Bayes Algorithm | Naive Bayes Classifier Examples*, no date)
[21] ('Decision Trees: An Overview', 2015)
[22] (George Nicholson and Anthony Trinh, 2022a)

AUC/RUC Score

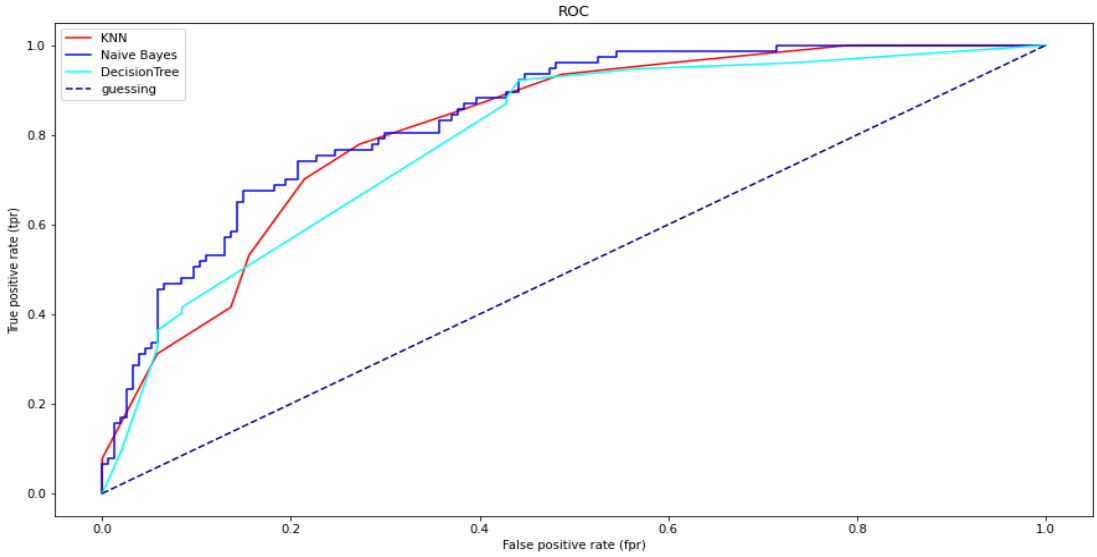


Figure 7 - AUC/ROC Score[23]

```
ROC Score
-------------------------
KNN: 81.0%
Naive Bayes: 84.0%
DecisionTree: 79.0%
```

Figure 8 - ROC Score[24]

The AUC-ROC curve is a performance measure of a classification problem with various threshold settings. ROC is a probability curve, and AUC is a measure or measure of separability. This shows the extent to which the model can distinguish between classes. The higher the AUC, the better the model predicts 0 class as 0 and 1 class as 1. For our cause, Naïve Bayes performs slightly better than both models.

[23] (George Nicholson and Anthony Trinh, 2022c)
[24] (George Nicholson and Anthony Trinh, 2022k)

Precision

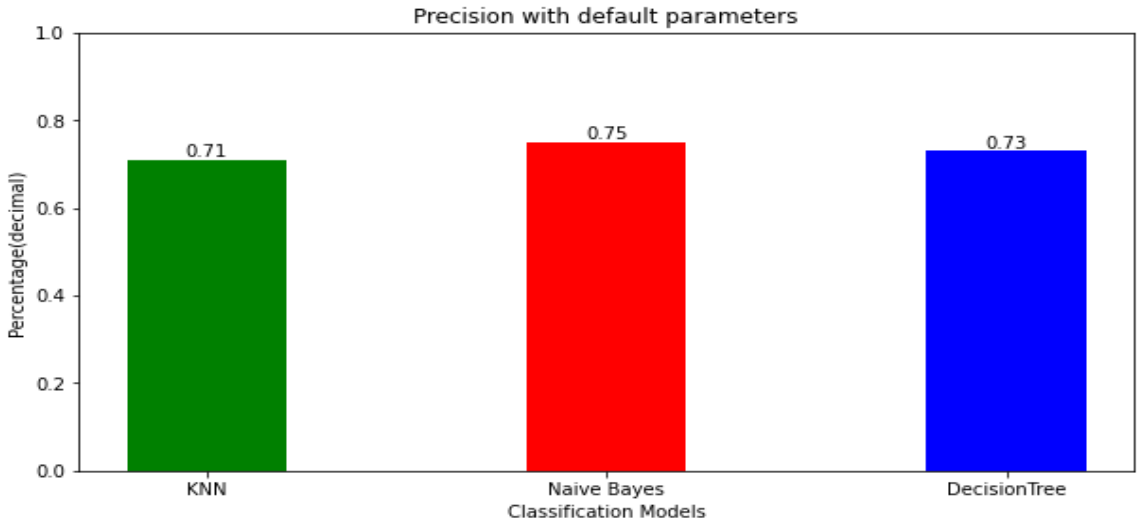


Figure 9 - Precision with default parameters[25]

