

2022

Assignment 2



Jasmin Chiu (19083325) & Alex Parker
(20107336)

Auckland University of Technology

1/1/2022

Contents

Table of Figures	3
Introduction	4
Data Exploration.....	4
Attributes, Instances and Data types.....	4
Summary Statistics	5
Features of dataset using visualisation.....	6
Pre-processing.....	8
Decision Tree Classifier.....	9
a) Building the model.....	9
b) Two adjusted parameters	11
Finding max depth	11
Finding max nodes.....	12
c) Confusion matrix.....	13
d) Feature importance	14
Artificial Neural Network (ANN).....	15
a) Five most significant features	15
b) One hidden layer	17
c) Error and loss value.....	18
d) Two hidden layers	19
e) Accuracy variation	20
Performance Comparison.....	20
Appendix.....	21

Table of Figures

Figure 1: Distribution of class.....	5
Figure 2: Summary statistics of dataset.....	5
Figure 3: Boxplots of Age, SystolicBP, HeartRate.....	6
Figure 4: Histograms of BodyTemp and BS.....	7
Figure 5: Count of null and NA values	8
Figure 6: Baseline decision tree	9
Figure 7: Final decision tree	10
Figure 8: max_depth and averaged 10-Fold cross-validation (CV) score	11
Figure 9: max_leaf_nodes and averaged 10-Fold cross-validation (CV) score	12
Figure 10: Confusion matrix of the test split	13
Figure 11: Confusion matrix metrics	13
Figure 12: Feature importance graph.....	14
Figure 13: F-score table	14
Figure 14: ANOVE F scores.....	15
Figure 15: ANOVE F scores.....	15
Figure 16: Top 5 features	16
Figure 17: Accuracy of different iteration values from 1-50.....	17
Figure 18: Iterations and loss value.....	18
Figure 19: Neurons classification table.....	19
Figure 20: Neurons classification table.....	Error! Bookmark not defined.

Introduction

By 2030, the United Nations wants to reduce global maternal mortality ratio to less than 70 per 100,000 live births. To assist in the accomplishment of this target, this report aims to investigate risk factors that contribute to maternal mortality in Bangladesh.

Data Exploration

Attributes, Instances and Data types

The data, collected in 2020, has been sourced from hospitals, community clinics, maternal health cares from rural areas of Bangladesh.

The dataset contains the following attributes that are known to be significant risk factors for maternal mortality:

Age	Any ages of a woman during pregnancy - in years
SystolicBP	Systolic Blood Pressure, upper value of Blood Pressure - in mmHg
DiastolicBP	Diastolic Blood Pressure, lower value of Blood Pressure - in mmHg
BS	Blood Sugar, blood glucose levels - in mmol/L
HeartRate	Normal resting heart rate - in beats per minute
BodyTemp	Body temperature - in Fahrenheit (°F)
Risk Level	Predicted Risk Intensity Level during pregnancy, considering HeartRate

The dataset contains 1014 instances and 6 features with 1 class:

Name of attribute	Data Type	Input/Output
Age	Quantitative– Discrete data	Input
SystolicBP	Quantitative - Discrete data	Input
DiastolicBP	Quantitative - Discrete data	Input
BS	Quantitative - Continuous data	Input
BodyTemp	Quantitative - Continuous data	Input
HeartRate	Quantitative - Discrete data	Input
Risk Level	Qualitative/categorical data	Output

