**Assignment 5**

**Author:** Hamna Ashraf, Student # 8826836

**Course:** PROG8431

**Question 1: Make sure that the data is free from outliers or unnecessary data.**

A close-up of a white background

Description automatically generated

There are no missing values in the dataset as no column has any null values.

**Question 2**

**2.1: Create numeric correlations and comment on what you see. Are there co-linear variables?**

A graph with numbers and lines

Description automatically generated

From this correlation matrix, we can see that there is a moderate positive correlation of features ‘Marrow’ and ‘Brain’ with the target feature which is ‘Out’ with correlation of 0.33 and 0.47 respectively. This means that these features have a direct relationship with the presence of tumor compared to others.

**2.2: Identify the two most significant predictors of tumors and provide statistical evidence**

As we already saw in question 2.1 earlier that the features Brain and Marrow has the highest correlation with feature Out (tumor). However, it is required to show other statistical evidences as well so we can use contingency tables for it.

A screenshot of a computer

Description automatically generated

The Chi-squared value shows that the Brain feature has a much stronger association with presence of tumor than Marrow however, Marrow feature also has a high chi-squared stats and has association with presence of Tumor.

A very small p-value also suggests that both Brain and Marrow are significant predictors of feature tumor.

**Question 3**

**3.2: Summarize the results in a Confusion Matrix**

**A close up of numbers

Description automatically generated**

True Negatives: 121

True Positives: 111

False Negatives: 39

False Positives: 29

**3.3: Based on the confusion matrix, calculate and comment on:**

a. Accuracy b. Specificity c. Sensitivity d. Precision

**A screenshot of a computer code

Description automatically generated**

**Accuracy:** The accuracy score is high which suggests that the performance is good.

**Specificity:** The model identifies patterns without tumors, however, there are still improvements that can be made.

**Sensitivity:** Sensitivity, which is also called recall, needs to be higher for these types of problems. We don’t want our model to predict no tumor for patient that actually has a tumor. The value we got is high which is almost 80% but can still be improved.

**Precision:** This tells us that 74% of the times our model prediction about having a tumor is correct.

**3.4: Create the ROC curve and calculate the AUC. Comment on how you interpret them.**

**A graph with a red line

Description automatically generated**

ROC Curve:

The ROC curve of the model rises quickly towards the top left corner. This indicates a high true positive rate which is a characteristic of a good classifier.

AUC:

As AUC value is 0.878, the model shows a strong ability to classify the positive class correctly. This means that the model has the capability to give a correct prediction is case of a tumor.

**3.5: Calculate the time (in seconds) it took to train the model and the prediction time.**

A number and numbers on a white background

Description automatically generated

The training time is found to be 0.2476561 and prediction time is 0.001332045.

**Question 4**:

**4.1:**

The categorial variables are converted into factors.

**4.3:**

**A white background with black text

Description automatically generated**

True Negatives: 122

True Positives: 114

False Negatives: 36

False Positives: 28

**4.4:**

A screenshot of a computer code

Description automatically generated

**Accuracy:** This is high and 78.67% of the predictions made by the model are correct. It is a good prediction.

**Specificity:** 77% of the times model identifies patterns without tumor i.e it calculates the proportion of actual negatives correctly.

**Sensitivity:** This has a high recall value which is around 80% which is good but we need to make the recall value as high as we can.

**Precision:** This tells us that 76% of the times our model prediction about having a tumor is correct.

**4.5:**

Training time: 0.1041009

Prediction Time: 0.00395298

**Question 5:**

**5.1:**

The transformation is already done.

**5.3:**

A white background with black text

Description automatically generated

True Negatives: 120

True Positives: 115

False Negatives: 35

False Positives: 30

**5.4:**

**A screenshot of a computer code

Description automatically generated**

**Accuracy:** This is high and 78.33% of the predictions made by the model are correct. It is a good prediction.

**Specificity:** 77% of the times model identifies patterns without tumor i.e it calculates the proportion of actual negatives correctly.

**Sensitivity:** This has a high recall value which is around 79% which is good but we need to make the recall value as high as we can.

**Precision:** This tells us that 76% of the times our model prediction about having a tumor is correct.

**5.5:**

Training time: 0.007757187

Prediction Time: 0.008302927

**Question 6:**

**6.1: Which classifier is most accurate?**

Naïve Bayes and Linear Discriminant Analysis (LDA) both exhibit higher accuracy rates than Logistic Regression, with Naïve Bayes achieving an accuracy of approximately 78.67% and LDA achieving around 78.33%, surpassing the Logistic Regression accuracy of roughly 77.33%.

**6.2 Which classifier is most suitable when processing speed is most important?**

In terms of training time, LDA is notably the most efficient, taking just about 0.0077 seconds, which is significantly less than the time taken by Logistic Regression, approximately 0.2476 seconds, and Naïve Bayes, which is around 0.1041 seconds. On the other hand, when it comes to prediction time, Logistic Regression leads with a minimal 0.001 seconds, while Naïve Bayes follows with 0.003 seconds, and LDA lags at 0.008 seconds.

**6.3 Which classifier minimizes Type 1 errors?**

Specificity, which indicates the proportion of true negatives correctly identified and thus reflects the rate of Type 1 errors, or false positives, is highest for LDA at 77.42%, followed by Naïve Bayes at 77.22%, with Logistic Regression at 75.63%.

**6.4 Which classifier minimizes Type 2 errors?**

Sensitivity, conversely, represents the proportion of true positives correctly identified, relating to the rate of Type 2 errors, or false negatives. Here, Naïve Bayes tops the chart with a sensitivity rate of 80.28%, LDA comes in next with approximately 79.31%, and Logistic Regression has the lowest at roughly 79.29%.