

Migraine Classification Using Data Mining Techniques

Milestone: Final Project Report

Group 1

Aditya Kakde
Dhruv Bhardwaj

+1-857-313-4704 (Aditya Kakde)
+1-857-313-4048 (Dhruv Bhardwaj)

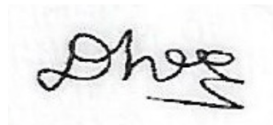
kakde.a@northeastern.edu
bhardwaj.dh@northeastern.edu

Percentage of Effort Contributed by Aditya: _____ 50% _____

Percentage of Effort Contributed by Dhruv: _____ 50% _____



Signature of Student 1: _____



Signature of Student 2: _____

Submission Date: _____ 5th Apr 2023 _____

Table of Contents

- 1. Problem Setting**
- 2. Problem Description**
- 3. Data Sources**
- 4. Data Description**
- 5. Exploratory Data Analysis**
- 6. Feature Selection**
- 7. Model Performance Evaluation**
- 8. Model Selection**
- 9. Conclusion**

Problem Setting

Migraine is not just a bad headache. It's a disabling neurological disorder with different symptoms and different treatment approaches compared to other headache disorders. The American Migraine Foundation estimates that at least 39 million Americans live with migraine, but because many people do not get a diagnosis or the treatment, they need the actual number is probably higher. Although migraine is a common disease with substantial impact, it is under diagnosed and under treated. Paucity of relevant literature to begin migraine diagnosis and proper classification of the type of migraine most often leads to under diagnosis and under treatments or maybe even delay the diagnosis.

Problem Definition

Every migraine episode presents with a series of 'features' or symptoms. Doctors often use these feature sets to predict the type of migraine the patient is suffering from in order to prescribe relevant treatment. More often than not, this process of identifying the type of migraine takes a substantial amount of time or often leads to misdiagnosis. The aim of this project is to study the various features that a migraine episode brings along with it and create a classification model using statistical and data mining techniques. The key question that needs to be answered is: Given these set of symptoms, what type of migraine is the patient presenting ?

Data Sources

The dataset is available on <https://www.kaggle.com/code/azazurrehmanbutt/migraine-prediction-dnns-99-acc/data>

Data Description

Database comprising 400 medical records of users diagnosed with various pathologies associated with migraines. Data were recorded by trained medical personnel at the American Migraine Foundation during the first quarter of 2013. The compiled database contains information regarding symptoms or variable of interest required for the classification of migraines.

The dataset contains information on 400+ migraine episodes along with 24 features. Presence of a feature is denoted by '1' while absence is marked as '0'. The response variable of this study would be the type of migraine based on the 24 features.

Attribute Information:

- 1) Age: Patient's age
- 2) Duration: duration of symptoms in last episode in days
- 3) Frequency: Frequency of episodes per month
- 4) Location: Unilateral or bilateral pain location (None - 0, Unilateral - 1, Bilateral - 2)
- 5) Character: Throbbing or constant pain (None - 0, Thobbing - 1, Constant - 2)
- 6) Intensity: Pain intensity, i.e., mild, medium, or severe (None - 0, Mild - 1, Medium - 2, Severe - 3)
- 7) Nausea: Nauseous feeling (Not - 0, Yes - 1)
- 8) Vomit: Vomiting (Not - 0, Yes - 1)
- 9) Phonophobia: Noise sensitivity (Not - 0, Yes - 1)
- 10) Photophobia: Light sensitivity (Not - 0, Yes - 1)
- 11) Visual: Number of reversible visual symptoms
- 12) Sensory: Number of reversible sensory symptoms
- 13) Dysphasia: Lack of speech coordination (Not - 0, Yes - 1)
- 14) Dysarthria: Disarticulated sounds and words (Not - 0, Yes - 1)
- 15) Vertigo: Dizziness (Not - 0, Yes - 1)
- 16) Tinnitus: Ringing in the ears (Not - 0, Yes - 1)
- 17) Hypoacusis: Hearing loss (Not - 0, Yes - 1)
- 18) Diplopia: Double vision (Not - 0, Yes - 1)
- 19) Visual defect: Simultaneous frontal eye field and nasal field defect and in both eyes (Not - 0, Yes - 1)
- 20) Ataxia: Lack of muscle control (Not - 0, Yes - 1)
- 21) Conscience: Jeopardized conscience (Not - 0, Yes - 1)
- 22) Paresthesia: Simultaneous bilateral paresthesia (Not - 0, Yes - 1)
- 23) DPF: Family background (Not - 0, Yes - 1)
- 24) Type: Diagnosis of migraine type (Typical aura with migraine, Migraine without aura, Typical aura without migraine, Familial hemiplegic migraine, Sporadic hemiplegic migraine, Basilar-type aura, Other) [247, 60, 20, 24, 14, 18, 17]