Summary Statistics

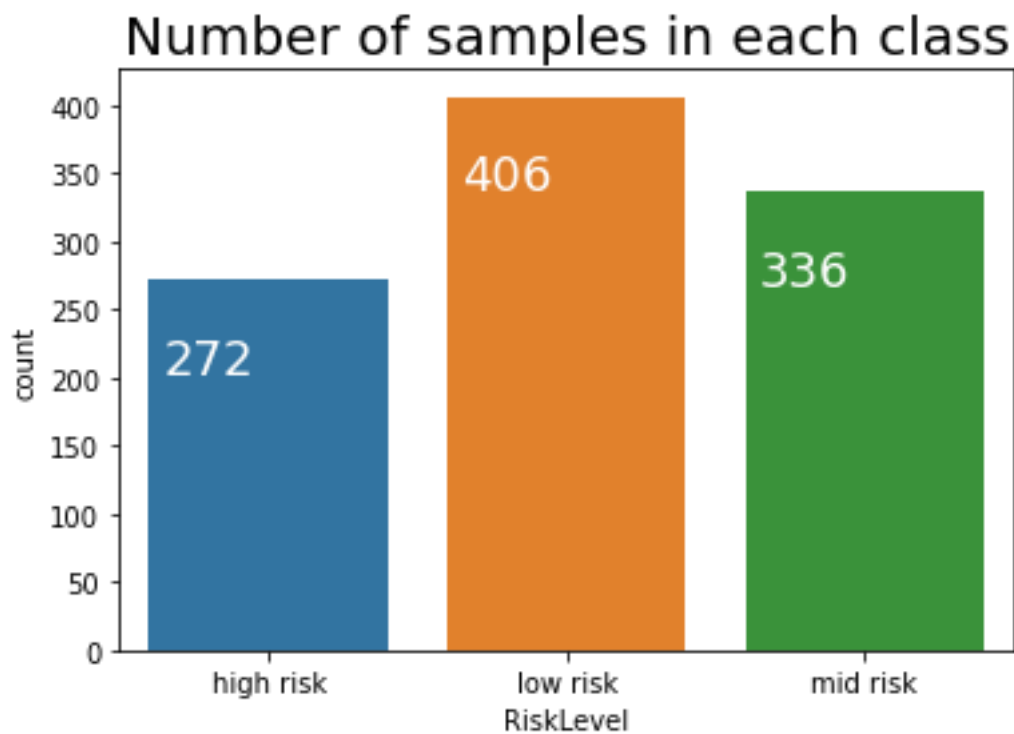


Figure 1: Distribution of class

This report aims to investigate how 'RiskLevel' can be predicted based on other attributes. The distribution of the values is not perfectly uniform, and the values vary with different means, so some standardization or normalization may be needed prior to modelling.

	Age	SystolicBP	DiastolicBP	BS	BodyTemp
count	1014.000000	1014.000000	1014.000000	1014.000000	1014.000000
mean	29.871795	113.198225	76.460552	8.725986	98.665089
std	13.474386	18.403913	13.885796	3.293532	1.371384
min	10.000000	70.000000	49.000000	6.000000	98.000000
25%	19.000000	100.000000	65.000000	6.900000	98.000000
50%	26.000000	120.000000	80.000000	7.500000	98.000000
75%	39.000000	120.000000	90.000000	8.000000	98.000000
max	70.000000	160.000000	100.000000	19.000000	103.000000

	HeartRate
count	1014.000000
mean	74.301775
std	8.088702
min	7.000000
25%	70.000000
50%	76.000000
75%	80.000000
max	90.000000

Figure 2: Summary statistics of dataset

From Figure 1, it is evident that there are no negative values in this dataset as all the minimum values are greater than zero. The counts for all the attributes are the same, meaning that there are no missing values in any of the instances.

Features of dataset using visualisation

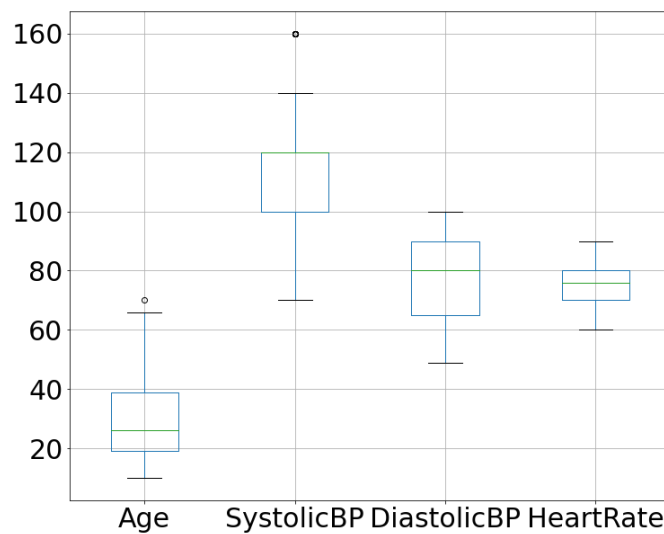


Figure 3: Boxplots of Age, SystolicBP, HeartRate

From Figure 2, it is apparent that each attribute (Age, SystolicBP, DiastolicBP, HeartRate) contains at least 1 outlier, which will have to be resolved before creating a decision tree. SystolicBP and HeartRate do not appear to be skewed; however, Age is. This is most likely due to the age group or demographic this study is concerned with. The range of ages for this dataset is between 10 and 70, but a large proportion of these ages is between 20 and 40 – a large proportion of the women in this dataset are adults.

Given that the BS and BodyTemp attributes are continuous variables, it is more suitable to plot them as a histogram. Below are the histograms for the BS and BodyTemp attributes

