Capstone Project Final Report

Craig Leavitt

2022-06-21

1. Introduction

This report is prepared for the final capstone project of the Data Science certificate program offered by HarvardX. The report details analysis and modeling of a migraine classification dataset available for download on the Kaggle website (https://www.kaggle.com/datasets/weinoose/migraine-classification). Each observation in the dataset consists of twenty three (23) numerical variables and a final migraine type (string) classification. The dataset is rather small with only 400 observations.

1.1 Dataset Analysis

Number of Observations:

```
## [1] 400
```

Data Columns and Labels:

```
##  [1] "Age"         "Duration"    "Frequency"   "Location"    "Character"
##  [6] "Intensity"   "Nausea"      "Vomit"       "Phonophobia" "Photophobia"
## [11] "Visual"      "Sensory"     "Dysphasia"   "Dysarthria"  "Vertigo"
## [16] "Tinnitus"    "Hypoacusis"  "Diplopia"    "Defect"      "Ataxia"
## [21] "Conscience"  "Paresthesia" "DPF"         "Type"
```

```
## [1] "Typical aura with migraine"    "Migraine without aura"
## [3] "Basilar-type aura"             "Sporadic hemiplegic migraine"
## [5] "Familial hemiplegic migraine"  "Other"
## [7] "Typical aura without migraine"
```

Example Observations:

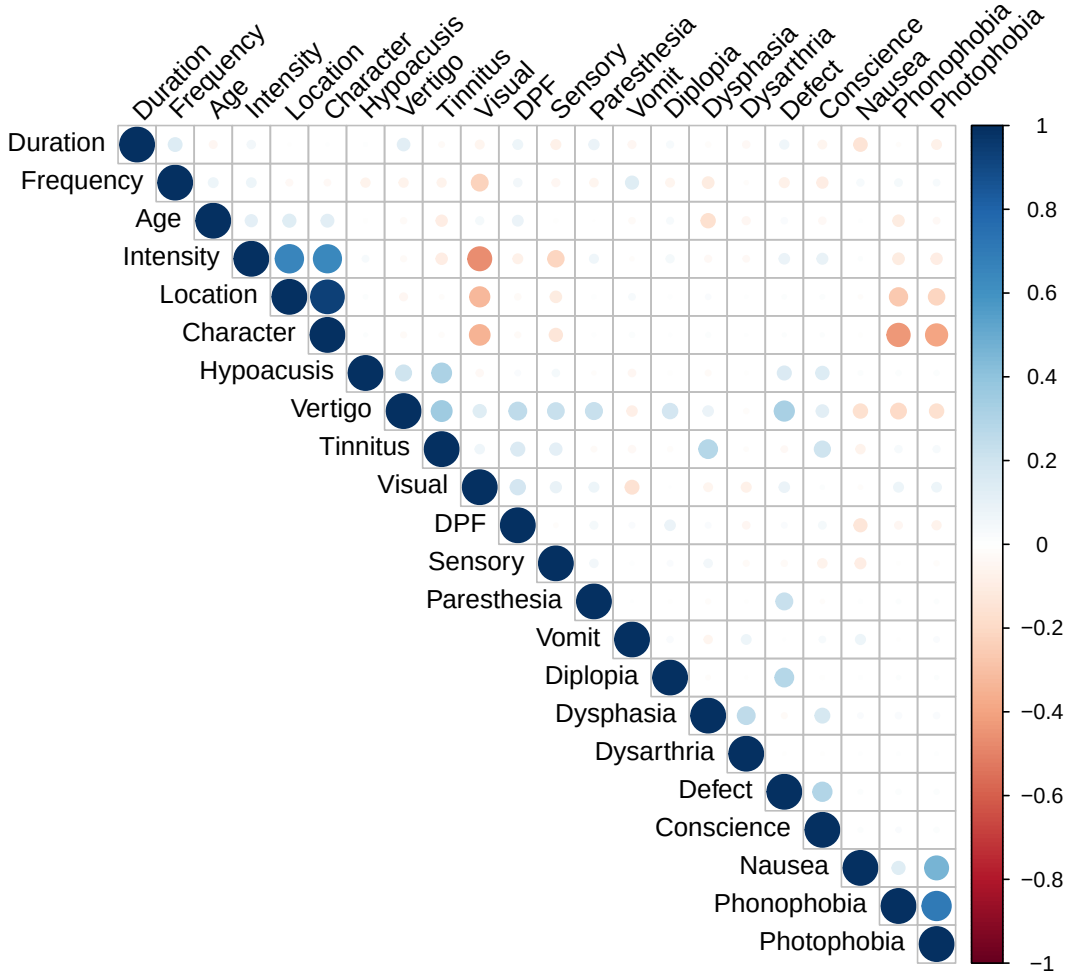
```
## # A tibble: 6 x 24
##     Age Duration Frequency Location Character Intensity Nausea Vomit Phonophobia
##   <dbl>    <dbl>     <dbl>    <dbl>     <dbl>     <dbl>  <dbl> <dbl>       <dbl>
## 1    30        1         5        1         1         2      1     0           1
## 2    50        3         5        1         1         3      1     1           1
## 3    53        2         1        1         1         2      1     1           1
## 4    45        3         5        1         1         3      1     0           1
## 5    53        1         1        1         1         2      1     0           1
## 6    49        1         1        1         1         3      1     0           1
## # ... with 15 more variables: Photophobia <dbl>, Visual <dbl>, Sensory <dbl>,
```

```
## #   Dysphasia <dbl>, Dysarthria <dbl>, Vertigo <dbl>, Tinnitus <dbl>,
## #   Hypoacusis <dbl>, Diplopia <dbl>, Defect <dbl>, Ataxia <dbl>,
## #   Conscience <dbl>, Paresthesia <dbl>, DPF <dbl>, Type <chr>
```