Exploratory Data Analytics

Couple of things were checked after loading the dataset. First thing that was checked was the descriptive statistics of each column in the dataset. The count column would give us a rough idea if there were any null values in any columns. Below is the description of the dataset.

	Age	Duration	Frequency	Location	Character	Intensity	Nausea	Vomit	Phonophobia	Photophobia	...
count	400.000000	400.000000	400.000000	400.000000	400.000000	400.000000	400.000000	400.000000	400.000000	400.000000	...
mean	31.705000	1.610000	2.365000	0.972500	0.977500	2.470000	0.987500	0.322500	0.977500	0.980000	...
std	12.139043	0.770964	1.675947	0.268186	0.277825	0.768490	0.111242	0.468019	0.148489	0.140175	...
min	15.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...
25%	22.000000	1.000000	1.000000	1.000000	1.000000	2.000000	1.000000	0.000000	1.000000	1.000000	...
50%	28.000000	1.000000	2.000000	1.000000	1.000000	3.000000	1.000000	0.000000	1.000000	1.000000	...
75%	40.000000	2.000000	4.000000	1.000000	1.000000	3.000000	1.000000	1.000000	1.000000	1.000000	...
max	77.000000	3.000000	8.000000	2.000000	2.000000	3.000000	1.000000	1.000000	1.000000	1.000000	...

8 rows x 23 columns

Fig 1. Gist of the Dataset Summary

We can see that all the columns have 400 values, which indicates that there are no null values in the dataset. However to support this assumption, we run a quick null count test on the dataset to get the following results.

```

Age      0
Duration 0
Frequency 0
Location 0
Character 0
Intensity 0
Nausea    0
Vomit     0
Phonophobia 0
Photophobia 0
Visual    0
Sensory   0
Dysphasia 0
Dysarthria 0
Vertigo   0
Tinnitus  0
Hypoacusis 0
Diplopia  0
Visual_defect 0
Ataxia    0
Conscience 0
Paresthesia 0
DPF       0
Type      0
dtype: int64

```

Fig 2. Features present in the data

This confirms that there are no null values present in any of the features present in the dataset.

In our problem statement, we are trying to classify the type of migraine based on the presence or absence of certain combination of features. For this, it is essential to know the various types of classes we have in our response variable ('Type'). Below we can the unique classes present in the response variable's domain.

```
array(['Typical aura with migraine', 'Migraine without aura',  
      'Basilar-type aura', 'Sporadic hemiplegic migraine',  
      'Familial hemiplegic migraine', 'Other',  
      'Typical aura without migraine'], dtype=object)
```

Fig 3. Different categories of Migraines

Let's look at the distribution of age in the dataset

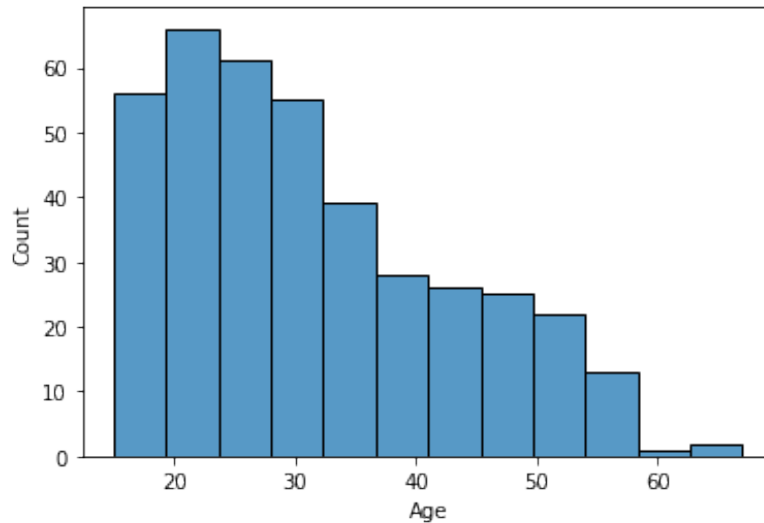


Fig 4. Age distribution in the dataset

We can see the most of the patients in the dataset are youngsters between the age of 20-30 years.

