A Comparison Of Naïve Bayes And Decision Tree For The Prediction Of Diabetes

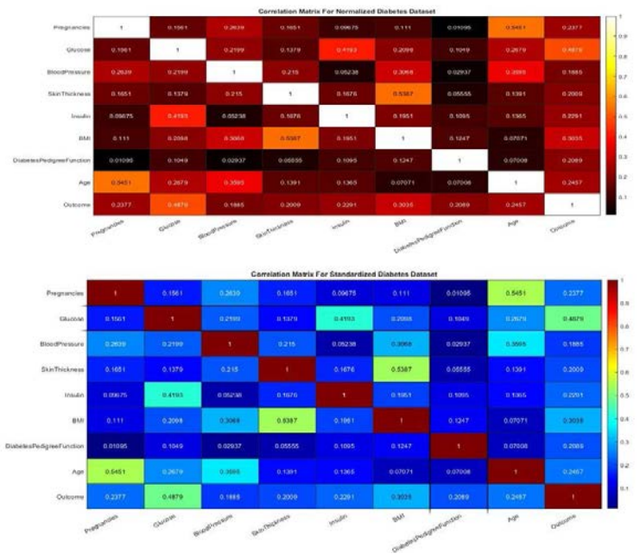
INM431 Machine Learning Coursework Poster **Name: Andreas Ioannides** **Email: Andreas.Ioannides@city.ac.uk**

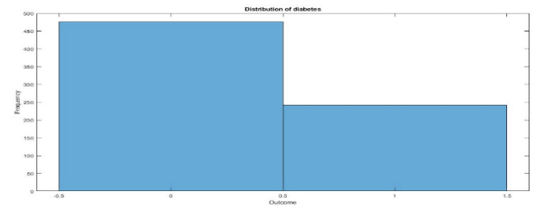
Description and Motivation:

This project aims to construct two supervised machine learning algorithms (Naïve Bayes and Decision Tree) to solve the classification problem of diabetes prediction. Performing these supervised algorithms on the publicly available Pima Indians Diabetes Dataset, we seek to compare and evaluate their performance. Diabetes is a long-term disease that affects a large number of the global adult population, thus the selected algorithms have been chosen due to their consistent effectiveness in classification tasks within the healthcare industry.[1]

Initial Analysis of Dataset:

For this project analysis, the Pima Indians Diabetes Dataset was found on Kaggle but was originally taken from the National Institute of Diabetes and Digestive and Kidney Diseases.[2] The dataset consists of 768 female patients aged from 21-81 years old. The dataset is formed by 8 feature variables which are the number of Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, and Age, and one target variable which is Outcome. The target variable is identified with a 1 if the patient is diagnosed with diabetes and a 0 if the patient is not diagnosed with diabetes.
Starting the project with some exploratory data analysis, the first thing we did was identify the unrealistic zero values in some of our feature variables. We decided to replace the 0 values with the corresponding median of each column without considering the 0 in our calculation. We then continued by inspecting if our data had any outliers and removed them, as well as visualizing all of our variables in a histogram which we can see to the right of this poster
When working with the Naïve Bayes model, we normalized our variables due to the normality assumption, but we did not normalize the pregnancies column since it is a count variable and did not normalize the outcome columns since it is binary. When working with the decision tree model, we standardized our variables to maximize model performance for our decision tree whilst again not standardizing the pregnancies and outcome columns.
Finally, we created a heatmap and summary statistics for both the normalized and standardized datasets which can be found to the right of this template
We can see that our dataset is imbalanced by classifying more patients without diabetes than with diabetes





Normalized Variables	Mean	Median	Standard Deviation
Pregnancies	4.3914	4	2.8935
Glucose	0.49039	0.45806	0.19138
Blood Pressure	0.48963	0.48571	0.16193
Skin Thickness	0.46204	0.46809	0.17356
Insulin	0.30585	0.29178	0.1474
BMI	0.40152	0.39914	0.18669
Diabetes Pedigree Function	0.26858	0.20607	0.20457
Age	0.25363	0.17021	0.2408

Standardized Variables	Mean	Median	Standard Deviation
Pregnancies	4.3914	4	2.8935
Glucose	2.28E-16	-0.16889	1
Blood Pressure	2.56E-16	-0.024206	1
Skin Thickness	1.21E-16	0.034831	1
Insulin	1.61E-16	-0.095467	1
BMI	-4.75E-16	-0.01277	1
Diabetes Pedigree Function	-2.66E-17	-0.30557	1
Age	-1.37E-16	-0.34642	1

Naïve Bayes

Naïve Bayes is a supervised learning algorithm that is mainly used for solving classification tasks such as our case with diabetes prediction[3]
The most distinct factor of Naïve Bayes is that it holds a strong assumption of independence, meaning it assumes that every single feature variable is independent not only from each other but also does not have any direct effect on the target variable
Naïve Bayes also assumes that no feature variable is more important than another and that all of the features have the same contributions to the outcome variable