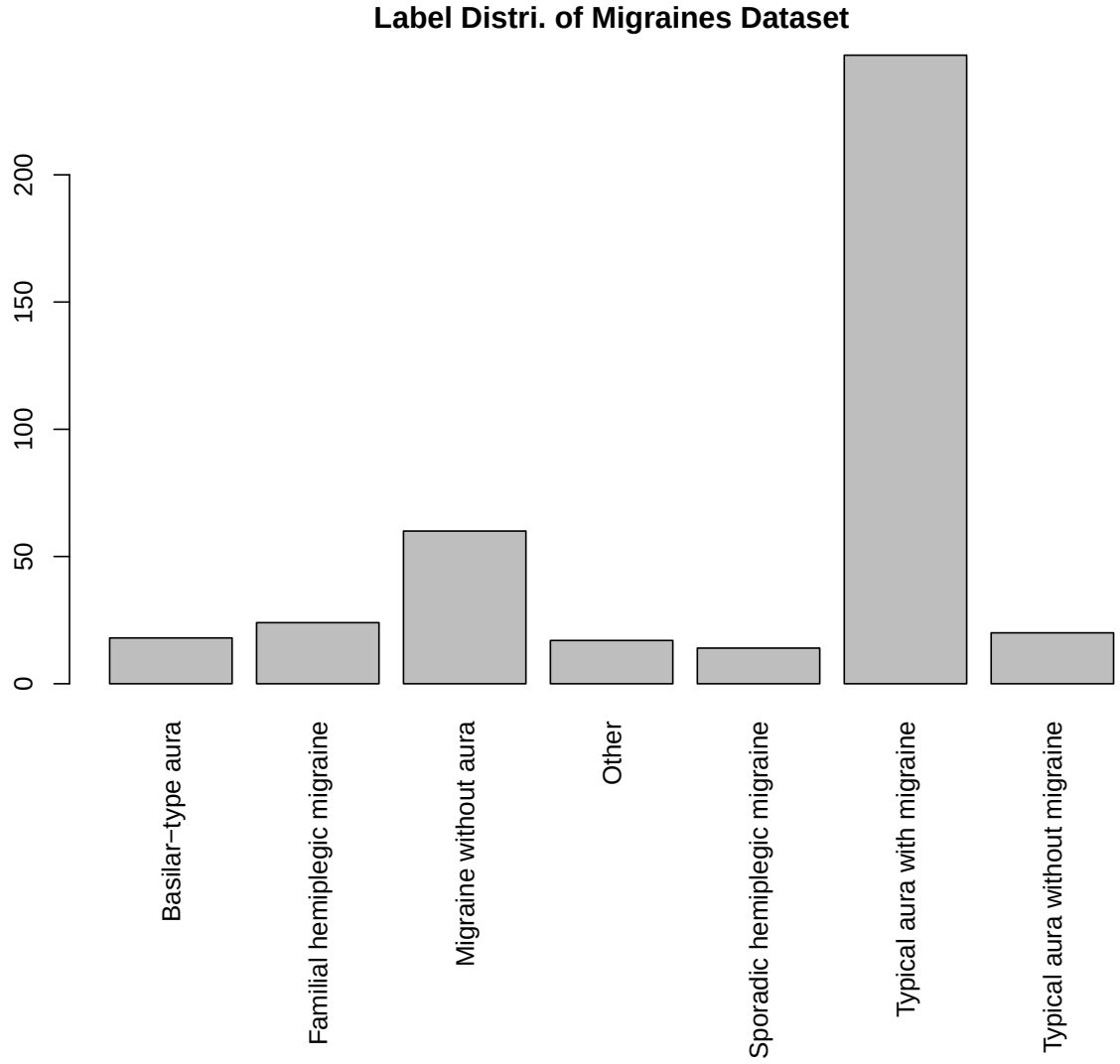
Data Characteristics:

```
##       Age          Duration      Frequency        Location
##  Min.   :15.0   Min.   :1.00   Min.   :1.000   Min.   :0.0000
##  1st Qu.:22.0   1st Qu.:1.00   1st Qu.:1.000   1st Qu.:1.0000
##  Median :28.0   Median :1.00   Median :2.000   Median :1.0000
##  Mean   :31.7   Mean   :1.61   Mean   :2.365   Mean   :0.9725
##  3rd Qu.:40.0   3rd Qu.:2.00   3rd Qu.:4.000   3rd Qu.:1.0000
##  Max.   :77.0   Max.   :3.00   Max.   :8.000   Max.   :2.0000
##    Character        Intensity        Nausea           Vomit
##  Min.   :0.0000   Min.   :0.00   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:1.0000   1st Qu.:2.00   1st Qu.:1.0000   1st Qu.:0.0000
##  Median :1.0000   Median :3.00   Median :1.0000   Median :0.0000
##  Mean   :0.9775   Mean   :2.47   Mean   :0.9875   Mean   :0.3225
##  3rd Qu.:1.0000   3rd Qu.:3.00   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :2.0000   Max.   :3.00   Max.   :1.0000   Max.   :1.0000
##   Phonophobia      Photophobia       Visual         Sensory
##  Min.   :0.0000   Min.   :0.00   Min.   :0.000   Min.   :0.0000
##  1st Qu.:1.0000   1st Qu.:1.00   1st Qu.:1.000   1st Qu.:0.0000
##  Median :1.0000   Median :1.00   Median :2.000   Median :0.0000
##  Mean   :0.9775   Mean   :0.98   Mean   :1.488   Mean   :0.3025
##  3rd Qu.:1.0000   3rd Qu.:1.00   3rd Qu.:2.000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.00   Max.   :4.000   Max.   :2.0000
##    Dysphasia        Dysarthria        Vertigo         Tinnitus
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.000   Min.   :0.00
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.00
##  Median :0.0000   Median :0.0000   Median :0.000   Median :0.00
##  Mean   :0.0375   Mean   :0.0025   Mean   :0.125   Mean   :0.06
##  3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.000   3rd Qu.:0.00
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.000   Max.   :1.00
##    Hypoacusis       Diplopia         Defect          Ataxia    Conscience
##  Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0   Min.   :0.0000
##  1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0   1st Qu.:0.0000
##  Median :0.000   Median :0.000   Median :0.000   Median :0   Median :0.0000
##  Mean   :0.015   Mean   :0.005   Mean   :0.015   Mean   :0   Mean   :0.0175
##  3rd Qu.:0.000   3rd Qu.:0.000   3rd Qu.:0.000   3rd Qu.:0   3rd Qu.:0.0000
##  Max.   :1.000   Max.   :1.000   Max.   :1.000   Max.   :0   Max.   :1.0000
##   Paresthesia          DPF           Type
##  Min.   :0.0000   Min.   :0.00   Length:400
##  1st Qu.:0.0000   1st Qu.:0.00   Class :character
##  Median :0.0000   Median :0.00   Mode  :character
##  Mean   :0.0075   Mean   :0.41
##  3rd Qu.:0.0000   3rd Qu.:1.00
##  Max.   :1.0000   Max.   :1.00
```