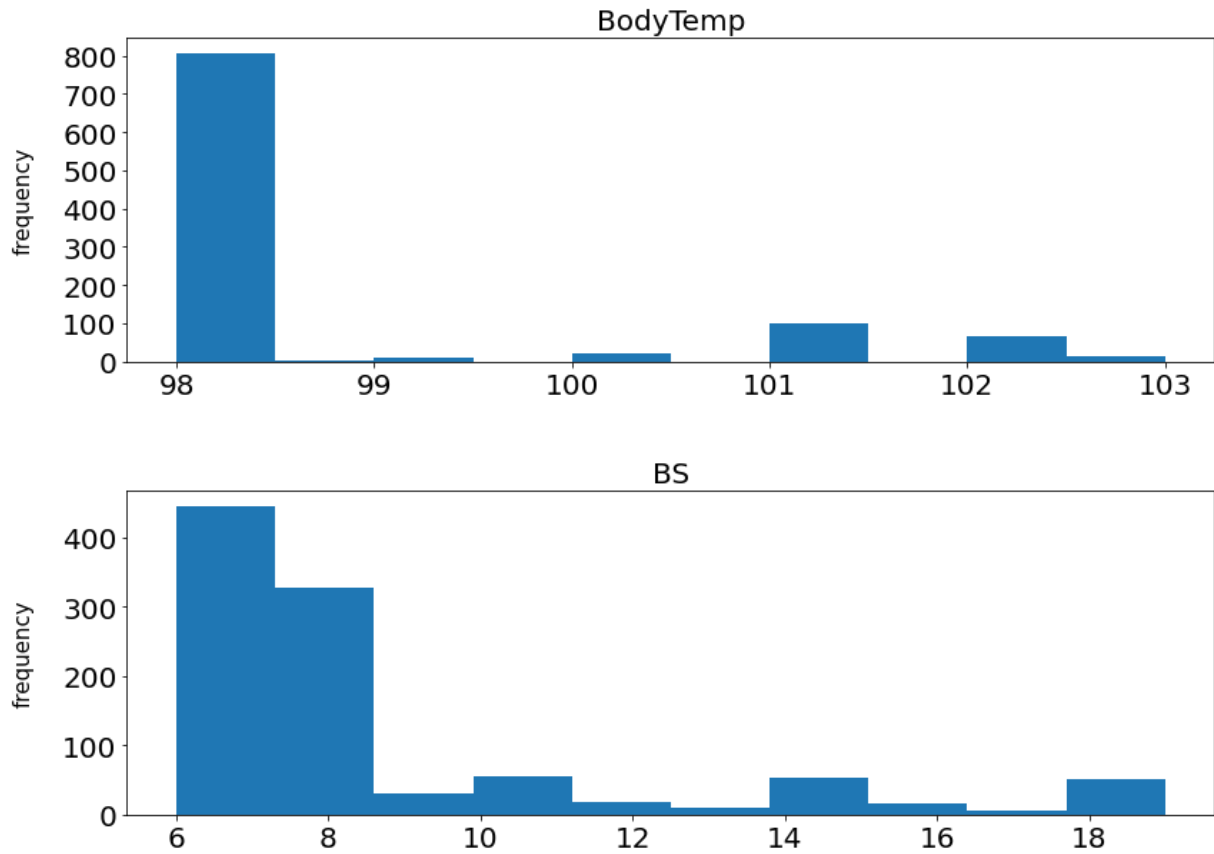


Figure 4: Histograms of BodyTemp and BS

Figure 3 shows the distribution of BodyTemp and BS.

The figure shows that BodyTemp is strongly skewed to the left. The range is between 90 to 103°F, but most of the instances of the BodyTemp attribute lie between 98 and 98.5°F.

BS is also skewed to the left, but the skew is less strong. Most of the instances lie between 6 to 9 mmol/L.

Pre-processing

Through the use of the 'isnull' and 'isna' functions, it is ensured sure that there are no missing or null data points. Given that there are no missing functions, no action needs to be taken.

```
Age          0
SystolicBP   0
DiastolicBP   0
BS           0
BodyTemp     0
HeartRate    0
RiskLevel    0
dtype: int64
```

Figure 5: Count of null and NA values

In preparation for building the decision tree, the dataset is split into a train and test set. 70% of the dataset is allocated for training, and 30% is allocated for testing.

As none of the attributes are ordinal, there is no need to encode them. Hence, it will not be included in the pre-processing section.

Decision Tree Classifier

a) Building the model

To begin, a baseline model of the classification tree is created. This is done with default parameters and the training sets from the pre-processing stage.

Below is the baseline tree, which had an overall accuracy of 0.71. In total, there are 291 tree nodes, which compresses all the nodes into one graph and essentially makes it unreadable.

