Binary Classification for Diabetes Detection

Arlette Slachmuylder, Deepika Parshvanath Velapure, Shatabdi Pal, Swetha Srihari

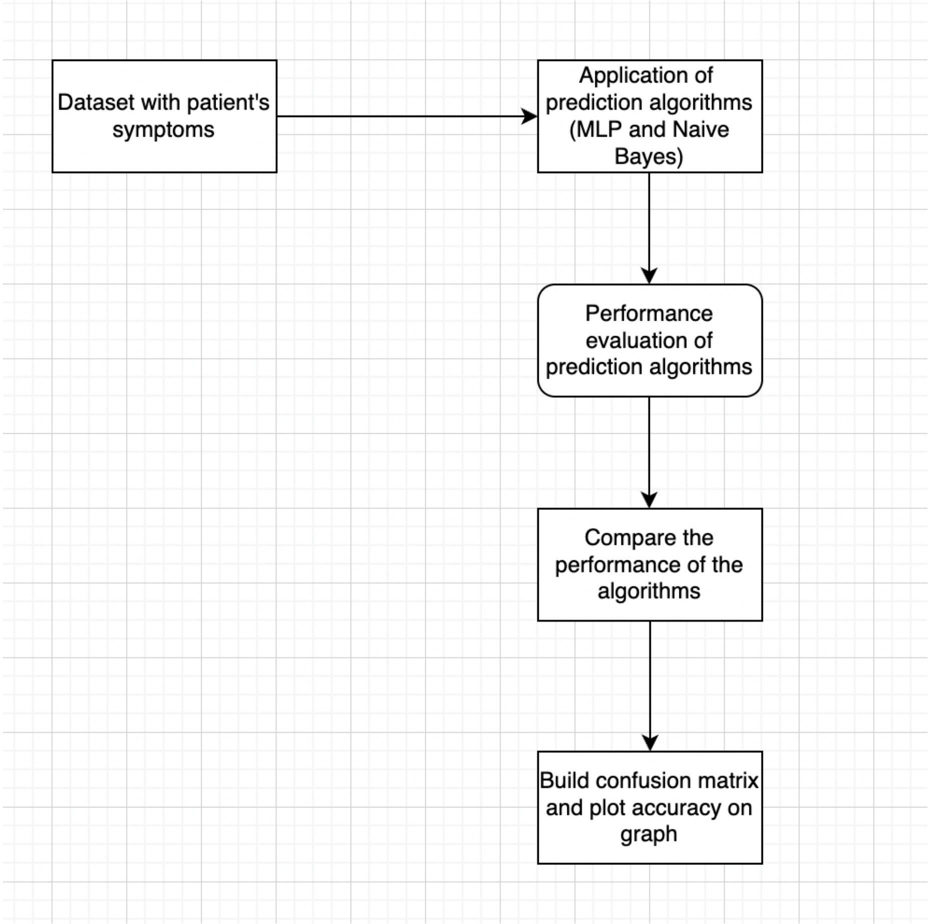
Maseeh College of Engineering and Computer Science, Portland State University

1 INTRODUCTION

The project is based on using Machine Learning techniques to predict diabetes based on diagnostic measures. Diabetes is a chronic (long-lasting) health condition that affects how the human body turns food into energy. Normally, the human body breaks down most of the food that is consumed into sugar (glucose) and releases it into the bloodstream. When the blood sugar goes up, it signals the pancreas to release insulin. Insulin acts like a key to let the blood sugar into the body's cells for use as energy. With diabetes, the body doesn't make enough insulin or cannot use it as well as it should. When there isn't enough insulin or cells stop responding to insulin, too much blood sugar stays in the bloodstream. Over time, this can cause serious health problems, such as heart disease, vision loss, and kidney disease. Thus, a diabetics prediction model will help any health professional in evaluating on a high level, whether a patient is diabetic or not based on the patient's general health details and thus, can consult them to take further steps.

Project Goal

The goal of the project is to determine and compare the performance and accuracy of Machine Learning techniques such as Multi-Layer Perceptron(MLP) and Naive Bayesian classifier, to predict if a patient is likely to be diabetic or not. The project will evaluate the performance and accordingly compare the accuracy of MLP and the Naive Bayesian classifier. It will also assess if additional features would be desired about a patient to make the model more accurate.