Let's look at the distribution of duration of the migraine episodes in the dataset.

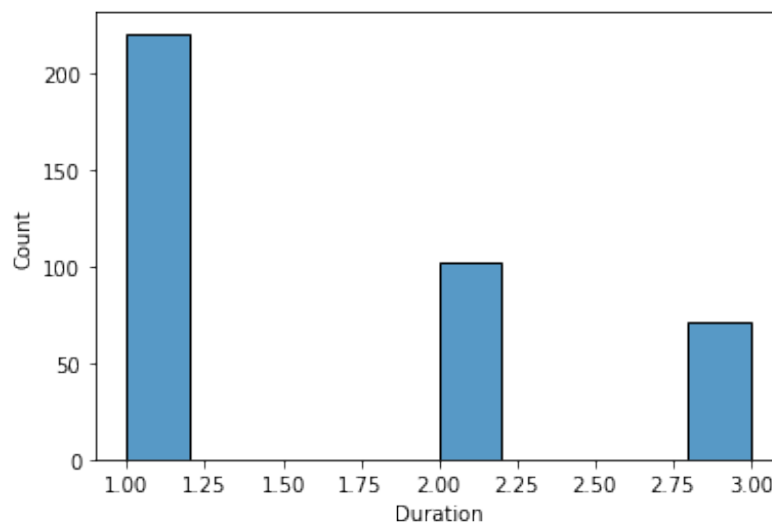


Fig 5. Distribution of Migraine Durations

Through this visualisation we can see that most of the patients have episodes that last about 1hr to 90mins.

Feature Selection

Since the dataset has 24 features, we need to reduce the features that do not have much importance when it comes to having an impact on the response variable. Thus for this we can follow 3 approaches :

1. Univariate Analysis
2. Feature Importance
3. Correlation of features using heatmaps

Univariate Analysis

According to Univariate Analysis the top 20 features of this dataset are presented below

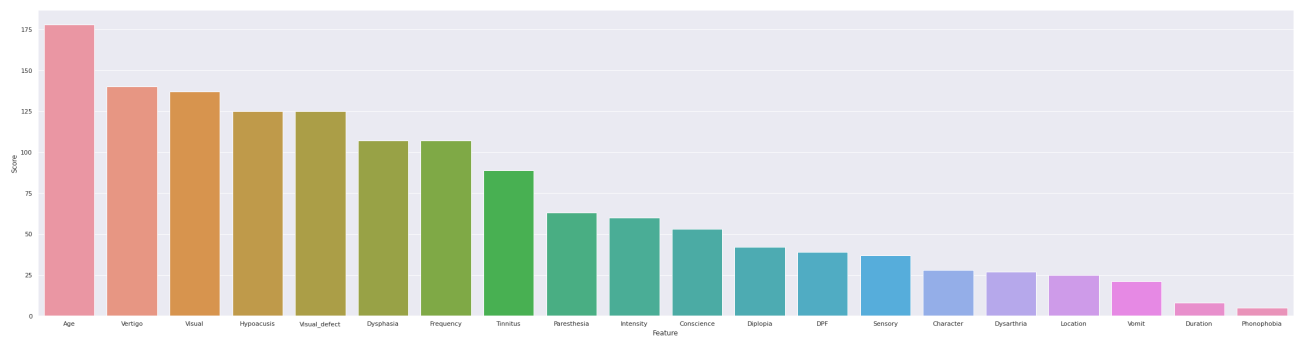


Fig 6. Top 20 Features present in the Dataset

Statistical tests can be use to select the features that have the strongest relationship with the output variable. The above example was computed using the Chi-squared test for non-negative features to select the top 10 features from the dataset.

Feature Importance

The below graph presents the graph of the weight of importance of each feature in the dataset.