Variable Correlations The correlation plot shows that most variables have low correlation, wiht the exceptions of Intensity - Character, and Phonophobia and Photophobia.



```
##    Length     Class      Mode
##       400 character character
```

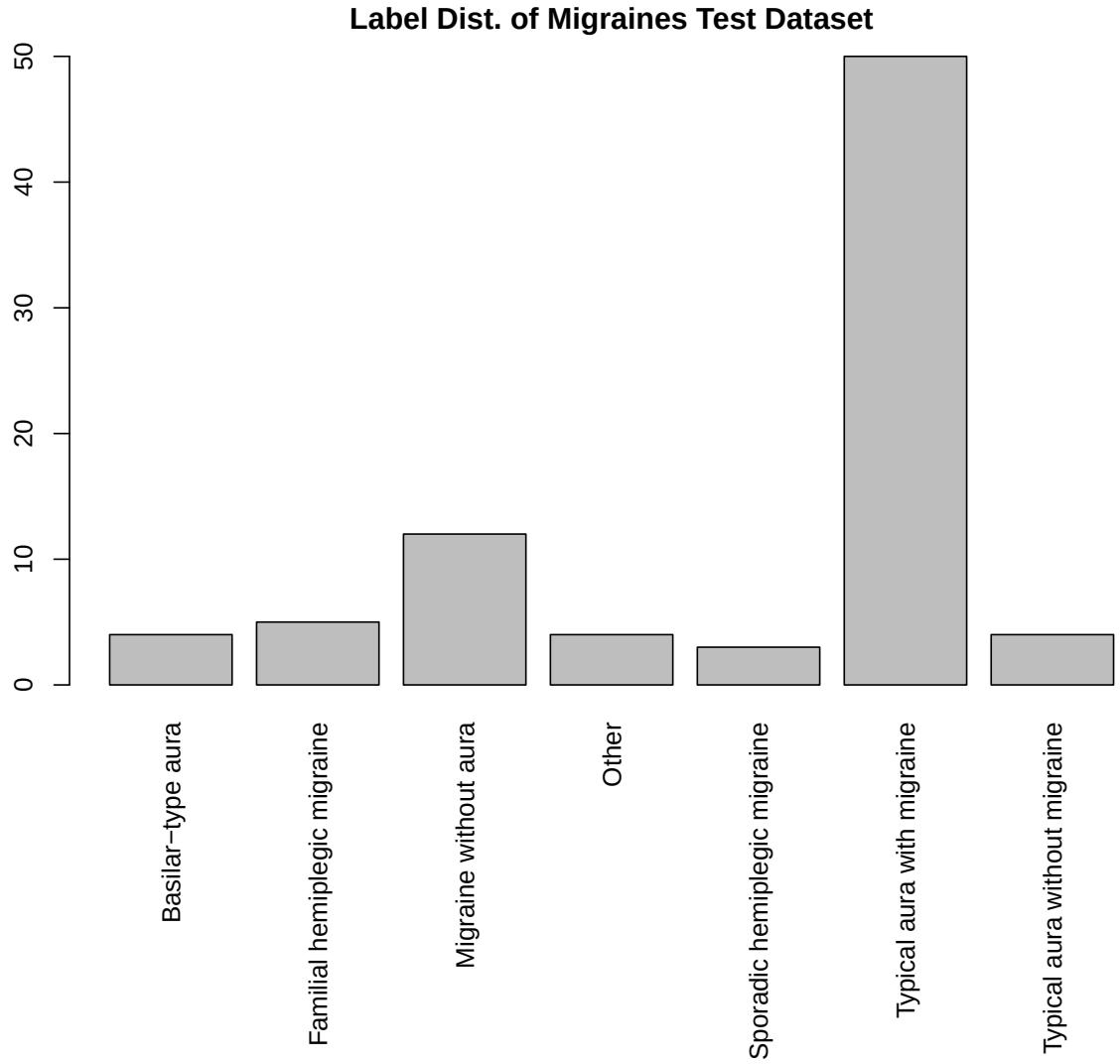


1.2 Data Partitioning and Preparation

The dataset was partitioned into an 80/20 train test split:

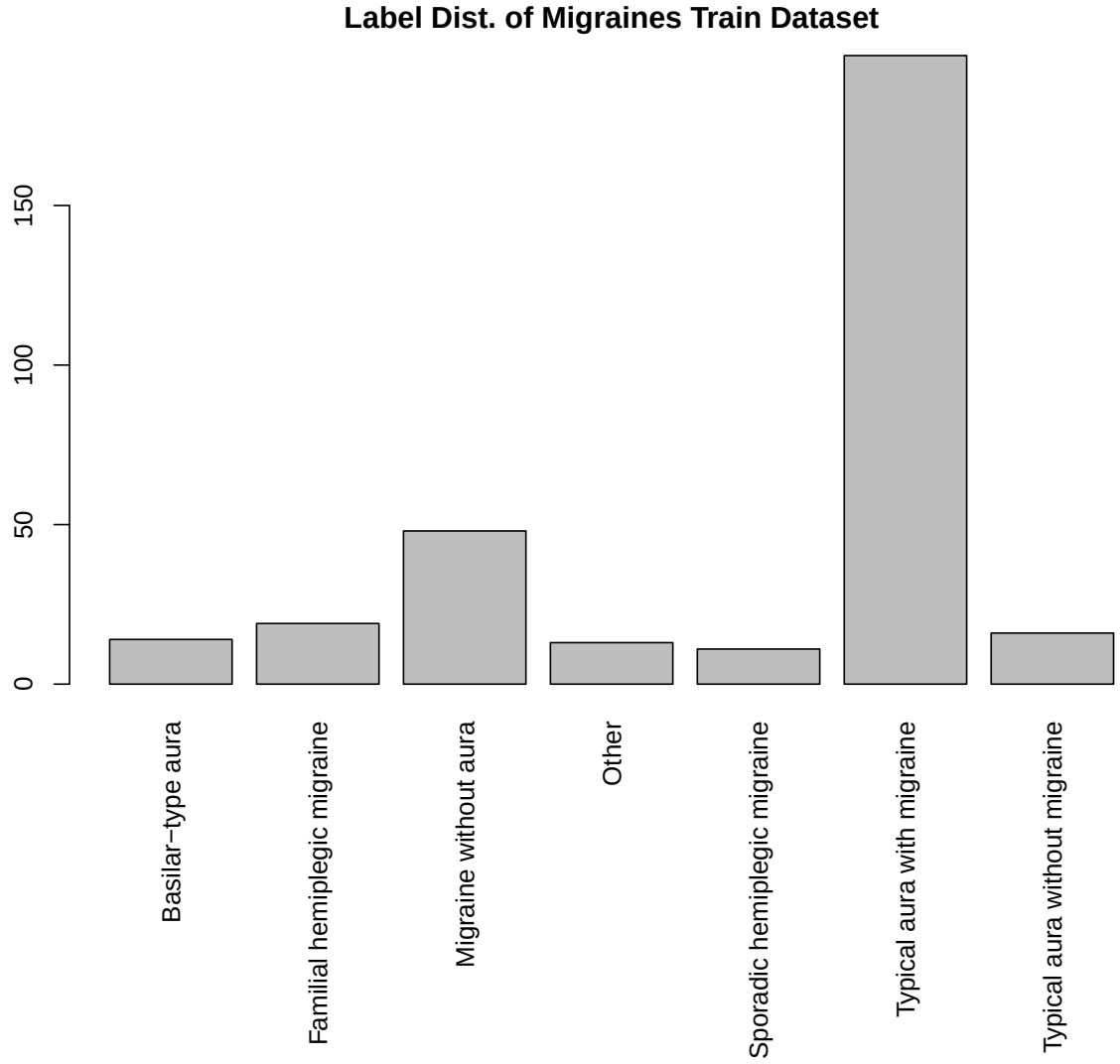
Test Data Characteristics:

```
##    Length     Class      Mode
##        82 character character
```