Dataset with patient's symptoms
Application of prediction algorithms (MLP and Naive Bayes)
Performance evaluation of prediction algorithms
Compare the performance of the algorithms
Build confusion matrix and plot accuracy on graph

.

Software and Programming Language

The models will be written using Python and the following Python libraries: Pandas, Numpy, SKLearn and Matplotlib.

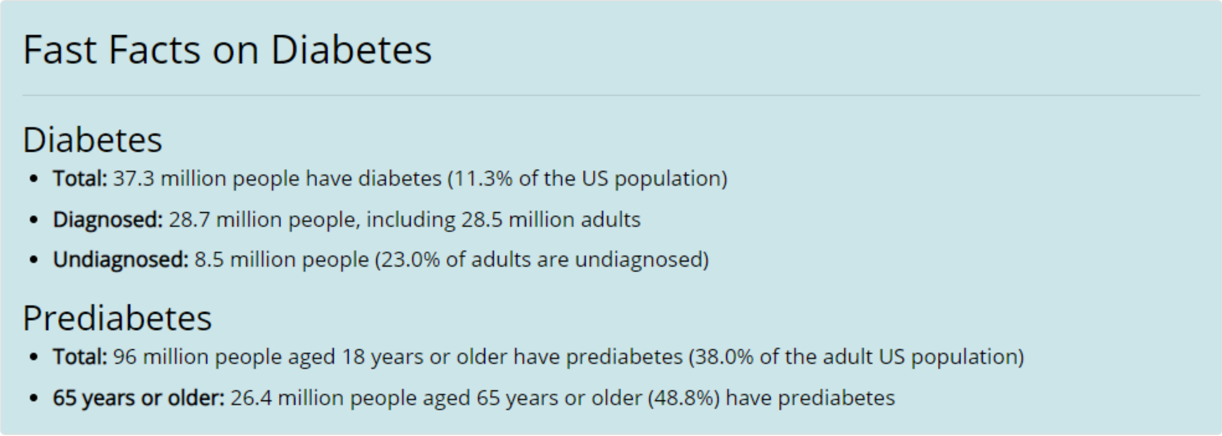
2 RELATED WORKS

The articles below summarize the great health cost of diabetes and its impact on quality of life. They also detail the efforts of applying machine learning to the health data collected, to predict diabetes.

Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8500744/

A comprehensive review of machine learning techniques on diabetes detection: https://link.springer.com/article/10.1186/s42492-021-00097-7

Below shows the statistics of diabetes in US



Fast Facts on Diabetes

Diabetes

- **Total:** 37.3 million people have diabetes (11.3% of the US population)
- **Diagnosed:** 28.7 million people, including 28.5 million adults
- **Undiagnosed:** 8.5 million people (23.0% of adults are undiagnosed)

Prediabetes

- **Total:** 96 million people aged 18 years or older have prediabetes (38.0% of the adult US population)
- **65 years or older:** 26.4 million people aged 65 years or older (48.8%) have prediabetes

As diabetes is increasingly becoming common in people belonging to all the age-groups, we want to build a model which can help any health professional to identify if a patient has diabetes based on their medical report.

3 DATASET DESCRIPTION

The data provided by the National Institute of Diabetes and Digestive and Kidney Diseases consisting of a labeled dataset of 390 patients that includes 16 features of a patient will be used to train and test the neural networks to predict if the patient is diabetic or not. The features are common data points that are available on most patient's general medical records. Though the dataset has a relatively low number of features, health predictions need to be sophisticated in order to be of value to a doctor and indicate as accurately as possible when further testing is needed. Therefore, the sophistication of a neural network function is the first model of choice to see how accurate a model can be built without overfitting the data.

The features for the prediction are their BMI, weight, age, glucose, and so on. The model will do binary classification for prediction. The model generates 1 for diabetes and 0 for non-diabetes.

Thus, this project will evaluate the performance and accordingly compare the accuracy of MLP and the Naive Bayesian classifier. It will also assess if additional features would be desired about a patient to make the model more accurate.