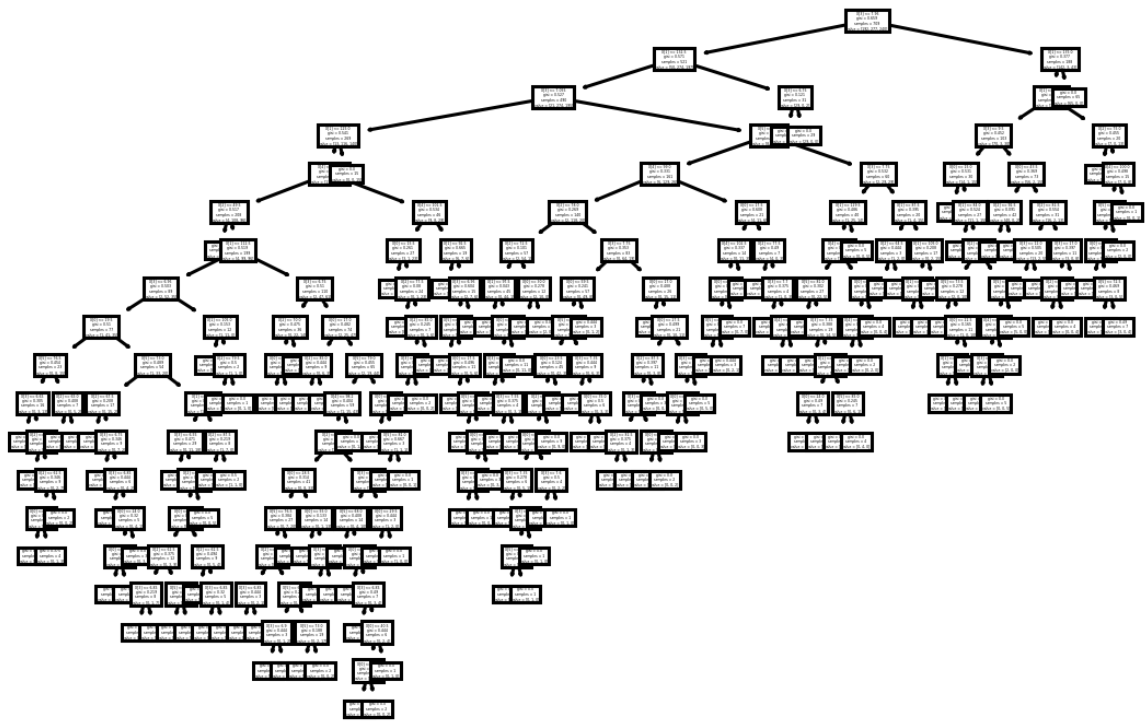


Figure 6: Baseline decision tree

The tree below is optimized to use far less tree nodes and in turn, more readable.

Decision tree trained on all the iris features using max depth=3

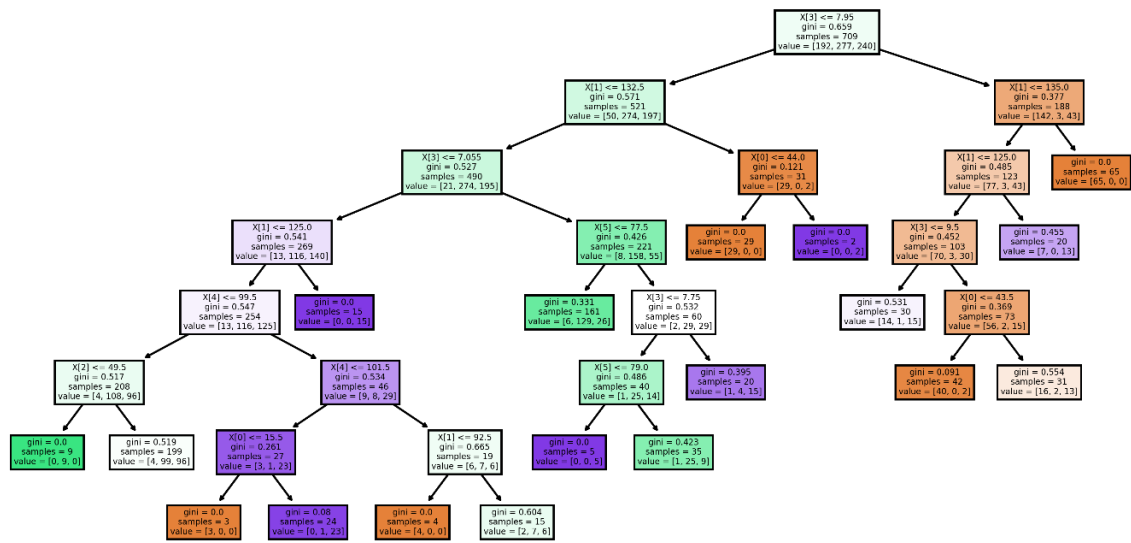


Figure 7: Final decision tree

The final decision tree contains 35 nodes, and the overall accuracy remains at 71%.

b) Two adjusted parameters

To optimize this graph, two parameters of the Decision Tree Classifier need to be altered.

The first one is the `max_depth` parameter - the `max_depth` parameter is responsible for controlling the max depth of the tree. By default, the tree will expand until all the leaves are pure. Below is a graph showing the optimal `max_depth`.