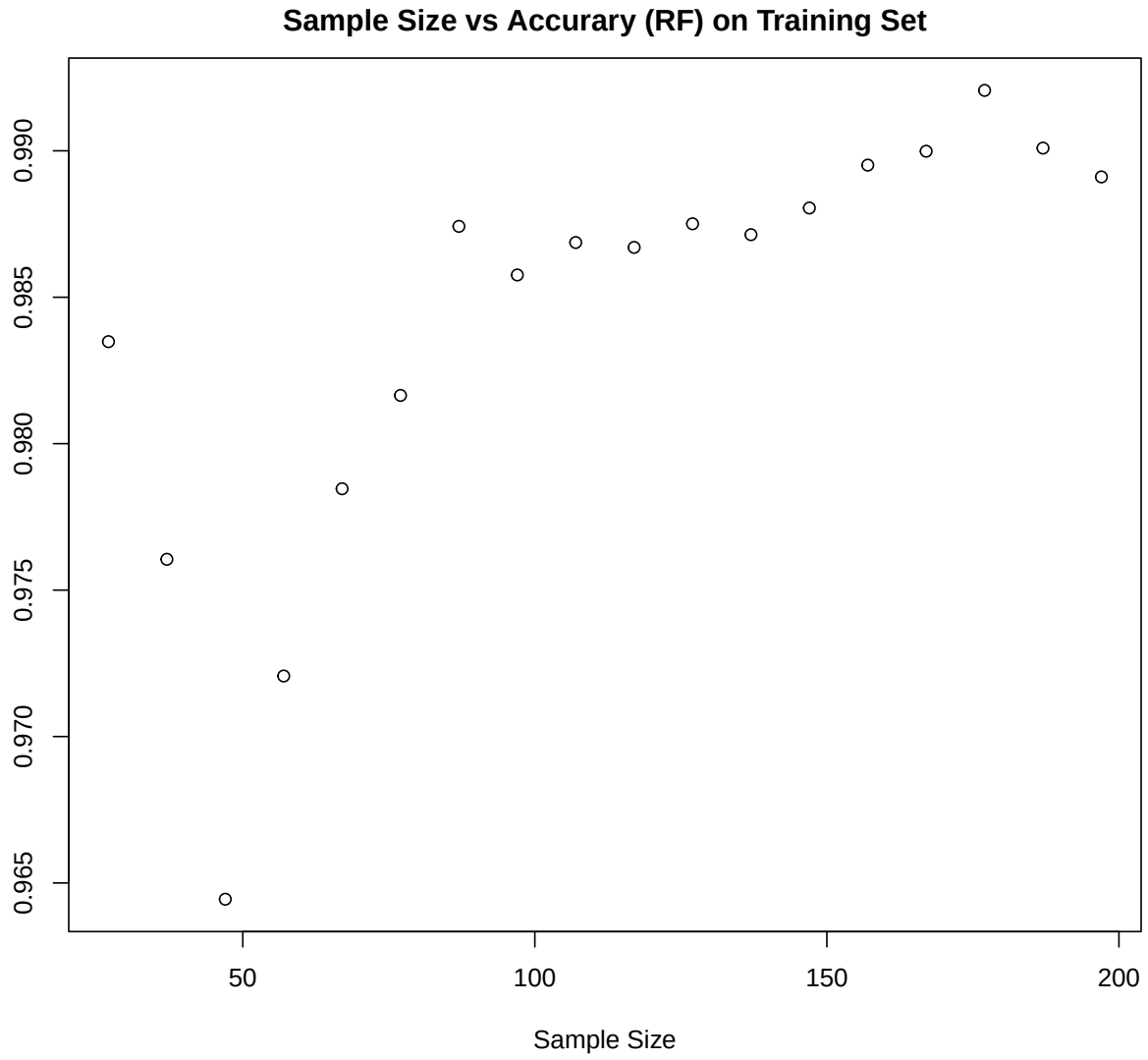


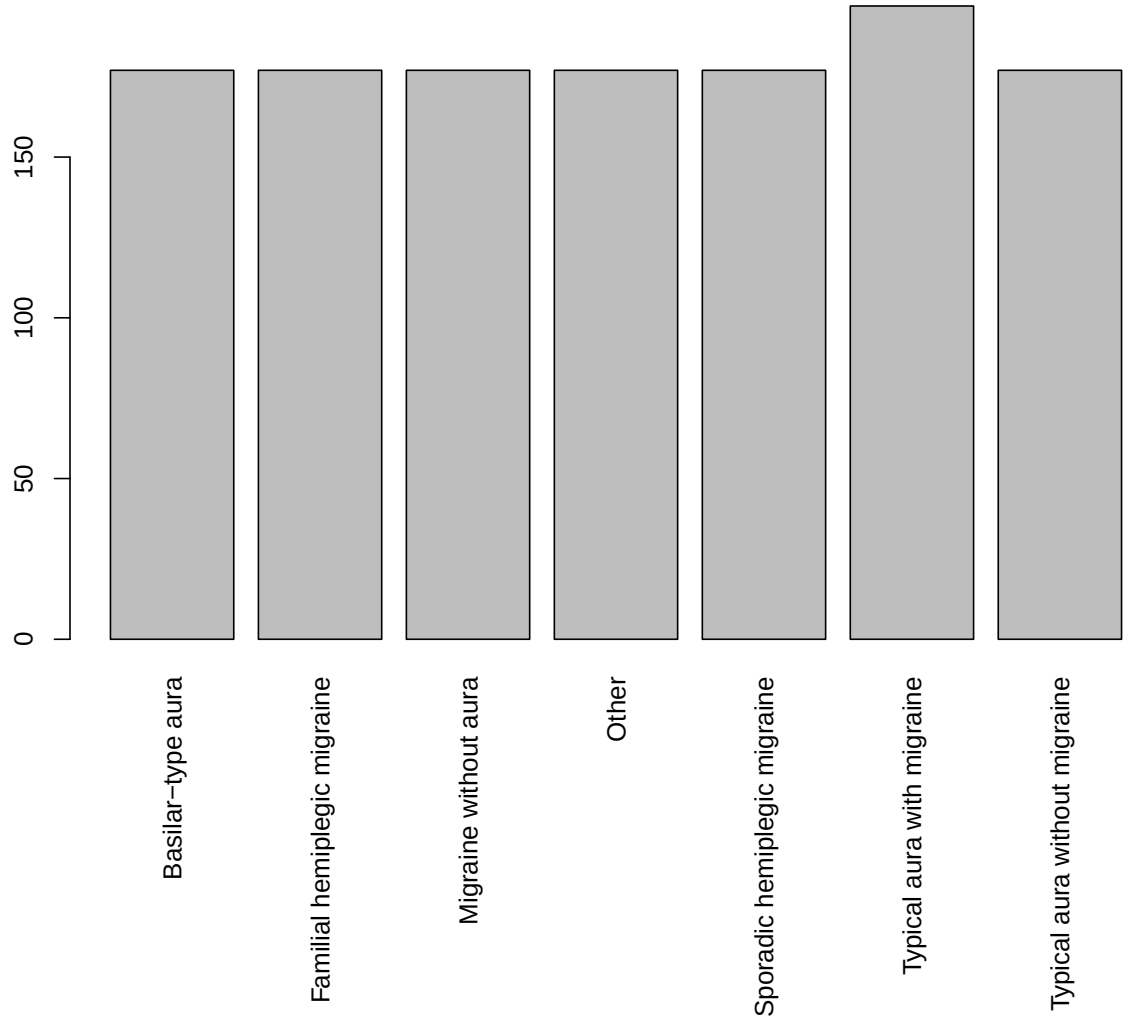
Train Dataset Characteristics:

```
##    Length     Class      Mode
##       318 character character
```



There is an imbalance in the class labels which could lead to model training issues, the training dataset will be rebalanced by resampling from the smaller label classes. A number of different sample sizes were tried and the best performance among the models was obtained by trialing the sample sizes on an RF model.