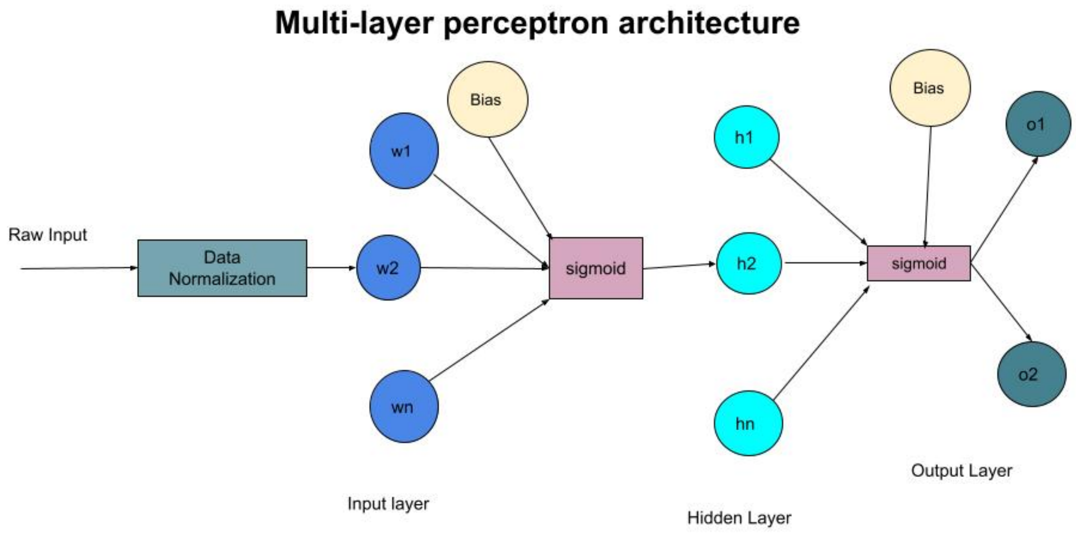
4 BINARY CLASSIFIER FOR DIABETES DETECTION

Binary Classification

Binary classification refers to those classification tasks that have two class labels. Typically, classification tasks involve one class that is the normal state while the other class is the abnormal state. In the context of this report, the absence of diabetes is seen as the normal state, while being diagnosed with diabetes is seen as the abnormal state. As is typical with this model, the absence of diabetes will be assigned a 0 and the diagnosis will be assigned a 1.[1]

[1] https://machinelearningmastery.com/types-of-classification-in-machine-learning/

4.1 Multilayer Perceptron



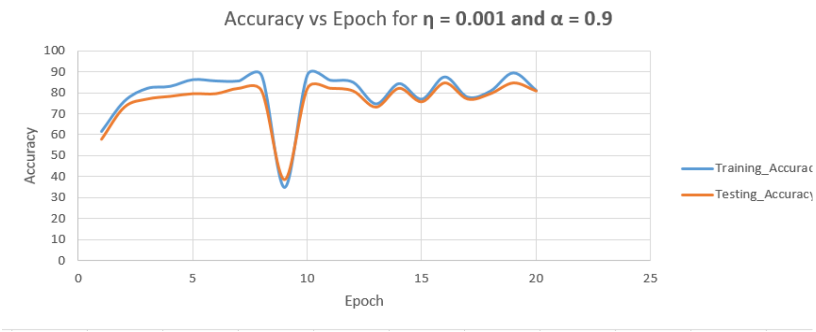
Sigmoid function $f(x) = \frac{1}{1 + \exp(-x)}$

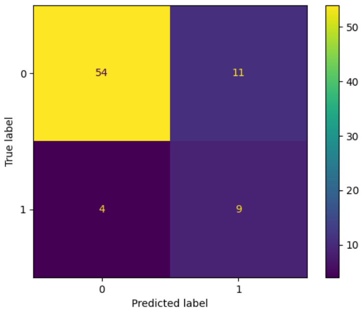
In building a multilayer perceptron model, several different model combinations were evaluated, to determine what produced the best results. All contained a single layer of hidden neurons and a single output neuron. Data prep was done to turn the gender fields consisting of text equalling “male” or “female” into two different columns that were identified as a 1 or 0. Next, all the data was normalized, so that the range of each feature would be similar to each other, and not be given greater weight by the neurons. Different models were tested, by varying the number of epochs, the number of hidden neurons, the learning rate, the momentum and the exclusion of dependent features.

The best accuracy was found to be a model that was built off of 20 epochs, 10 internal neurons, a removal of dependent features, a learning rate of .01 and a momentum of .9. This produced an accuracy of 94.87. A similar model that differed only in having 50 epochs rather than 20 had the same accuracy, but the decision matrix held many more false positives. The decision matrix for 20 epochs was shown to have 6 false positives and negatives as compared to the 50 epochs that had double that number.