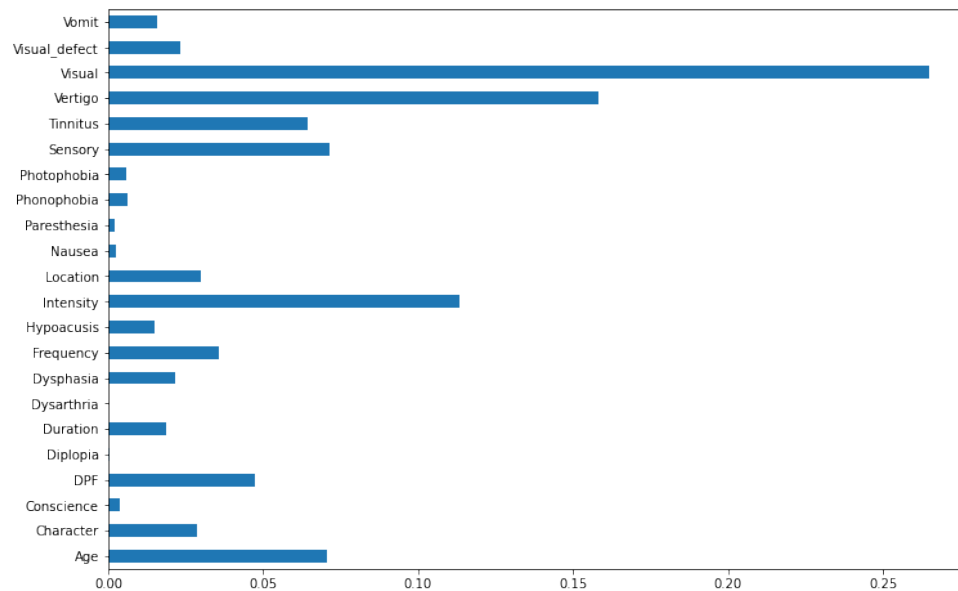


Fig 7. Feature Importance Chart

From the graph those which have very low importance can be omitted from consideration while developing the model. Feature importance gives you a score for each of the feature in the dataset. Higher scores means more important the feature is towards your output variable.

So through this analysis we can infer that features like Diplopia, Nausea, Phonophobia, Photophobia, Conscience, Dysarthria are low importance features and thus can be excluded from the dataset while model building.

Model Performance Evaluation

Three algorithms were chosen to go ahead with for making the models. According to the previous step Decision Trees, Gaussian NB and Random Forest algorithms gave the best results for the dataset.

Decision Tree –

The Decision Tree model was built to classify the 7 types of migraine episodes. Using sklearn.tree library with the help of DecisionTreeClassifier the models were built.

The data was split into training and testing sets in a 70:30 fashion. The model was trained on the 70% of the data and was fitted on the same. For the model evaluation phase we used the remaining 30% of data.

	precision	recall	f1-score	support
0	1.00	0.83	0.91	6
1	0.83	1.00	0.91	5
2	0.93	1.00	0.96	13
3	1.00	0.75	0.86	4
4	1.00	1.00	1.00	5
5	0.99	1.00	0.99	78
6	1.00	0.88	0.93	8
accuracy			0.97	119
macro avg	0.96	0.92	0.94	119
weighted avg	0.98	0.97	0.97	119

Fig 8. Decision tree Classification Report

The above classification report shows the various performance evaluation metrics like Precision, Recall, F1-Score and Support for all the 7 classes. We can see the overall accuracy of the model was found around 0.97 or 97% using a Decision Tree Model.

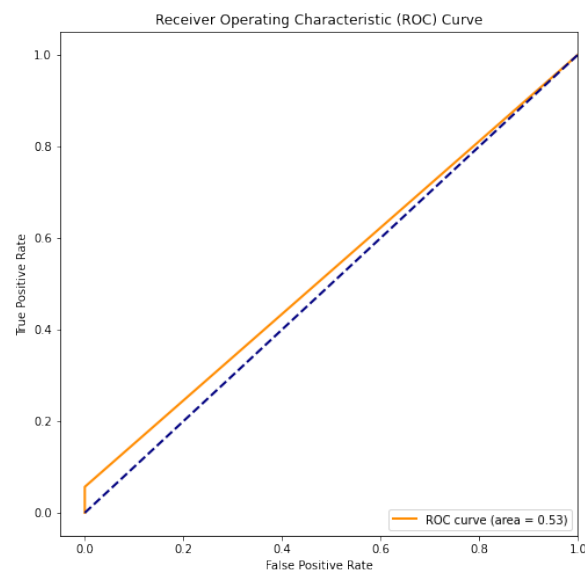


Fig 9. ROC Curve for Decision Tree

Random Forest Model –

The Random Forest model was built to classify the 7 types of migraine episodes. Using sklearn.ensemble library with the help of RandomForestClassifier the models were built.