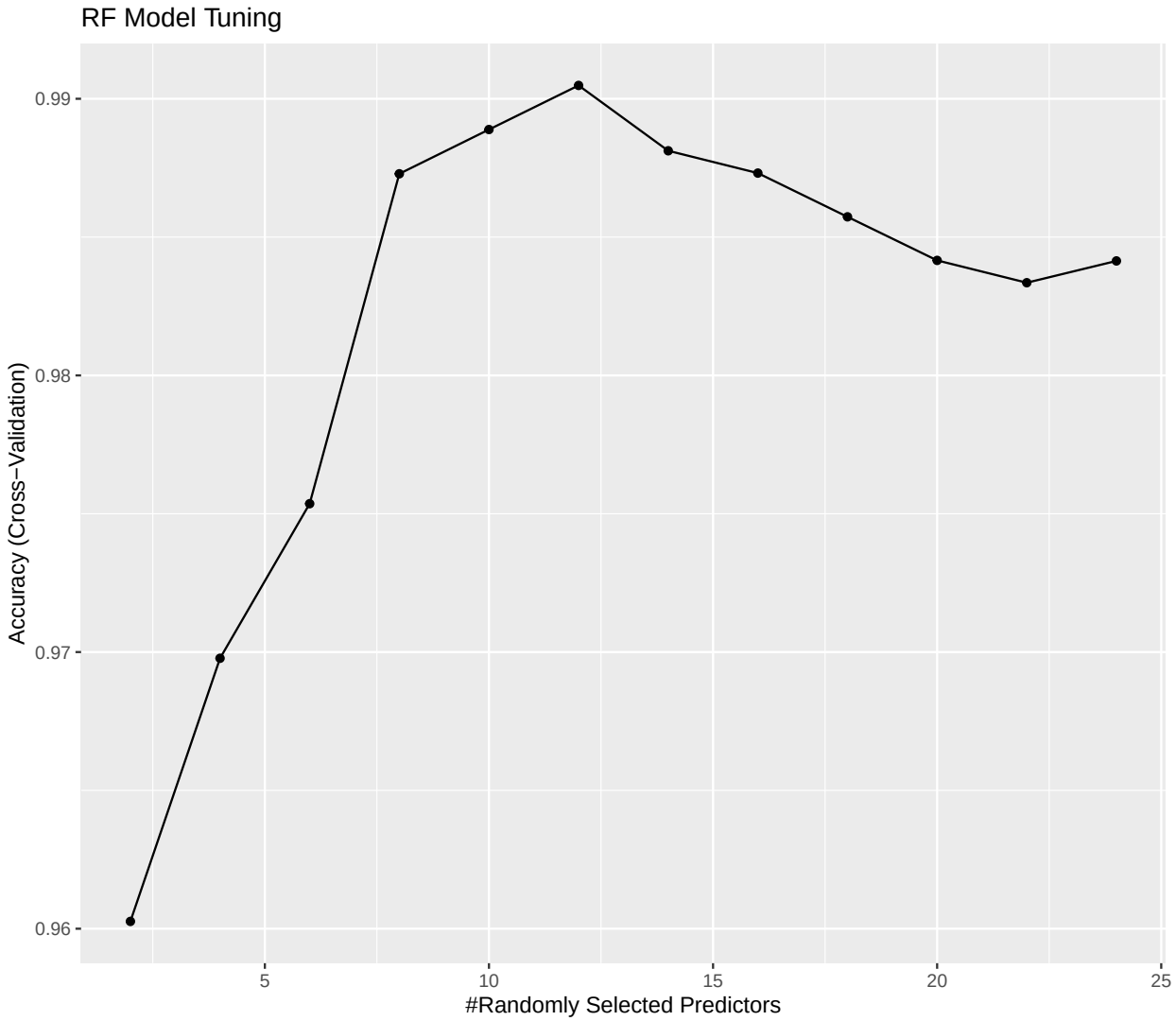

Label Distri. of Balanced Training Set
0
50
100
150
Basilar−type aura
Familial hemiplegic migraine
Migraine without aura
Other
Sporadic hemiplegic migraine
Typical aura with migraine
Typical aura without migraine

2. Modeling and Analysis

All models show below use 10 crossfold validation with a 0.9 probability. All models were separately tuned on the training set, then tested against the holdout set for F1, and overall accuracy. Detailed results are given for each model.

