Project Report 8

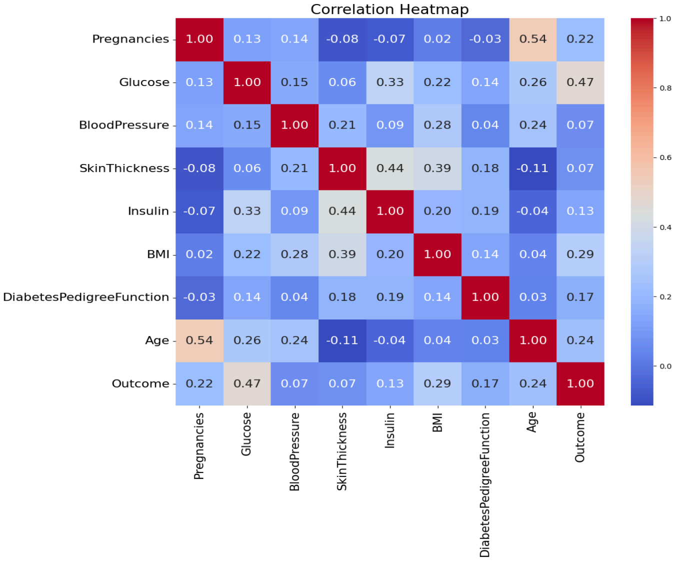
Sadra Jafari and Caden Maxwell

Introduction

The aim of this project is to create decision trees with different depths (3 and 9) and neural networks with different hidden layers to predict diabetes based on 8 numerical features named: pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, etc, with a target column which has 0 for not having and 1 for having diabetes (here 1 is more important for us). With 768 records without NaN values. The stakeholders of the analysis can be the doctors to help them in diagnosis and patients for quick check-ups. The models are created with 80% training data and tested with 20% test data. We found 0.68 for precision, 0.61 for recall, and 0.64 for f-score with the decision tree with depth 3 and 0.66 for precision, 0.70 for recall, and 0.68 for f-score with depth 9 (slightly better). Moreover, we created different neural network models (with hidden layers: (10, 14, 6), (5, 8, 11, 15), and (5, 8)). The simple model with hidden layer size (5, 8) with the precision of 0.67, recall of 0.73, and f-score of 0.70 performed the best among other neural network models. Github repository and presentation links.

Dataset

In this project, we used a diabetes dataset[1] of 768 patients without NaN values. The dataset comprises eight possible diabetes predictors, including numeric features like pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, and age. The dataset also denotes whether or not the patient actually has (1) or does not have (0) diabetes. Based on the provided correlation matrix (below) we selected the top 5 correlated features for two decision tree models and used all the features for neural network models. Based on this we can see that in terms of Outcome, the most correlated value is Glucose with a value of 0.47, and then BMI with a value of 0.29.



Analysis technique

For the decision tree part, we split the data into 80% for training to make the model with different max depths (3 and 9) tested the models with 20% of testing data, and found

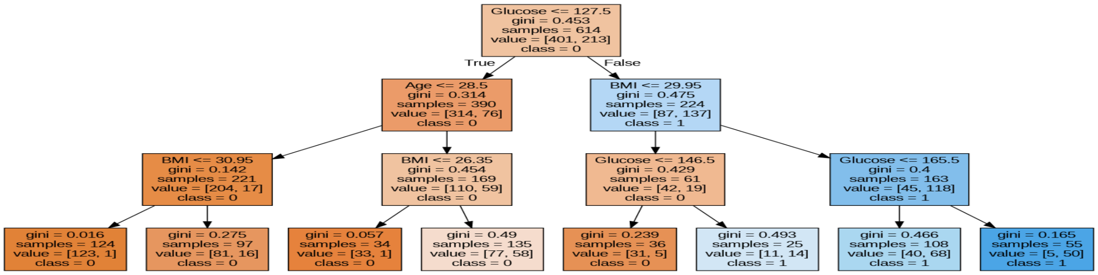
[1] https://www.kaggle.com/datasets/jillanisofttech/diabetes-disease-updated-dataset/data

performance metrics precision, recall, and f-score then we will talk about bias and variance based on number of samples on the leaves. Also, we visualized the decision trees that we made. Additionally, we show the feature importance and confusion matrix of both decision trees. It needs to be mentioned for the decision tree creation we used the top 5 correlated features which were: "Glucose", "BMI", "Age", "Pregnancies", and "DiabetesPedigreeFunction."