Precision is the ratio between the True Positives and all the Positives. These four indications of the predictions are right or wrong are:

- TN/ True Negative: when a case was negative and predicted negative
- TP/ True Positive: when a case was positive and predicted positive
- FN/ False Negative: when a case was positive but predicted negative
- FP/ False Positive: when a case was negative but predicted positive

Again, Native Bayes performs better than the other two models.

Recall

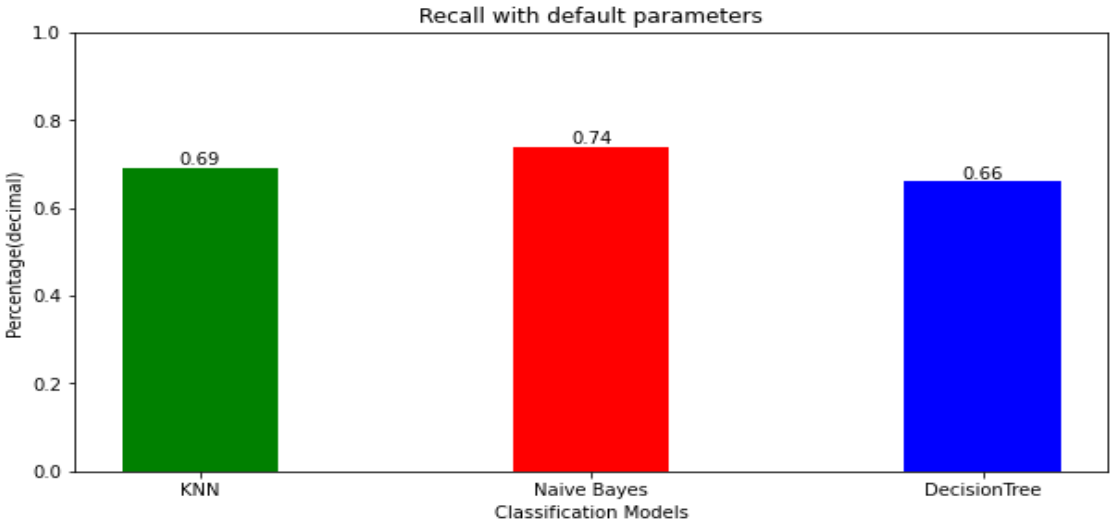


Figure 10 - recall with default parameters[26]

The recall is the measure of our model correctly identifies True Positives. Here, Naïve Bayes was able to identify true positives much better than the other two models.

[25] (George Nicholson and Anthony Trinh, 2022h)
[26] (George Nicholson and Anthony Trinh, 2022j)

F1 - Score

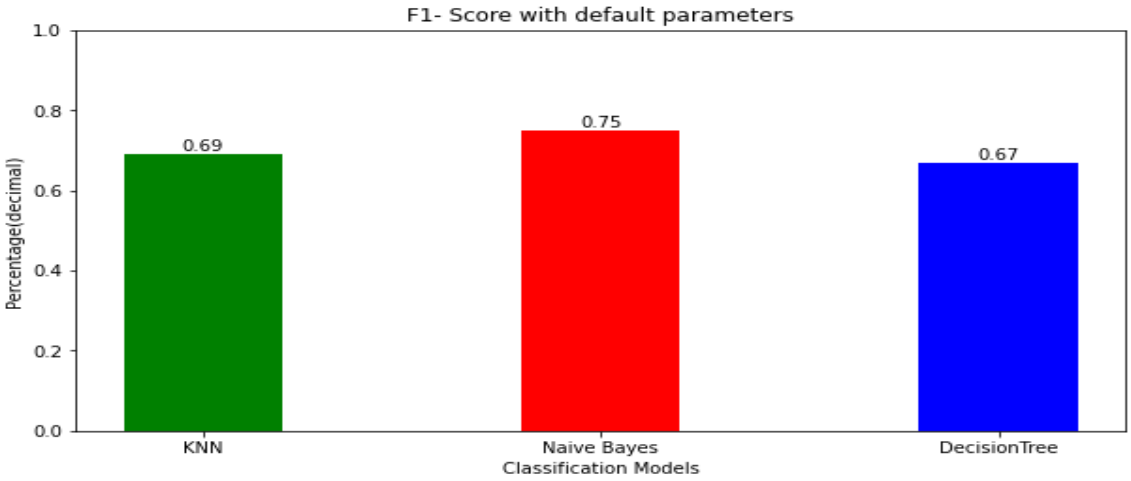


Figure 11 - f1 score with default parameters[27]