Advantages:
Naïve Bayes has very fast and accurate calculations helping it perform better than the other algorithms[4]
Naïve Bayes also performs better with smaller datasets such as the Pima Indians Diabetes dataset that we are working with

Disadvantages:
When dealing with real-life data, there is almost always some form of dependence between the feature variables so the assumption of independence is violated[5]
Naïve Bayes is not a suitable algorithm if there is multicollinearity between the feature variables
Naïve Bayes does not perform well when the data is noisy and has problems with overfitting

Decision Tree

The Decision Tree algorithm also falls under the category of supervised machine learning algorithms and it performs well in both classification and regression tasks
Decision Trees consist of branches, nodes, and leaves. Branches represent different commands of the decision tree, nodes represent the feature variables, and leaves represent the 0-1 binary outcome. The latter is one of the reasons why we have used it for our classification task[6]

Advantages:
Decision Trees are both easy to perform and interpret. This is mainly because we can draw the decision tree graph and already have an estimate of what outcome to expect
Decision Trees work well with categorical data as well as numerical data so they have fewer limitations on what they can perform
They can be performed instantly without any need to normalize, standardize, or make any significant changes to the processing of the data[7]

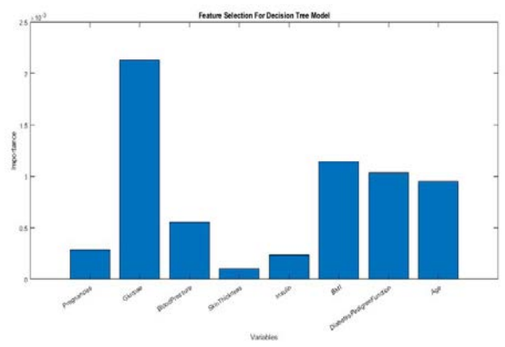
Disadvantages:
Decision trees can have a lengthy time of training, testing, and optimizing the model which does not make it suitable if we want to perform a fast task
If our dataset has problems with overfitting, then the decision tree algorithm will not be relevant and can easily provide inaccurate results
Decision trees perform better with larger and more complex datasets

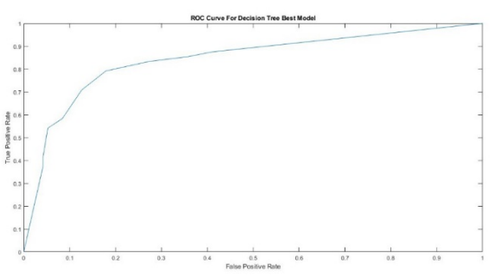
Hypothesis Statement:

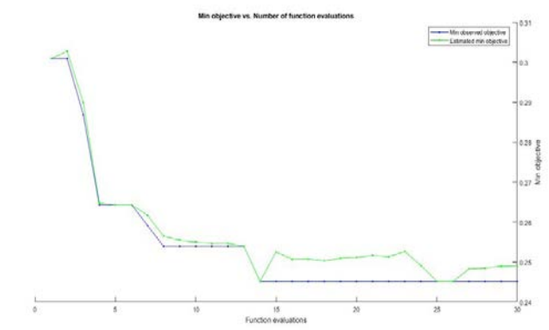
I believe that both supervised learning algorithms that I chose, will effectively classify the prediction of diabetes and give accurate and reliable results. I believe that both models will provide meaningful insight into the prediction of diabetes, however, the decision tree model will perform better since we know that it can handle both categorical as well as numerical data. In my opinion, I believe that I have performed the necessary data pre-processing such as normalizing the dataset for the Naïve Bayes model and standardizing the dataset for the Decision Tree model, so that both models have the best foundations to perform as best as they can. After training and testing both models as well as tuning their hyperparameters, we will calculate and compare a variety of model evaluation methods to see which of the two algorithms will perform better.

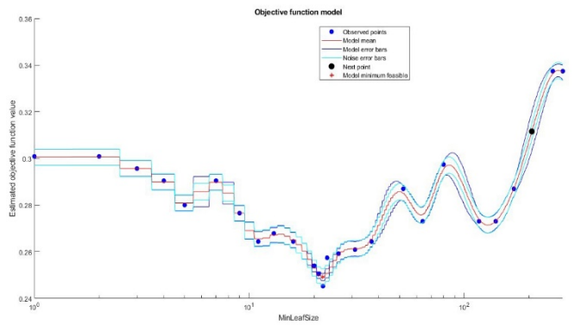
Methodology:

