CPC251 Project Part I

HANA SHAH BINTI FAIZAL SHAH (153583), JAZREEL OOI SUE CHING (150740), LOGEN NAIDU S/O SURESH (152860), SITI ADIBAH BINTI ZAINI (152978)

CARDIOTOCOGRAPH

Most fetal deaths happen during the antepartum period which is the period before childbirth. Hence, antenatal assessment of fetal well-being is necessary in order to reduce fetal deaths.

Cardiotocography (CTG) is the most regularly utilised procedure in the modern world to assess fetal well-being. CTG is an ultrasound transducer put on the mother's abdomen that continuously records the fetal heart rate [1]. Its main goal is to prevent negative fetal outcomes.

Despite CTGs having a high sensitivity, it has low specificity which means that it is highly capable of detecting which fetuses are well but not so good at detecting which fetuses are unwell [2]. Therefore, knowing how to evaluate and analyse the state of a fetus promptly and precisely is critical.

Aim: To build an effective and reliable predictive model that is able to classify the fetal state of patients accurately.

Dataset Description

The dataset was obtained based on 2126 cardiotocograms processed and the respective diagnostic features were measured. The dataset was classified by three expert obstetricians and it shows the data on normal, suspicious and pathological fetuses. The dataset contains 23 columns including the cardiotocography information of 2126 fetuses. The target is identified. The target's column name is "NSP" which is the fetal state class code where 1 denotes a normal fetus, 2 denotes a suspicious fetus and 3 denotes a pathological fetus.

	LB	AC	FM	UC	ASTV	MSTV	ALTV	MLTV	DL	DS	...	Max	Nmax	Nzeros	Mode	Mean	Median	Variance	Tendency	CLASS	NSP
0	120	0	0	0	73	0.5	43	2.4	0	0	...	126	2	0	120	137	121	73	1	9	2
1	132	4	0	4	17	2.1	0	10.4	2	0	...	198	6	1	141	136	140	12	0	6	1
2	133	2	0	5	16	2.1	0	13.4	2	0	...	198	5	1	141	135	138	13	0	6	1
3	134	2	0	6	16	2.4	0	23.0	2	0	...	170	11	0	137	134	137	13	1	6	1
4	132	4	0	5	16	2.4	0	19.9	0	0	...	170	9	0	137	136	138	11	1	2	1

Table 1: Samples of the dataset

Data Analysis

The dataset is analyzed in order to gain insight from the dataset. Data visualization technique such as scatter plot is used to analyze the relationship between 22 features. Examples of the plotting are given in Figure 1.

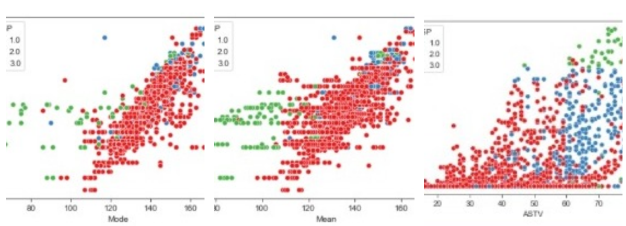

Figure 1: (Left) LB against Mode. (Centre) LB against Mean. (Right) ALTV against ASTV

In order to determine the relationship between the independent variables which are the features and the dependent variable which is the target, bivariate analysis is performed. The bivariate analysis is performed by performing feature selection using ANOVA to get the relevant features that can be used to build an accurate model. ANOVA is chosen as the feature selection method as the features are numerical data and the target is categorical data. Out of 22 features, 11 are selected as the most relevant features based on the scores obtained. Based on the accuracy of both models, these 11 features yield one of the highest accuracies. Hence, 11 features are selected. Table 2 shows the selected features.

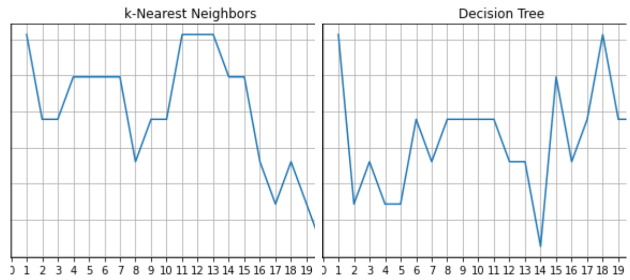

Figure 2: (Left) k-Nearest Neighbors Model Accuracy against Number of Features. (Right) Decision Tree Model Accuracy against Number of Features.

Feature	Type	Value/Statistics
CLASS	Discrete Numerical	Range: 1 - 10 Mean: 4.510 Std: 3.027
MLTV	Continuous Numerical	Range: 0.0 – 50.7 Mean: 8.188 Std: 5.628
DL	Discrete Numerical	Range: 0 - 16 Mean: 1.570 Std: 2.4992
DP	Discrete Numerical	Range: 0 - 4 Mean: 0.126 Std: 0.464
Mode	Discrete Numerical	Range: 60 - 187 Mean: 137.452 Std: 16.381
Mean	Discrete Numerical	Range: 73 - 182 Mean: 134.610 Std: 15.594
Median	Discrete Numerical	Range: 77 - 186 Mean: 138.090 Std: 14.467
AC	Discrete Numerical	Range: 0 - 26 Mean: 2.722 Std: 3.561
Variance	Discrete Numerical	Range: 0 - 269 Mean: 18.808 Std: 28.978
DS	Discrete Numerical	Range: 0 - 1 Mean: 0.0033 Std: 0.0573
LB	Discrete Numerical	Range: 106 - 160 Mean: 133.304 Std: 9.841