F1 – Score is the measure of Precision and recall combined. This represents that overall Native Bayes outperforms both KNN and Decision Tree. As a result, the naive Bayes-based classifier model is a powerful and powerful predictor. They solve diagnostic and predictive problems and provide a valuable perspective for understanding and evaluating many learning algorithms. Explicitly calculate Robust against hypothetical probabilities and noise Input data. That's why; a naive Bayes-based classifier Models improve accuracy in predicting diabetes This study.

Hyperparameters

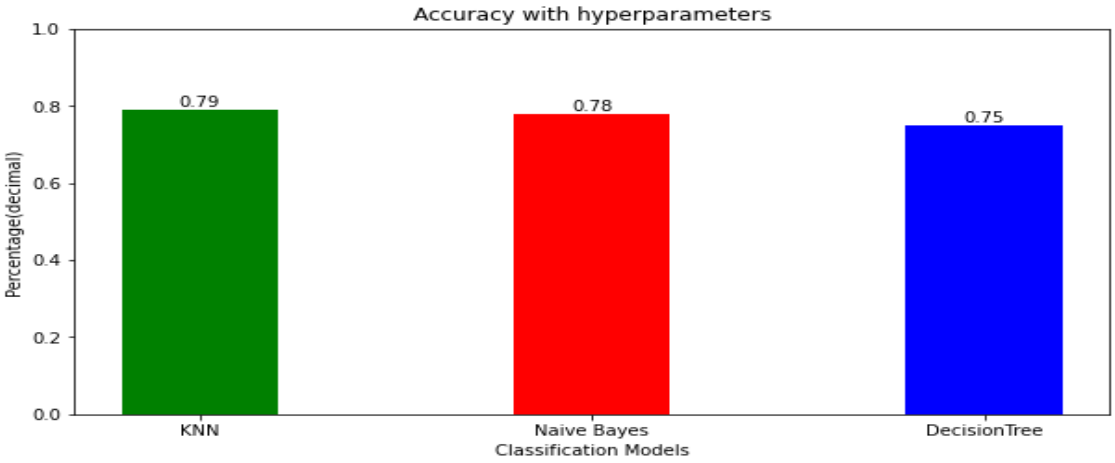


Figure 12 - Accuracy with Hyperparameters[28]

Hyperparameters it the process of changing the parameters of a machine learning model. By tuning different settings on our dataset, we can get different results. There are three methods of tuning the parameters. They are:

- By Hand
- RandomSearchCV
- GridSearchCV

For our dataset, it was concluded that we use GridSearchCV as it is a brute force method that finds all the combinations of the parameters that we passed in. We increased the KNN and Naïve Bayes using hyperparameters, but the decision tree didn't improve.

[27] (George Nicholson and Anthony Trinh, 2022f)
[28] (George Nicholson and Anthony Trinh, 2022b)

Section 3: Conclusions (max 1 page)

From our exploration of the given dataset and the processing within our machine learning models, we were able to predict whether or not a person, based on the medical information provided, had diabetes using all of the evaluation metrics to help evaluate the performance of a given model, these metrics being f1Score, Precision and recall, these metrics for each classification algorithm were compared, and in each of the comparison, Naïve Bayes Classification model exceeded the other classification models in all accuracy scores and as this was the most accurate classification model used within the dataset. This classification model presented the findings of 173 accurate predictions out of 231 people.

With the hopes of making our model more accurate and thus achieving more accurate results, the obvious flaw in our dataset was the initial inaccuracies and reduction in our current dataset to achieve accurate predictions. With more medical outcomes, the more precise we could create our machine learning model. Since we had to Impute and average missing data to salvage the dataset, this hindered the model accuracy by an amount unknown but still workable to achieve an overall accuracy of 75%. Regarding our current model, the additions of hyperparameters and other parameters would be beneficial possibly with predicting in this dataset. This could be done by hand, which would've taken an extremely long time in comparison to other methods; however, also by random search and grid search, we went with grid search as a way to increase our model; however, the more parameters added, the longer it takes to find the results due to the increase of the number of combinations needed to be run through the algorithm.