Finding max depth

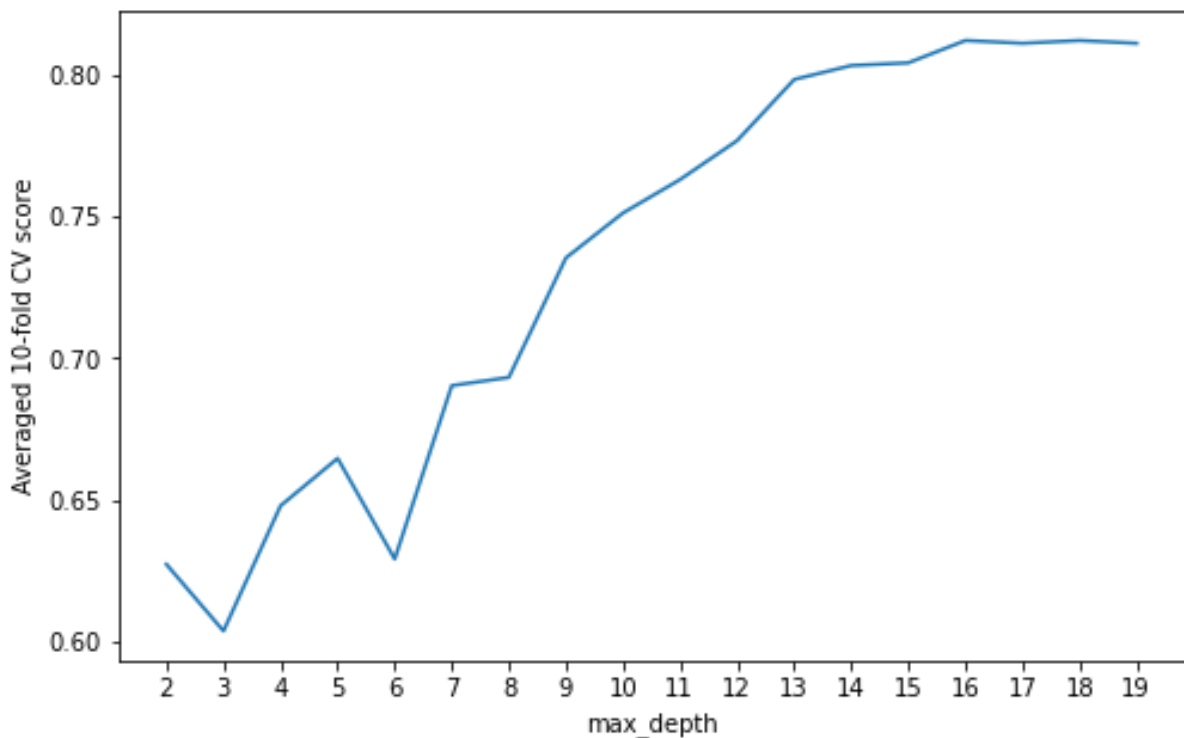


Figure 8: max_depth and averaged 10-Fold cross-validation (CV) score

In Figure 7, as the `max_depth` increases, the accuracy also increases. It peaks at 16, showing that 16 is the most optimal `max_depth` value for the model.

Finding max nodes

Once the optimal max_depth is found, the most optimal leaf nodes for the tree can also be found.

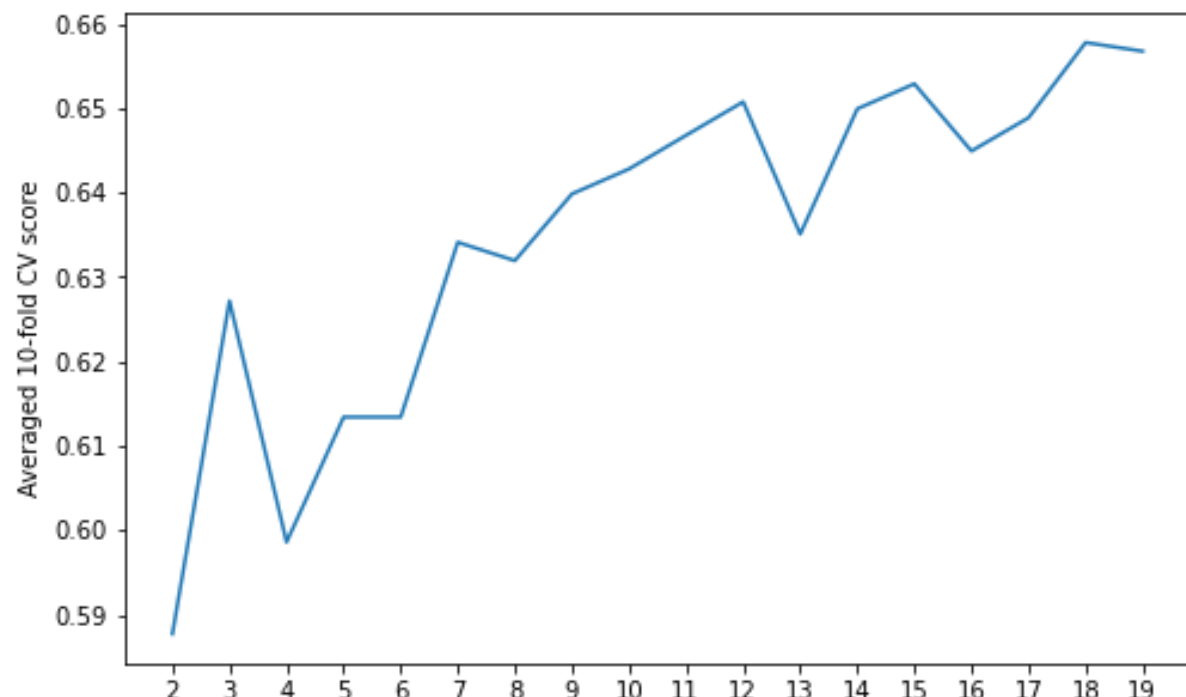


Figure 9: max_leaf_nodes and averaged 10-Fold cross-validation (CV) score

Similarly, as the max_leaf_nodes value increases, the accuracy also increases. The graph peaks at 18, showing that 18 is the most optimal max_leaf_nodes value for the model.

The max_leaf_nodes parameter is responsible for controlling the maximum number of leaf nodes the decision tree can have.