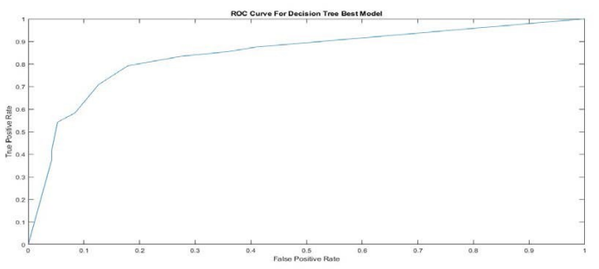
After our data pre-processing and exploratory data analysis which consisted of replacing zeroes with median, removing outliers, and normalizing and standardizing our variables, we partitioned our datasets into training and testing sets using a cross-validation partition and the holdout method. We assigned 80% of our data to the training set and 20% of our data to the test set. We started by training our initial Naive Bayes model and predicted on the test set. We then calculated the major evaluation metrics such as precision, recall, classification accuracy, f1 score, and the AUC curve. Then we proceeded by doing the same process for the decision tree algorithm. We split our dataset and assigned 80% to the training test and 20% to the test set for consistency and proceeded with calculating the same performance metrics as the Naïve Bayes model.
Then we performed some feature selection on our decision tree model and plotted a bar chart which can be found to the right of this page. We did not do this for the Naïve Bayes model due to the assumption of independence
Finally, we performed hyperparameter optimization using the auto function in MATLAB so it could choose the best-tuned parameters for each model. Our last step was to calculate revised performance metrics with the hope of improving the accuracy of our models.











Evaluations Before Hyperparameter Optimization					
	Accuracy	Percision	Recall	F1	AUC
Naïve Bayes	0.8112	0.7692	0.625	0.6897	0.8822
Decision Tree	0.7343	0.625	0.5208	0.5682	0.7016

Evaluations After Hyperparameter Optimization						
	Accuracy	Percision	Recall	F1	Resubloss	AUC
Naïve Bayes	0.8252	0.7949	0.6458	0.7126	0.2904	0.8866
Decision Tree	0.8042	0.7778	0.5833	0.6667	0.1861	0.8435

Analysis And Critical Evaluation Of The Results

By firstly comparing the Naïve Bayes and Decision Tree algorithms before hyperparameter optimization, we can see that the Naïve Bayes model performed relatively better than the Decision Tree model in our classification task since it had higher scores than the Decision Tree model in all of the evaluations we have done. The Naïve Bayes model had on average about a 12% better classification score than the Decision Tree model, and a significant 18% better performance on the AUC score.
However, after hyperparameter tuning, we can observe that the Decision Tree model had a significant improvement in all of the evaluation metrics, especially in the AUC, Precision, and Recall, scores. On the other hand, even though the Naïve Bayes still showed an improvement in classifying diabetes prediction, it was a relatively small improvement. Overall, this shows us that hyperparameter optimization indeed helped both models to improve their performance. Pre hyperparameter optimization, the Naïve Bayes model consistently performed better than the Decision Tree model, however post hyperparameter optimization, both models performed well, with the Decision Tree model showing significant improvement.
Considering the feature importance bar chart, we can see that Glucose has the highest score showing us that glucose is a strong feature variable for the prediction of diabetes. On the other hand, feature variables such as Skin Thickness and Insulin have a very low score making them have a lower impact on the prediction of diabetes.

Lessons Learned And Future Work

Whilst performing this study we have learned that the data pre-processing step which consisted of normalizing and standardizing variables, handling missing values, and detecting and removing outliers was a critical step that had to be done to maximize the performance of each model in predicting diabetes accurately and consistently. We have seen that both models have a few similar but also different advantages and disadvantages making us see that independence is a very important factor for the Naïve Bayes model. This assumption justifies that the Naïve Bayes model is better for smaller datasets and the Decision Tree model for larger datasets because it can handle both numerical and categorical variables. The feature importance chart on the Decision Tree model clearly showed that glucose is a very important variable when it comes to diabetes prediction.
For future work, it would be better to include the k-fold cross-validation technique to tackle the problem of imbalanced data that we currently have. Performing feature engineering on the models might have improved their classification accuracy even more. We could use ensemble methods such as gradient descent to improve model performance as well as perform SMOTE to deal with our imbalanced dataset. Finally, we could have found a real-life report that compared these algorithms on our dataset, and see how our algorithms compare against those.

REFERENCES

[1]"Facts & figures," International Diabetes Federation. Accessed: Dec. 15, 2023. [Online]. Available: https://idf.org/about-diabetes/diabetes-facts-figures/

[2]"Pima Indians Diabetes Database." Accessed: Dec. 15, 2023. [Online]. Available: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

[3]"What are Naive Bayes classifiers? | IBM." Accessed: Dec. 15, 2023. [Online]. Available: https://www.ibm.com/topics/naive-bayes

[4]S. Ray, "Naive Bayes Classifier Explained: Applications and Practice Problems of Naive Bayes Classifier," Analytics Vidhya. Accessed: Dec. 15, 2023. [Online]. Available: https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

[5]A. Raj, "The Not So Naive Bayes," Medium. Accessed: Dec. 15, 2023. [Online]. Available: https://towardsdatascience.com/the-not-so-naive-bayes-b795eaa0f69b

[6]"Decision Tree Algorithm in Machine Learning - Javatpoint." Accessed: Dec. 15, 2023. [Online]. Available: https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm

[7]D. K, "Top 5 advantages and disadvantages of Decision Tree Algorithm," Medium. Accessed: Dec. 15, 2023. [Online]. Available: https://dhirajkumarblog.medium.com/top-5-advantages-and-disadvantages-of-decision-tree-algorithm-428ebdf99d9a