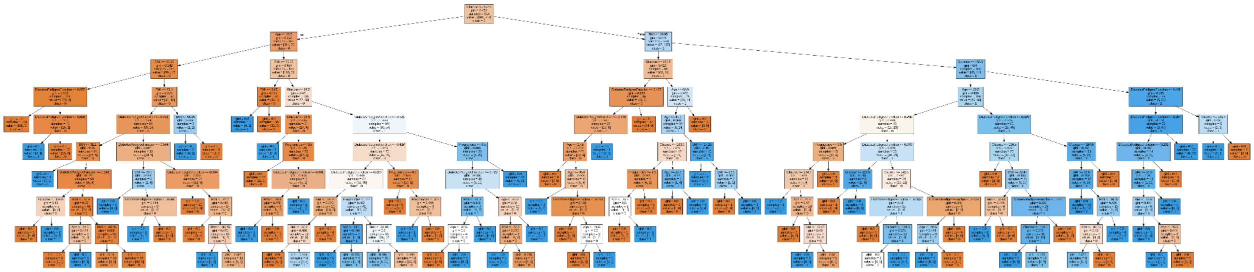
Next, we utilized three neural network models (using all the features) with slightly different parameters (hidden layers: (10, 14, 6), (5, 8, 11, 15), and (5, 8)), all built on a Multi-Layer Perceptron architecture. These models also were deemed suitable for our analysis, as they can capture complex, non-linear relationships with ease. The confusion matrices and structure of models are shown as well and the features are standardized for neural networks.

Result

Using the decision tree with a max depth of 3, we were able to predict whether samples had diabetes with precision of 0.68, recall of 0.61, and f-score of 0.64. And got 0.79 for precision, 0.83 for recall, and 0.81 for f-score for class label 0. The tree for this model is shown below.



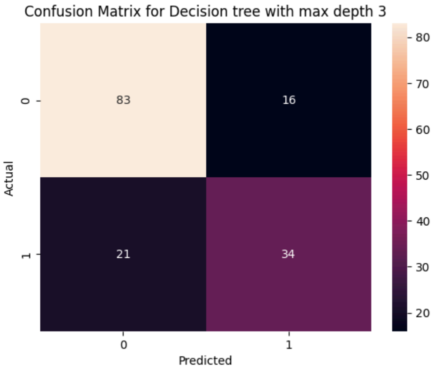
We can see that the number of samples in the leaves of the tree is pretty high, meaning that our model is slightly underfitting, with a high bias and low variance. However, in the next decision tree with a max depth of 9, we can see that there is a high variance because the number of samples in each leaf is low and may cause overfitting.

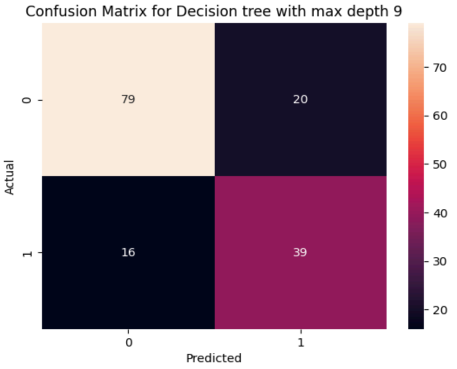


For this model, we found 0.83 and 0.66 for precision, 0.79 and 0.70 for recall, and 0.81 and 0.68 for f-score for class labels 0 and 1.

In the following table, we can see the feature importance of each of the trees. The table shows that in the tree with max depth 3, the Pregnancies and DiabetesPedigreeFunction have 0 values, meaning they weren't helpful (used) at all for deciding whether a person had diabetes. However, the table shows the DiabetesPedigreeFunction as a much more helpful indicator of diabetes in the tree with max depth 9. Additionally, in both trees, Glucose and BMI have the highest importance score, which is reflective of the high correlations they had in the correlation matrix above.

	Glucose	BMI	Age	Pregnancies	DiabetesPedigreeFunction
Importance Score max depth 3	0.597	0.262	0.140	0.000	0.000
Importance Score max depth 9	0.364	0.265	0.165	0.036	0.167


Confusion Matrix for Decision tree with max depth 3


Confusion Matrix for Decision tree with max depth 9

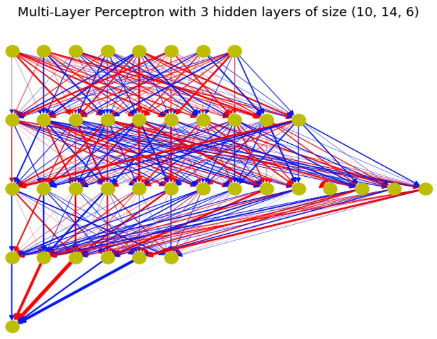
As depicted by these two confusion matrices, both models had very similar results. The decision tree with a max depth 3 predicted 117 patients correctly, but also predicted 16 false positives and 21 false negatives. The model with a max depth of 9 predicted 118 cases correctly but gave 16 false negatives and 20 false positives in the case of having diabetes (0 is negative and 1 is positive).