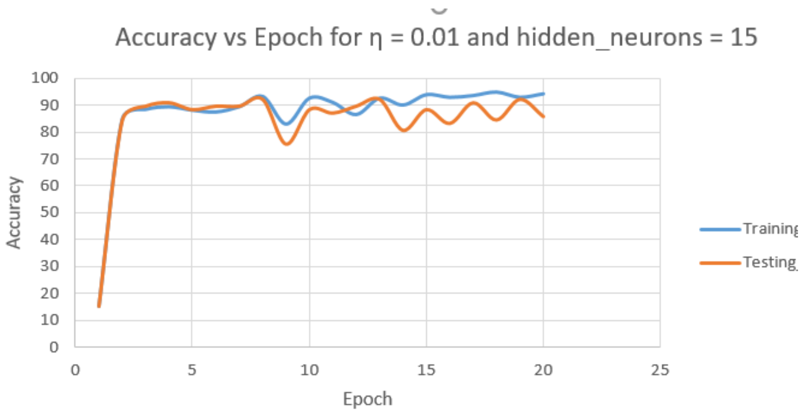
The three features that were removed were ratios that were a composite of two other columns in the data. This probably resulted in giving artificial weight to these features. Removing them brought the model from 88% accuracy to close to 95%.

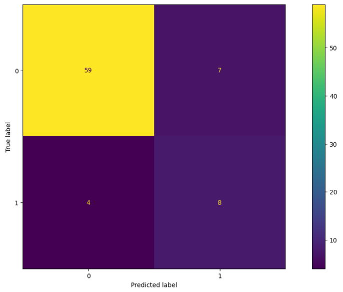
The full results of all the trials are shown below.

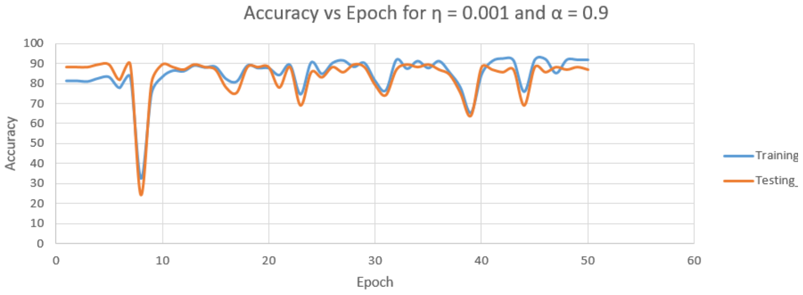
<table>
<tr><th>Epoch</th><th># Hidden Neurons</th><th>Learning Rate</th><th>Momentum</th><th># Features</th><th>Accuracy</th></tr>
<tr><td>20</td><td>10</td><td>.001</td><td>.9</td><td>16</td><td>80.77</td></tr>
<tr><td colspan="6">


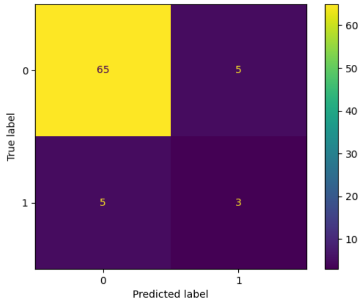


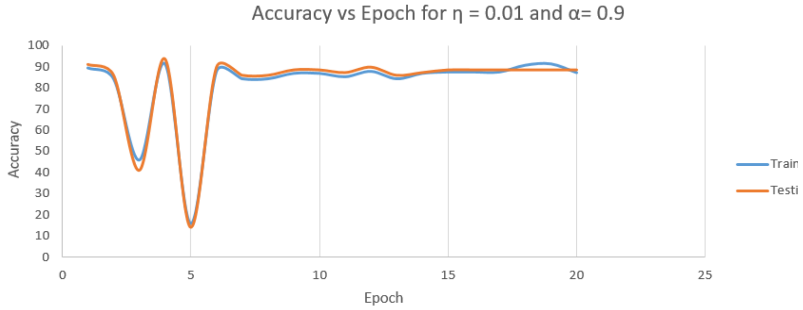
</td></tr>
</table>

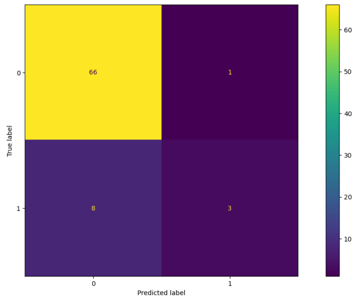
<table>
<tr><th>Epoch</th><th># Hidden Neurons</th><th>Learning Rate</th><th>Momentum</th><th># Features</th><th>Accuracy</th></tr>
<tr><td>20</td><td>15</td><td>.01</td><td>.9</td><td>16</td><td>85.90</td></tr>
<tr><td colspan="6">