Table 2: The selected features

Data Modeling

Two predictive models are built using Decision Tree and K-Nearest Neighbors algorithms. The models are evaluated using the hold-out method. The ratio of the split is 80% training set and 20% test set. The parameters of the predictive models are given in Table 2.

```
Confusion Matrix:
[[325   1   0]
 [  1  66   1]
 [  1   1  30]]
Classification Report:
              precision    recall  f1-score   support

           1       0.99      1.00      1.00       326
           2       0.97      0.97      0.97        68
           3       0.97      0.94      0.95        32

    accuracy                           0.99       426
   macro avg       0.98      0.97      0.97       426
weighted avg       0.99      0.99      0.99       426

Accuracy:
0.9882629107981221
```

Figure 3: Results of classification using K-Nearest Neighbour model

```
Confusion Matrix:
[[326   0   0]
 [  2  66   0]
 [  0   1  31]]
Classification Report:
              precision    recall  f1-score   support

           1       0.99      1.00      1.00       326
           2       0.99      0.97      0.98        68
           3       1.00      0.97      0.98        32

    accuracy                           0.99       426
   macro avg       0.99      0.98      0.99       426
weighted avg       0.99      0.99      0.99       426

Accuracy:
0.9929577464788732
```

Figure 4: Results of classification using Decision Tree model

The predictions are represented by the columns of the 3x3 confusion matrix, and the rows are the actual values that predict whether the case is normal, suspect, or pathologic.

The main diagonal is generated by the K-Nearest Neighbour model (325, 66, 30). The model correctly predicted 325 normal cases, but mispredicted one suspicious case as a normal case, and zero as pathologic cases. Following that, it correctly predicted 66 suspicious cases while mispredicting one as normal and one as a pathologic case. Finally, the K-Nearest Neighbour model correctly predicted 30 pathologic cases while mispredicting one as normal and one as suspicious case.

Meanwhile, the Decision Tree model's main diagonal (326, 66, 31) correctly predicts 326 normal cases, 66 suspicious cases, and 31 pathologic cases. The number of predictions differs slightly from the K-Nearest Neighbour model. The Decision Tree model correctly predicted 326 normal cases while mispredicting none as suspicious or pathologic cases. It correctly predicted 66 suspicious cases while incorrectly predicting two as normal and zero as pathologic cases. Finally, the Decision Tree model predicted 31 pathologic cases correctly while mispredicting zero as normal and one as a suspicious case.

The Decision Tree model outperforms the K-Nearest Neighbour model when the two results are compared. The number of incorrect predictions is lower than that of the K-Nearest Neighbour model. The decision tree incorrectly predicted three cases, and the K-Nearest Neighbour model incorrectly predicted five.

By using the K-Nearest Neighbour model, the precision value for the three classes are 0.99, 0.97 and 0.97 with an average of 0.98. The recall value obtained for the three classes are 1.00, 0.97 and 0.94. On the other hand, by using the Decision Tree model, the precision value for the three classes are 0.99, 0.99 and 1.00 with an average of 0.99. The recall value obtained for the three classes are 1.00, 0.97 and 0.97.

We have obtained the same value of precision for the first class, which is 1.00 by using the Decision Tree model and the K-Nearest Neighbour model. Thus, the results obtained mean that both models have the same ability to correctly classify the first class. Furthermore, we have obtained a higher value of precision for the second class, which is 0.99 by using the Decision Tree compared to 0.97 by using the K-Nearest Neighbour model. Therefore, the results show that the Decision Tree model has a better ability to correctly classify the second class. Lastly, we have obtained a higher value of precision for the third class, which is 1.00 by using the Decision Tree model compared to 0.97 by using the K-Nearest Neighbour model.

Hence, the results above tell us that the Decision Tree model has a better ability to correctly classify the third class. Both are non-parametric methods. The decision tree allows for automatic feature interaction, whereas KNN does not. Because of

KNN's expensive real-time execution, the decision tree is faster.

The same recall value for the first class, which is 1.00 was obtained by using the K-Nearest Neighbour and the Decision Tree model. This result shows us that both models have predicted the same number of outcomes correctly. Furthermore, the same recall value for the second class, which is 0.97 was obtained by using the K-Nearest Neighbour model and the Decision Tree model. This result shows us that both models have predicted the same number of outcomes correctly. Lastly, a lower recall value for the third class, which is 0.94 was obtained by using the K-Nearest Neighbour model compared to 0.97 by using the Decision Tree model. This result shows us that the K-Nearest Neighbour model has predicted a lower number of outcomes correctly.

The Decision Tree algorithm is the best algorithm to be used since the values of the F1 score for every class are closer to 1.0. This means that this model will give us a better expected performance compared to the K-Nearest Neighbour model.

References

[1] National Library of Medicine, National Center of Biotechnology Information. (September 12, 2015). *Antenatal cardiotocography for fetal assessment*.

[2] The Royal Women's Hospital Victoria, Australia. (July 27, 2020). *CTG Interpretation and Response*.

Credits:

Created with an image by Claudio Divizia - "light blue paper texture background"