If these values obtained for this dataset were to be used for another dataset at random, it will NOT improve the accuracy. This is because the values used in this decision tree were specifically picked because they optimize this specific tree. All datasets are different, and will have their respective parameters, so the values used on this decision tree are not interchangeable with others.

After passing the optimal max_depth and max_leaf_nodes values into the decision tree, the final optimised tree is obtained with a total of 35 nodes and an overall accuracy of 71%.

c) Confusion matrix

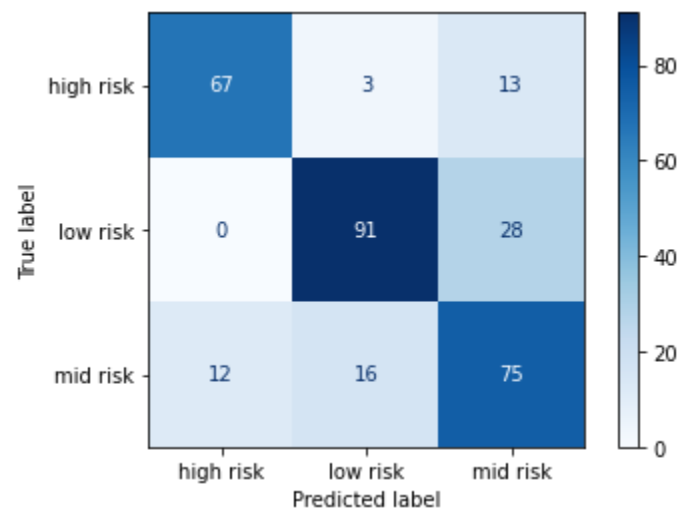


Figure 10: Confusion matrix of the test split

	High Risk	Mid Risk	Low Risk
Accuracy	0.888525	0.727869	0.786885
Precision	0.780822	0.609524	0.732283
Recall	0.760000	0.603774	0.750000
F1 score	0.770270	0.606635	0.741036

Figure 11: Confusion matrix metrics

Accuracy is the ratio of correctly classified instances – true positives and true negatives – out of all instances. It represents the probability that an instance is correctly classified. This model has accuracies that range from 73% to 89%. This means that 73% to 89% of all instances are correctly classified. For the implementation in this report, this is more than adequate. However, if being used in the medical field, the results need to be more reliable, with less variance.

Precision is the ratio of correctly classified positives – true positives – out of all classified positives – true and false positives. It represents the probability that an instance that is classified as positive is correct. This model has precisions that range from 61% to 78%. This means that 61% to 78% of all positives are correctly classified, which means that the model is less reliable at correctly identifying positives.

Recall is the ratio of correctly classified positives – true positives – out of all positives – true positives and false negatives. It represents the probability that an instance that is positive is correctly classified. This model has recalls that range from 60% to 76%. This means that 60% to 76% of all positives are classified as positive, which means that the model is less reliable at identifying positives.

d) Feature importance

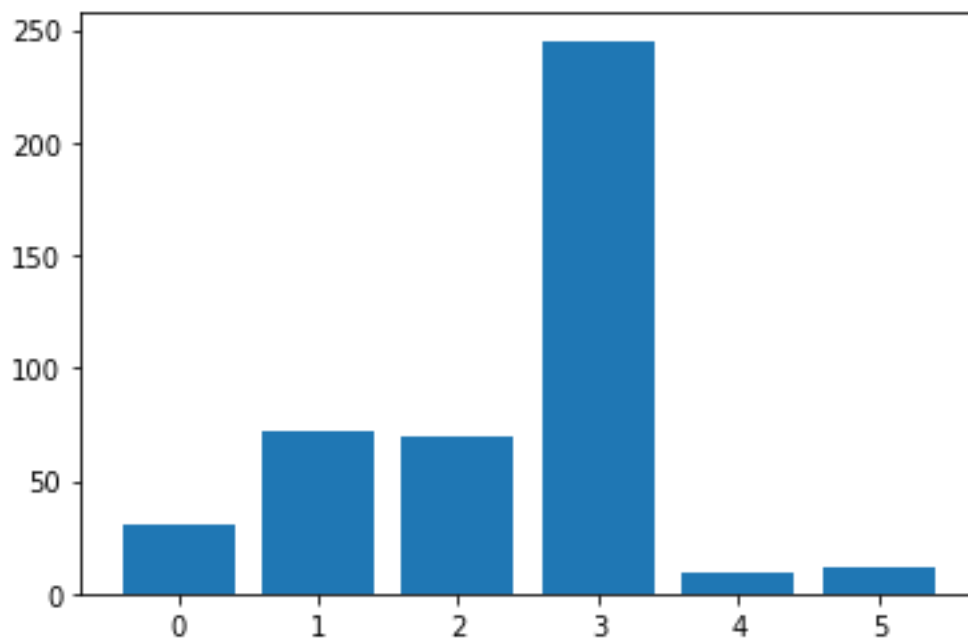


Figure 12: Feature importance graph

	F-score
Age	30.610827
SystolicBP	71.571118
DiastolicBP	69.610876
BS	245.025734
BodyTemp	9.106083
HeartRate	11.454246

Figure 13: F-score table

From the figures above, it is clear that BS is considered to be the most important to the class as it has the highest F-score.