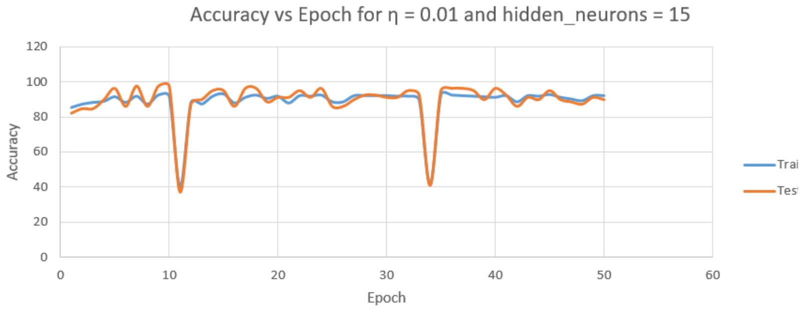
</td></tr>
<tr><td>50</td><td>10</td><td>.001</td><td>.9</td><td>16</td><td>87.18</td></tr>
<tr><td colspan="6">


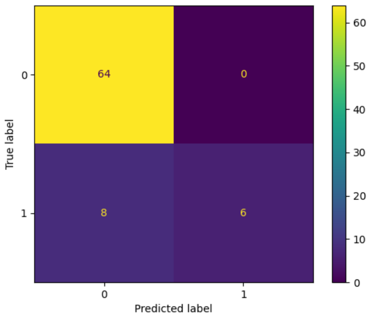


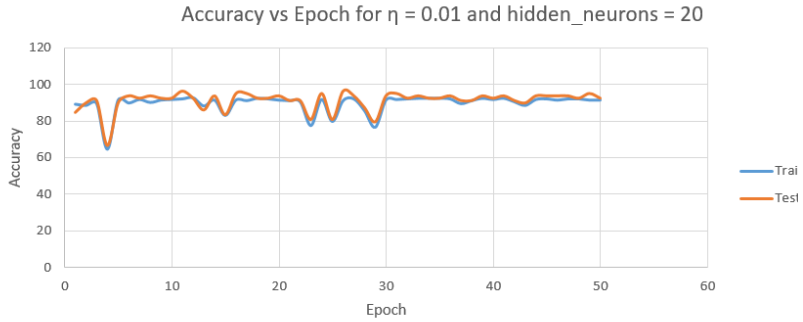
</td></tr>
<tr><td>20</td><td>20</td><td>.01</td><td>.9</td><td>16</td><td>88.46</td></tr>
<tr><td colspan="6">