- Explain what you could achieve by running ML models on your dataset, and was it helpful to solve your problem?
- Any suggestion on improving your model and achieving better results?

Section 4: References

5 Best Machine Learning Classification Algorithms + Real-World Projects (2022) *Omdena | Building AI Solutions for Real-World Problems*. Available at: https://omdena.com/blog/machine-learning-classification-algorithms/ (Accessed: 18 May 2022).

5 Types of Classification Algorithms in Machine Learning (2020) *MonkeyLearn Blog*. Available at: https://monkeylearn.com/blog/classification-algorithms/ (Accessed: 18 May 2022).

Brownlee, J. (2017) 'Difference Between Classification and Regression in Machine Learning', *Machine Learning Mastery*, 10 December. Available at: https://machinelearningmastery.com/classification-versus-regression-in-machine-learning/ (Accessed: 18 May 2022).

Central tendency: Mean, median and mode (2020) *Scribbr*. Available at: https://www.scribbr.com/statistics/central-tendency/ (Accessed: 20 May 2022).

'Classification In Machine Learning | Classification Algorithms' (2019) *Edureka*, 4 December. Available at: https://www.edureka.co/blog/classification-in-machine-learning/ (Accessed: 21 May 2022).

'Decision Trees: An Overview' (2015) *Aunalytics*, 31 January. Available at: https://www.aunalytics.com/decision-trees-an-overview/ (Accessed: 22 May 2022).

George Nicholson (2022) *glucose and bmi from ANOVA*.

George Nicholson and Anthony Trinh (2022a) *Accuracy with default parameters*.

George Nicholson and Anthony Trinh (2022b) *Accuracy with Hyperparameters*.

George Nicholson and Anthony Trinh (2022c) *AUC/RUC score*.

George Nicholson and Anthony Trinh (2022d) *Computed N/A values Central Tendency of distribution*.

George Nicholson and Anthony Trinh (2022e) *Diabetic vs Non-Diabetic*.

George Nicholson and Anthony Trinh (2022f) *f1 score with default parameters*.

George Nicholson and Anthony Trinh (2022g) *Glusocse and Insulin*.

George Nicholson and Anthony Trinh (2022h) *Precision with default parameters*.

George Nicholson and Anthony Trinh (2022i) *Precomputed dataset Central Tendency of distribution*.

George Nicholson and Anthony Trinh (2022j) *recall with default parameters*.

George Nicholson and Anthony Trinh (2022k) *ROC Score*.

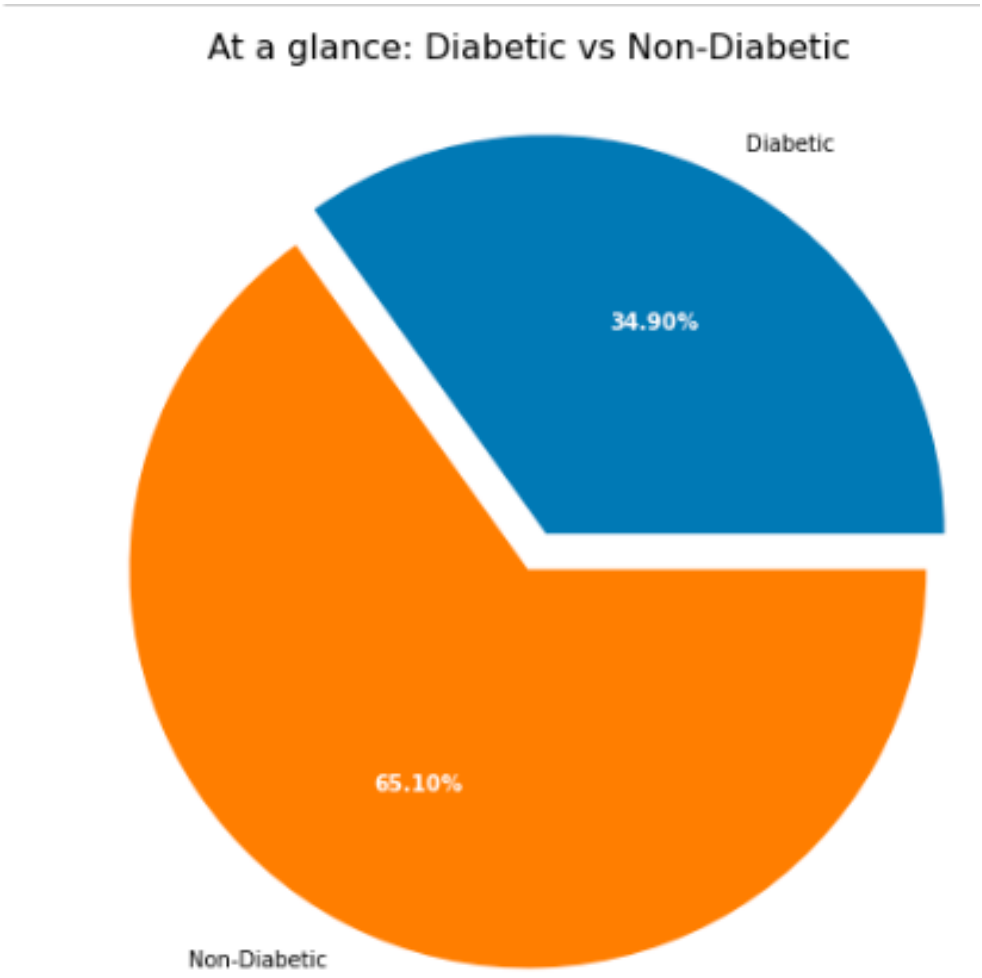
George Nicholson and Anthony Trinh (2022l) *training and testing data*.