The data was split into training and testing sets in a 70:30 fashion. The model was trained on the 70% of the data and was fitted on the same. For the model evaluation phase we used the remaining 30% of data.

	precision	recall	f1-score	support
0	1.00	0.83	0.91	6
1	0.83	1.00	0.91	5
2	0.93	1.00	0.96	13
3	1.00	0.50	0.67	4
4	1.00	0.60	0.75	5
5	0.96	1.00	0.98	78
6	1.00	1.00	1.00	8
accuracy			0.96	119
macro avg	0.96	0.85	0.88	119
weighted avg	0.96	0.96	0.95	119

Fig 10. Classification report for Random Forest

The above classification report shows the various performance evaluation metrics like Precision, Recall, F1-Score and Support for all the 7 classes. We can see the overall accuracy of the model was found around 0.96 or 96% using a Random Forest Model.

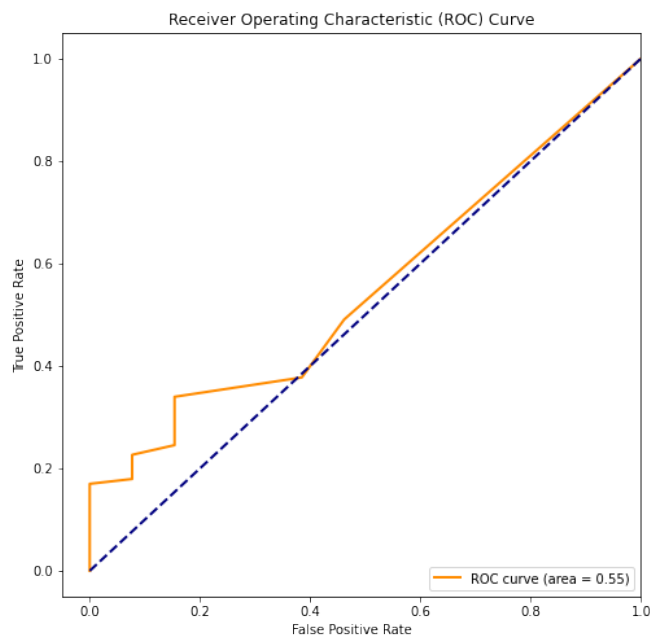


Fig 11. ROC Curve for Random Forest

Gaussian NB Model –

The Gaussian NB model was built to classify the 7 types of migraine episodes. Using `sklearn.naive_bayes` library with the help of `GaussianNB` the models were built.

The data was split into training and testing sets in a 70:30 fashion. The model was trained on the 70% of the data and was fitted on the same. For the model evaluation phase we used the remaining 30% of data.

	precision	recall	f1-score	support
0	1.00	0.67	0.80	6
1	0.71	1.00	0.83	5
2	0.93	1.00	0.96	13
3	0.67	0.50	0.57	4
4	0.83	1.00	0.91	5
5	1.00	0.99	0.99	78
6	1.00	1.00	1.00	8
accuracy			0.96	119
macro avg	0.88	0.88	0.87	119
weighted avg	0.96	0.96	0.96	119

Fig 12. Classification Report for Gaussian NB

The above classification report shows the various performance evaluation metrics like Precision, Recall, F1-Score and Support for all the 7 classes. We can see the overall accuracy of the model was found around 0.96 or 96% using a Gaussian NB Model.