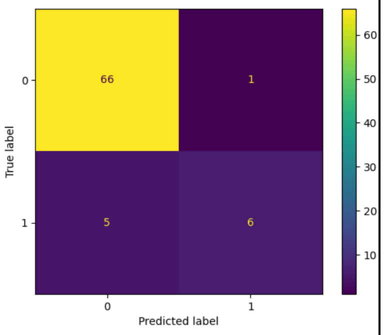


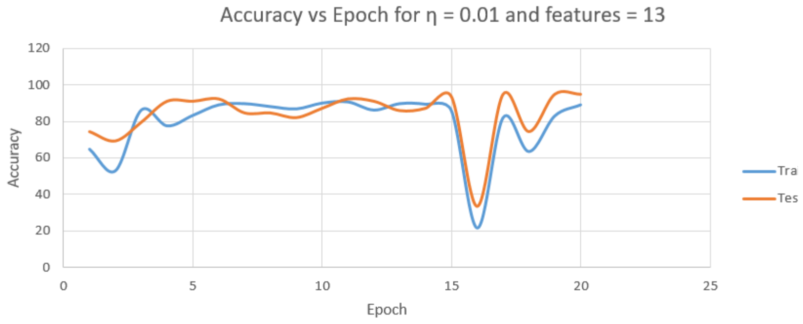
</td></tr>
</table>

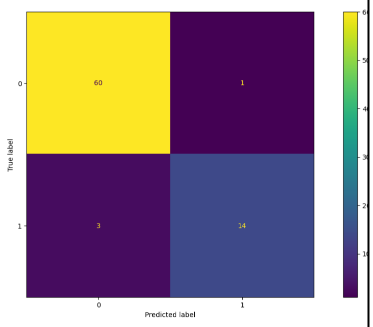
<table>
<tr><th>Epoch</th><th># Hidden Neurons</th><th>Learning Rate</th><th>Momentum</th><th># Features</th><th>Accuracy</th></tr>
<tr><td>50</td><td>15</td><td>.01</td><td>.9</td><td>16</td><td>89.74</td></tr>
<tr><td colspan="6">




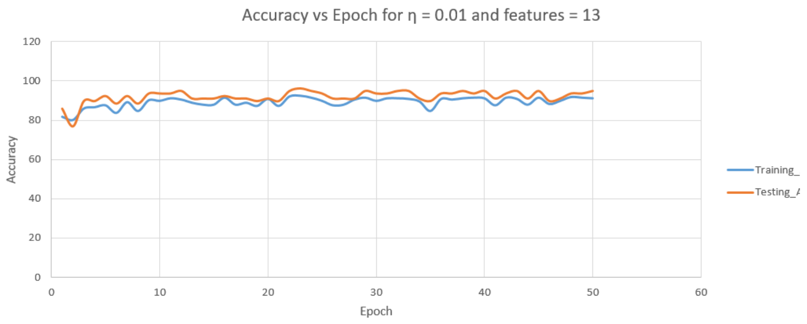
</td></tr>
<tr><td>50</td><td>20</td><td>.01</td><td>.9</td><td>16</td><td>92.31</td></tr>
<tr><td colspan="6">


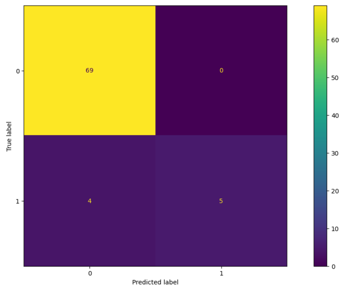


</td></tr>
<tr><td>20</td><td>10</td><td>.01</td><td>.9</td><td>13</td><td>94.87</td></tr>
<tr><td colspan="6">




</td></tr>
</table>

<table>
<tr><th>Epoch</th><th># Hidden Neurons</th><th>Learning Rate</th><th>Momentum</th><th># Features</th><th>Accuracy</th></tr>
<tr><td>50</td><td>10</td><td>.01</td><td>.9</td><td>13</td><td>94.87</td></tr>
<tr><td colspan="6">