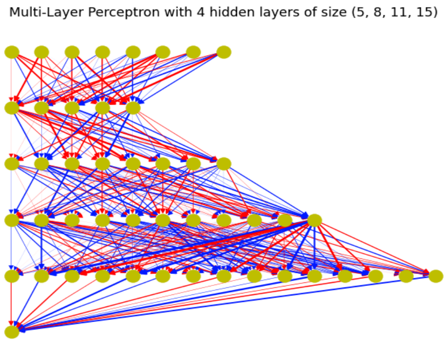
The decision trees usually start by splitting on the Glucose if it is more than 127.5, the tree would most of the time predict 1 on the other hand 0 (but not for certain). After that, the trees diverge and begin to split on different attributes. The structure of both trees is the same in our case. Generally, different nodes at the same level may split on different attributes depending on the dataset characteristics.

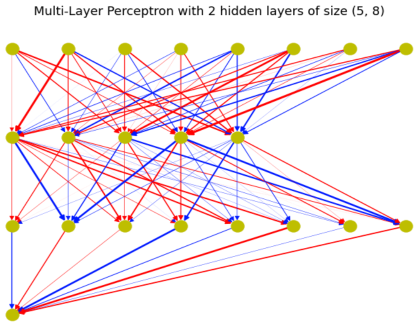
Similarly, the neural network models, with varying hidden layer sizes, demonstrated consistent performance across different configurations. The results for each of the three models are shown below.

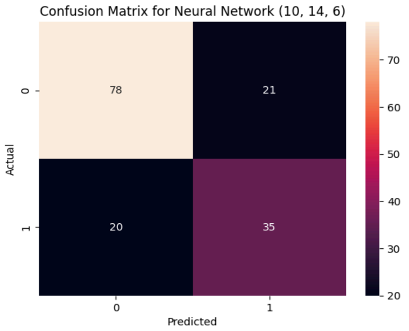
MLP Scores (results class labels 0 and 1 sequentially)			
Hidden Layer Sizes	(10, 14, 6)	(5, 8, 11, 15)	(5, 8)
Max Iterations	1000	1000	1000
Precision	0.80, 0.62	0.81, 0.65	0.84, 0.67
Recall	0.79, 0.64	0.80, 0.67	0.80, 0.73
F1-Score	0.79, 0.63	0.81, 0.66	0.82, 0.70

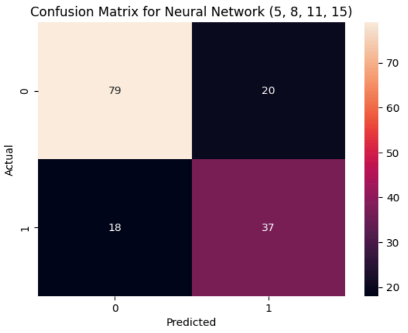
These results suggest that neural networks, while offering flexibility in modeling complex relationships, may not substantially outperform decision trees in this context. Notably, increasing the complexity of the neural network architecture did not consistently improve performance. In fact, having a simpler model was slightly more effective. Furthermore, there did not seem to be any correlation between the decision tree nodes and the edge weights of the neural networks, as visualized below.

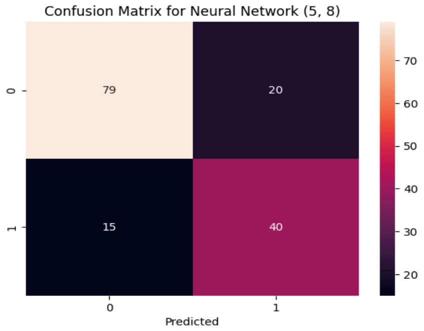

Multi-Layer Perceptron with 3 hidden layers of size (10, 14, 6)


Multi-Layer Perceptron with 4 hidden layers of size (5, 8, 11, 15)


Multi-Layer Perceptron with 2 hidden layers of size (5, 8)


Confusion Matrix for Neural Network (10, 14, 6)


Confusion Matrix for Neural Network (5, 8, 11, 15)


Confusion Matrix for Neural Network (5, 8)

We found that neural network (5, 8) is operating better than all other neural networks by predicting 119 cases correctly and 20 cases of false positives, and 15 cases of false negatives based on confusion matrices.

In summary, our study suggests that decision trees and neural networks offer the potential for predicting diabetes onset based on the dataset features. However, more research is necessary to improve the accuracy and reliability of these models in this field. It emphasizes the importance of continuously refining and validating predictive models to better guide healthcare decisions and other public health strategies.