In contrast, Age, BodyTemp, and HeartRate have no correlation to the class, as their scores are the lowest. This means that if these features were to be increased or decreased, there is expected to be no drastic change in the risk level of the patient.

Artificial Neural Network (ANN)

a) Five most significant features

Given that the features are quantitative, an “ANOVA F” value test will be carried out to identify the top five most significant features.

```
Feature 0: 33.702792  
Feature 1: 71.716968  
Feature 2: 66.112841  
Feature 3: 243.905624  
Feature 4: 11.296673  
Feature 5: 18.891135
```

Figure 14: ANOVA F scores

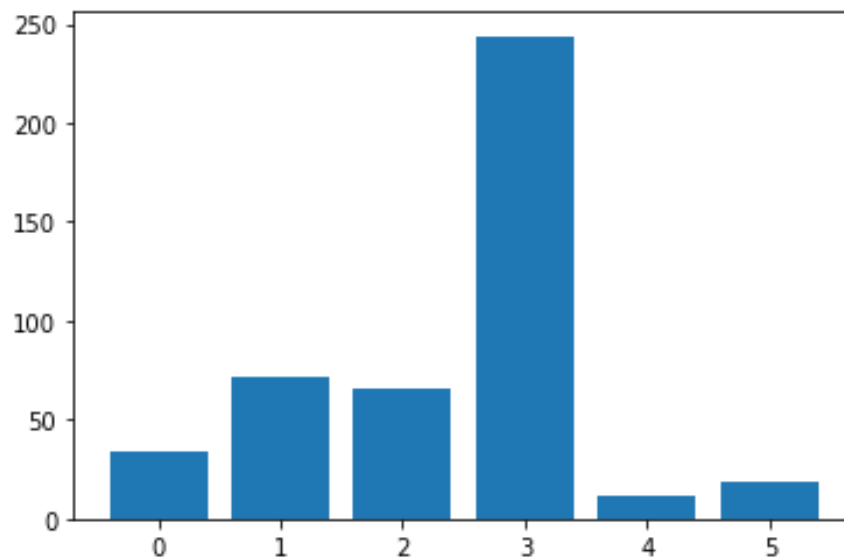


Figure 15: ANOVA F scores

Figure 14 and Figure 15 show the outputs of the ANOVA F-test. They show the significance of the features, with feature 3 showing the most significance out of all 6 features.

This, however, has little meaning unless the feature is known.

	F-score
BS	243.905624
SystolicBP	71.716968
DiastolicBP	66.112841
Age	33.702792
HeartRate	18.891135

Figure 16: Top 5 features

From the F-score table in Figure 15, BS scored an extremely high value, being much higher than every other feature and almost four times larger than SystolicBP. This means that BS has high significance to the class.

Likewise, DiastolicBP is quite high compared to some of the other features, with an F-score of 66. Age and HeartRate have the lowest F-scores, being 33 and 18 respectively. Given that HeartRate has a score of 18, may be reasonable to claim that the feature may not be significant or important to the class.

Compared to the feature importance from the decision tree classifier, there is not much difference between the two classifiers. There is slight variation in the scores, with the variation ranging from 0.2 to 7, but overall, the scores are very similar – the importance of the attributes is the same across both models.

b) One hidden layer

From using the MLPClassifier with default values, it is apparent that a max_iteration value of 50 works the best for this dataset, along with neurons.

MLP Accuracy: 37.38%
MLP Accuracy: 28.52%
MLP Accuracy: 36.39%
MLP Accuracy: 34.10%
MLP Accuracy: 40.00%
MLP Accuracy: 37.38%
MLP Accuracy: 37.05%
MLP Accuracy: 45.90%
MLP Accuracy: 36.39%
MLP Accuracy: 32.79%
MLP Accuracy: 33.11%
MLP Accuracy: 31.15%
MLP Accuracy: 30.49%
MLP Accuracy: 33.11%
MLP Accuracy: 44.92%
MLP Accuracy: 47.54%
MLP Accuracy: 39.02%
MLP Accuracy: 32.46%
MLP Accuracy: 47.54%
MLP Accuracy: 46.56%
MLP Accuracy: 48.20%
MLP Accuracy: 35.74%
MLP Accuracy: 40.33%
MLP Accuracy: 37.38%
MLP Accuracy: 49.84%
MLP Accuracy: 55.08%
MLP Accuracy: 53.11%
MLP Accuracy: 37.38%
MLP Accuracy: 27.54%
MLP Accuracy: 41.64%
MLP Accuracy: 40.66%
MLP Accuracy: 48.85%
MLP Accuracy: 38.36%
MLP Accuracy: 43.28%
MLP Accuracy: 41.31%
MLP Accuracy: 52.46%
MLP Accuracy: 41.31%
MLP Accuracy: 44.26%
MLP Accuracy: 48.85%
MLP Accuracy: 55.74%
MLP Accuracy: 46.23%
MLP Accuracy: 37.70%
MLP Accuracy: 55.74%
MLP Accuracy: 41.97%
MLP Accuracy: 48.20%
MLP Accuracy: 49.18%
MLP Accuracy: 46.56%
MLP Accuracy: 22.95%
MLP Accuracy: 33.11%
MLP Accuracy: 42.62%