Jimenez-del-Toro, O. *et al.* (2017) 'Chapter 10 - Analysis of Histopathology Images: From Traditional Machine Learning to Deep Learning', in Depeursinge, A., S. Al-Kadi, O., and Mitchell, J.R. (eds) *Biomedical Texture Analysis*. Academic Press (The Elsevier and MICCAI Society Book Series), pp. 281–314. doi:10.1016/B978-0-12-812133-7.00010-7.

Kartik Menon (no date) *Feature Selection In Machine Learning [2021 Edition] - Simplilearn*, *Simplilearn.com*. Available at: https://www.simplilearn.com/tutorials/machine-learning-tutorial/feature-selection-in-machine-learning (Accessed: 21 May 2022).

Learn Naive Bayes Algorithm | Naive Bayes Classifier Examples (no date). Available at: https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/ (Accessed: 22 May 2022).

What is Feature Selection? Definition and FAQs | HEAVY.AI (no date). Available at: https://www.heavy.ai/technical-glossary/feature-selection (Accessed: 21 May 2022).




```
Out of 768 subjects, 268 have diabetes, and 500 do not
```

Figure 13 - Diabetic vs Non-Diabetic

```
print(data_clean_copy.isnull().sum())
```

```
Pregnancies                 4
Glucose                     5
BloodPressure              35
SkinThickness             228
Insulin                   392
BMI                        14
DiabetesPedigreeFunction   11
Age                         5
Outcome                     0
dtype: int64
```

```
In [27]: p = data_clean.hist(figsize = (25,25))
```

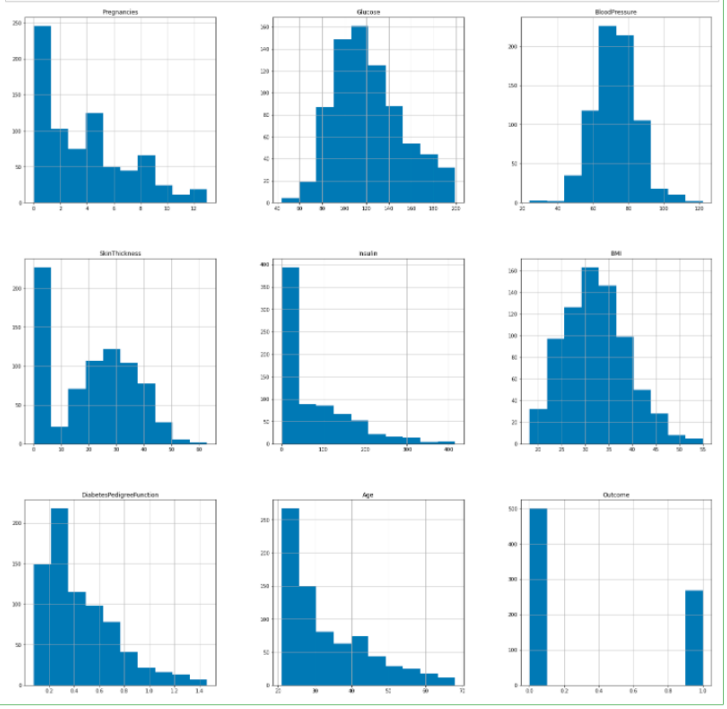


Figure 14 - Precomputed dataset Central Tendency of distribution

```
In [29]: data_clean_copy.head()
```