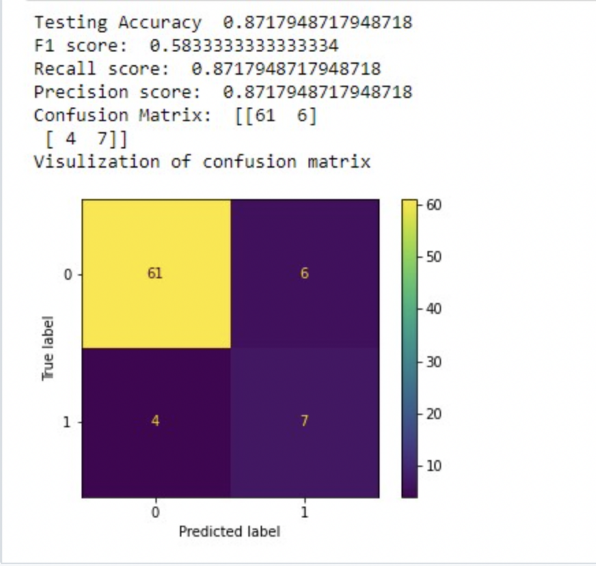


</td></tr>
</table>

4.2 Gaussian Bayesian Model

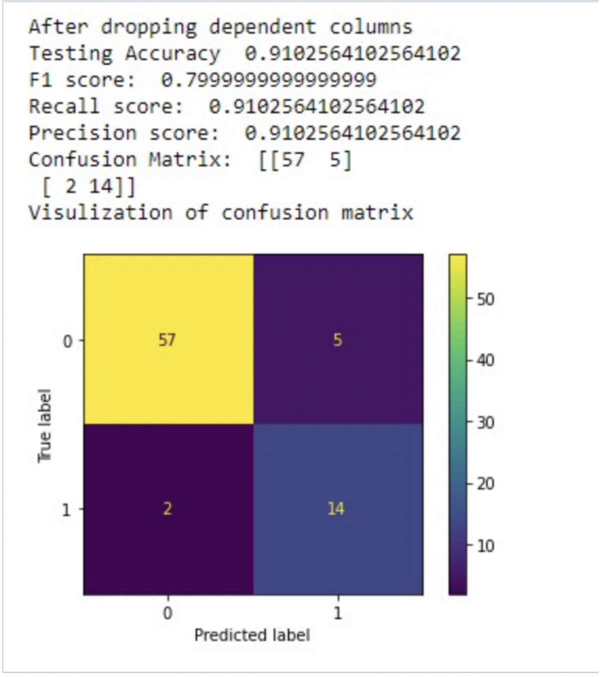
For the Bayesian model, data prep was done to modify the gender to be 0 for male and 1 for female. The data was split into training and testing sets with the training set consisting of 80% of the original dataset and the testing set consisting of the remaining 20%. The Gaussian Naive Bayes model was trained against the training set. A second training iteration was done without the dependent columns that were ratios based on the values of other columns. In this second iteration, the model achieved a higher accuracy. It went from 87% with the dependent columns to 91% without these columns. The following are the results obtained from testing this model against the test data :

With the dependent columns



```
Testing Accuracy  0.8717948717948718
F1 score:  0.5833333333333334
Recall score:  0.8717948717948718
Precision score:  0.8717948717948718
Confusion Matrix:  [[61  6]
 [ 4  7]]
Visulization of confusion matrix
```

Without the dependent columns



```
After dropping dependent columns
Testing Accuracy  0.9102564102564102
F1 score:  0.7999999999999999
Recall score:  0.9102564102564102
Precision score:  0.9102564102564102
Confusion Matrix:  [[57  5]
 [ 2 14]]
Visulization of confusion matrix
```

5 RESULTS AND DISCUSSION

This report presents the results of the diabetes prediction using both multilayer perceptron and Bayesian approaches. The performance of the classifier was demonstrated by accuracy, precision, recall and F1 score. Performance was evaluated by using all the features represented in the original dataset and dropping the dependent features. The classification task involves 1 (diabetes) and 0 (non-diabetes). The original dataset consists of 390 patients, consisting of both diabetes and non-diabetes. The observation was carried out by splitting the data into training and test sets with an 80-20 ratio. The reason behind choosing the Bayesian model is that we have a very small dataset. Multilayer perceptron requires a large dataset, and sometimes a large data set is not feasible. As per our knowledge on supervised learning, multilayer perceptrons give better performance than Bayesian models for large datasets. Yet, despite the smaller dataset available for this experiment, the same results were found to be true. The MLP achieved an accuracy close to 95 %, whereas the Bayes model achieved an accuracy of 84 – 87 %. With varying the number of hidden neurons, the learning rate and accuracy of MLP was improved by ~14% (80 to 94%). The Naïve Bayes classifier had performed the classification task as expected. The Naïve Bayes classifier has an advantage of being computationally easier than MLP.

This project is mainly a proof-of-concept demonstration of a machine learning algorithm. We haven't explored other algorithms due to shortage of time. Our observation is limited to MLP and Naïve Bayes classification. The classifier can be reimplemented by SVM and K-means algorithm to compare the performance with our approach.

6 CONCLUSIONS

The multilayer perceptron model performed slightly better by yielding 94% accuracy for detecting diabetes in comparison with the Gaussian Naive Bayes model which gave only 91% accuracy with the test data.

ACKNOWLEDGMENTS

Team member contributions:

MLP Model: Arlette Slachmuylder, Deepika Parshvanath Velapure

Bayes Model: Shatabdi Pal, Swetha Srihari

Final document: Arlette Slachmuylder, Deepika Parshvanath Velapure, Shatabdi Pal, Swetha Srihari

Also thanks to Dr. Anthony Rhodes and Pan Tan

REFERENCES

Binary Classification:

https://machinelearningmastery.com/types-of-classification-in-machine-learning/

Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8500744/

A comprehensive review of machine learning techniques on diabetes detection: https://link.springer.com/article/10.1186/s42492-021-00097-7

https://www.cdc.gov/diabetes/basics/index.html