Figure 17: Accuracy of different iteration values from 1-50

c) Error and loss value

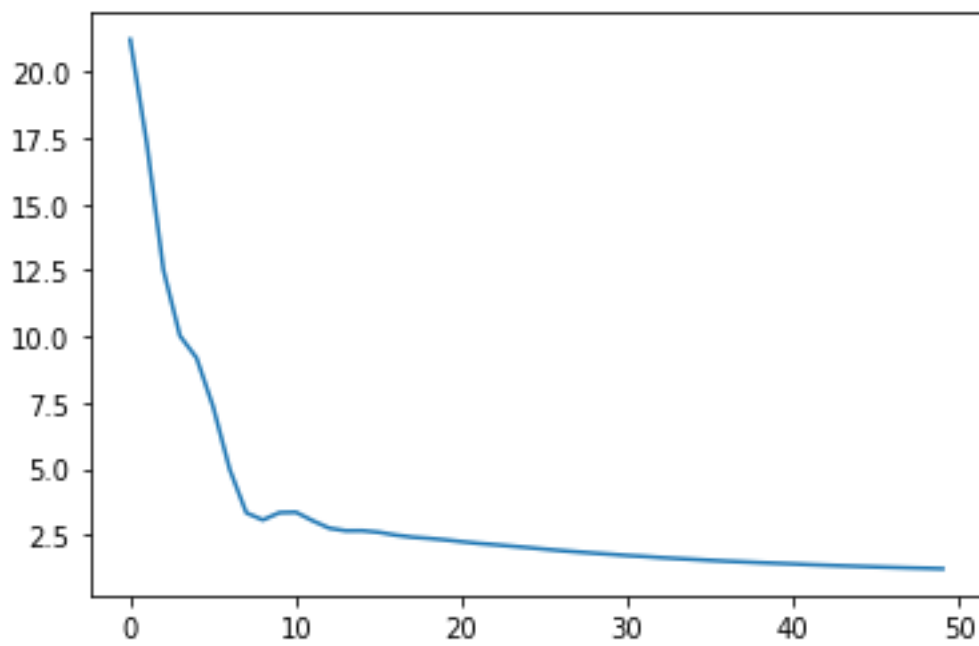



Figure 18: Iterations and loss value

Figure 17 shows the loss values of different numbers of iterations. The graph begins to flatten, and there is little change in the loss value, after iterations = 34.

d) Two hidden layers



	Neuron	MLP Accuracy
0	(24, 1)	0.272131
1	(23, 2)	0.337705
2	(22, 3)	0.337705
3	(21, 4)	0.383607
4	(20, 5)	0.268852
5	(19, 6)	0.413115
6	(18, 7)	0.373770
7	(17, 8)	0.383607
8	(16, 9)	0.357377
9	(15, 10)	0.340984
10	(14, 11)	0.324590
11	(13, 12)	0.485246
12	(12, 13)	0.481967
13	(11, 14)	0.426230
14	(10, 15)	0.357377
15	(9, 16)	0.514754
16	(8, 17)	0.573770
17	(7, 18)	0.459016
18	(6, 19)	0.393443
19	(5, 20)	0.504918
20	(4, 21)	0.393443
21	(3, 22)	0.436066
22	(2, 23)	0.390164
23	(1, 24)	0.390164

Figure 19: Neurons classification table

e) Accuracy variation

Given that the dataset may not be non-linear, there is a possibility that there may be an XOR problem in the dataset which may lead to the accuracy variation.

This means that there needs to be way to classify the non-linear data. This is solved by adding a hidden layer.

Likewise, it is possible to change the number of neurons in each layer to create a more complex model which could be able to pick up on smaller details, increasing the overall accuracy.

By adding more neurons per layer, therefore adding more parameters, the model is able to fit more complex functions.

Performance Comparison

When comparing the performance of the two classification models in terms of their accuracy score, it is clear that the Decision Tree classifier is more successful in terms of accuracy.

```
Accuracy score of our model with Decision Tree: 0.68  
Accuracy score of our model with Neural Network: 0.39
```

From the values above, the decision tree is 29% more accurate than the neural network.

Neither model has a very high accuracy rate, so they may not be the best classification methods. However, if one were to be selected to be used for this dataset, it would most likely be the decision tree as it has the higher accuracy score.

Overall, this report found blood pressure to be a potential determining factor in maternal mortality risk in Bangladesh. This inference may not apply to other countries – for instance, as Bangladesh is a developing country, the living conditions may also be a contributing factor. In other words, more data is needed from a larger population in different regions for this to apply to a more global situation.

Appendix

Below is the code used for this assignment. It will be handed in alongside the assignment document SHOULD something happen, i.e., you are unable to open the document from the appendix.

https://autuni-my.sharepoint.com/:u:/g/personal/fsp7259_autuni_ac_nz/ESI2j5DJWSRFjG_A86nlgLEBsH8eKpOG4ERC6r!pokp4-g?e=l3YpOv