Out[29]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6.0	148.0	72.0	35.0	120.0	33.6	0.627	50.0	1
1	1.0	85.0	66.0	29.0	120.0	26.6	0.351	31.0	0
2	8.0	183.0	64.0	29.0	120.0	23.3	0.672	32.0	1
3	1.0	89.0	66.0	23.0	94.0	28.1	0.167	21.0	0
4	0.0	137.0	40.0	35.0	168.0	43.1	0.366	33.0	1

```
In [96]: nan_metric_plot = data_clean_copy.hist(figsize = (25,25))
```

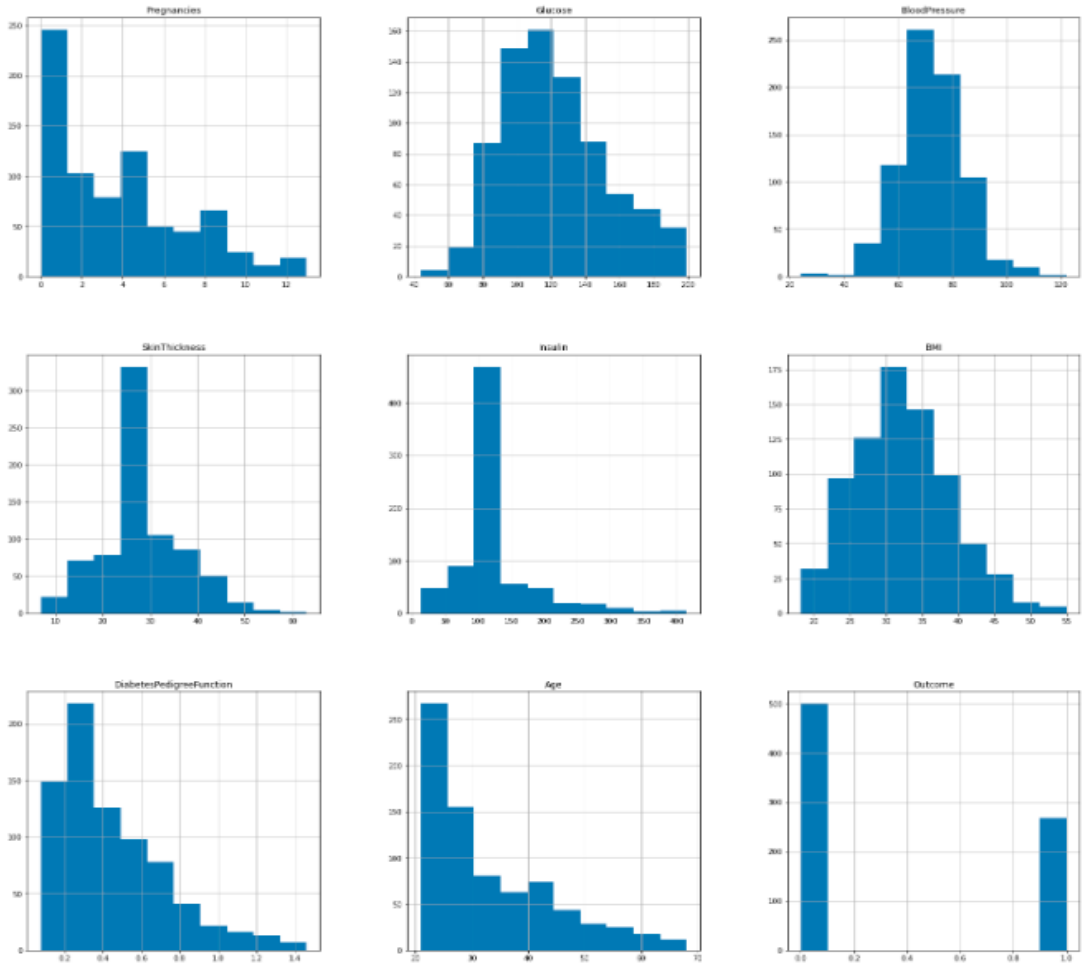


Figure 15 - computed N/A values Central Tendency of distribution