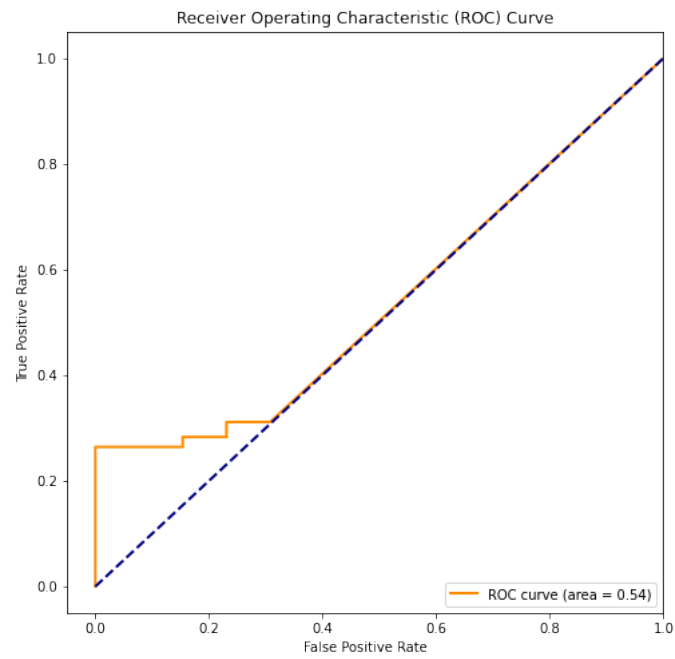


Fig 13. ROC Curve for Gaussian NB

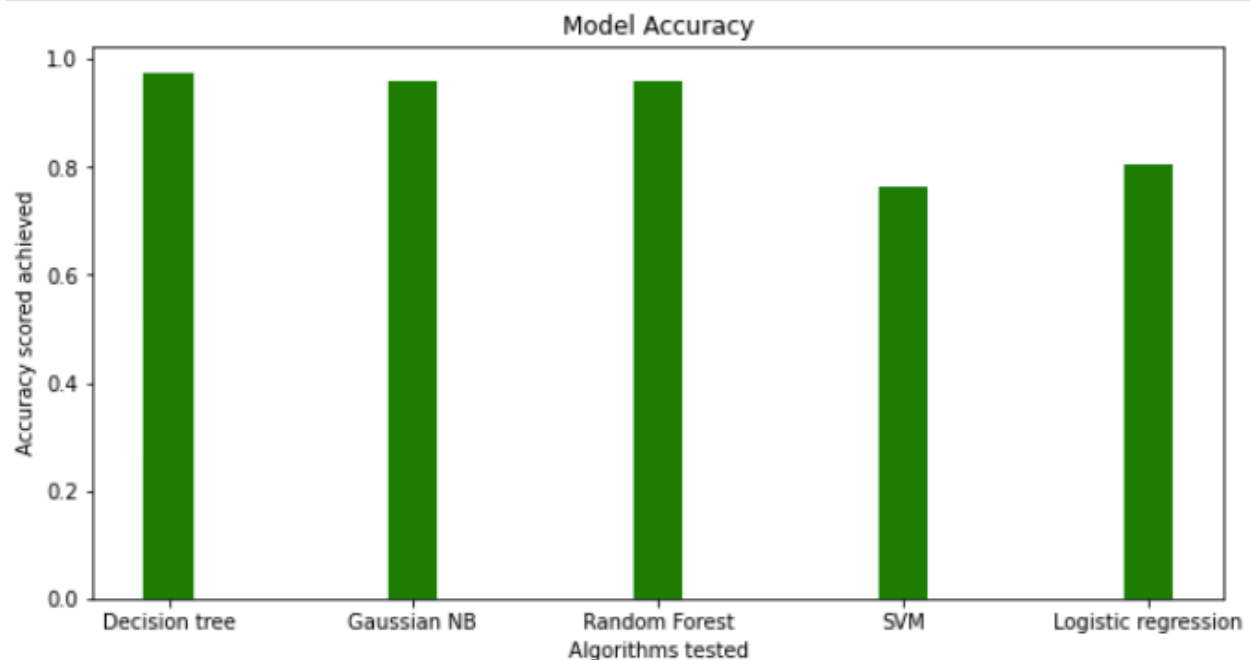
Therefore among the 3 models we can see that the Random Forest Model does the best.

Model Performance Comparison Table

Model	Accuracy
Decision Tree	97%
Random Forest	96%
Gaussian NB	96%

Model Selection

Since the problem is of a multi-class classification type, couple of models were tried out to see which gives the best accuracy for the cleaned data. Below we see the performance of various models that were tried out, compared to each other.



We can clearly see that models like Decision Trees, Random Forests and Gaussian Naïve Bayes give the best performance of greater than 0.95.

These models can be chosen as proper candidates for future modelling and tuning purposes.

Therefore we can choose – Decision trees, Random Forests and Gaussian Naïve Bayes as the models to go forward with in creating the models.

Conclusion

In conclusion we can see that the Decision Tree models performs the best among the three models that were tried out. The Decision Tree helps in categorising the datapoints, into one of the 7 categories of migraine based on the set of input predictors.

An extension of these models can be used to aid clinical analysis of migraines and help with the early diagnosis and prophylaxis of various types of migraines.