2.1 *Random Forest*

```
## rf variable importance
##
##   only 20 most important variables shown (out of 23)
##
##             Overall
## Age         100.000
## Visual       97.270
## Character    85.781
## Intensity    67.853
## DPF          57.385
## Location     54.822
## Vertigo      45.802
## Frequency    33.204
## Hypoacusis   24.684
## Photophobia  18.445
## Dysphasia    17.786
## Duration     16.891
## Tinnitus     16.433
## Sensory      16.022
## Phonophobia  11.605
## Conscience   10.759
## Defect        9.212
## Nausea        6.303
## Paresthesia   6.064
## Vomit         6.039
```



Test Set Accuracy and Overall F1 Score:

```
## [1] 0.9146341
```

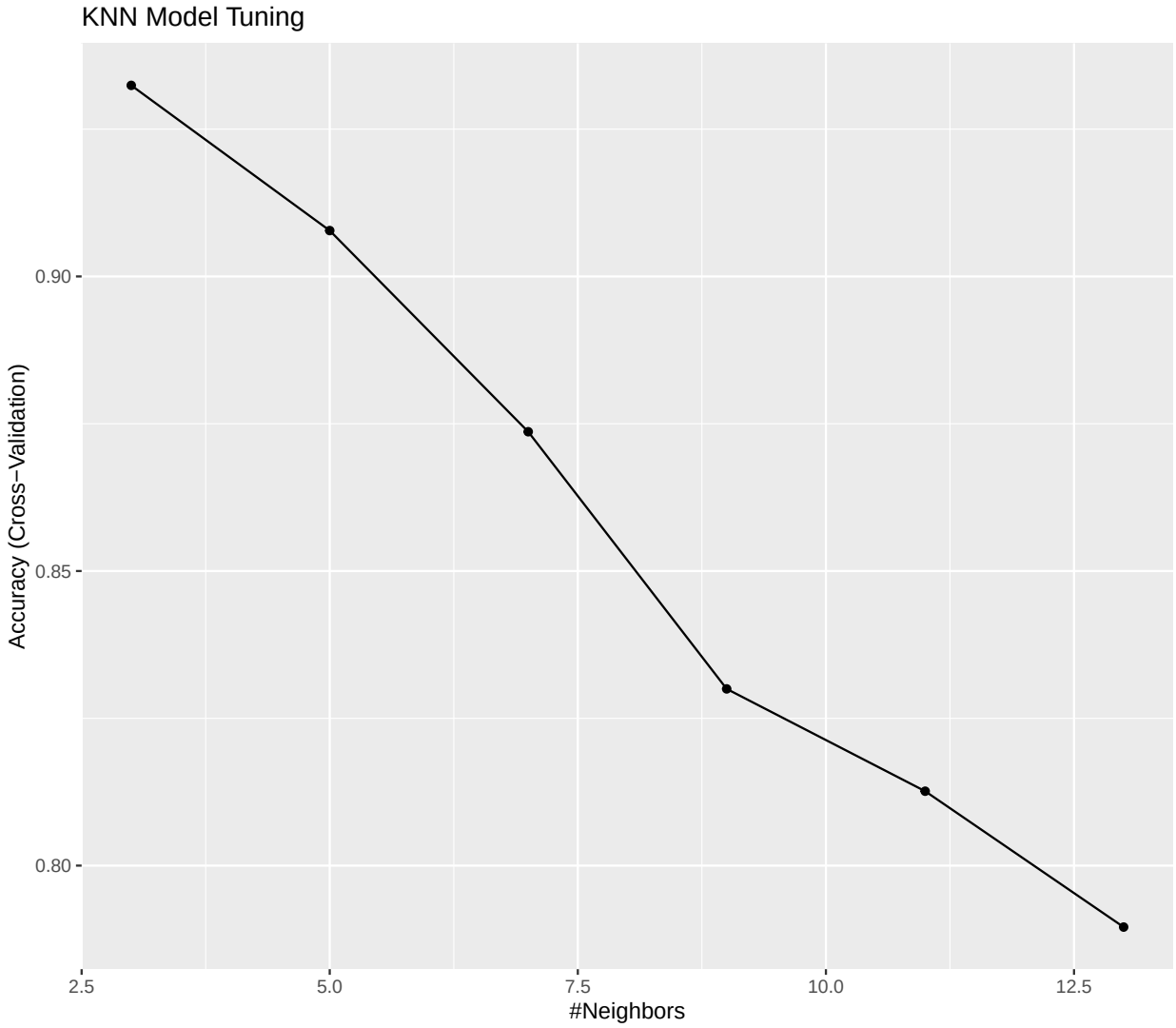
```
## [1] 0.8308905
```

Detailed Test Set Results

```
##                                     Sensitivity Specificity Pos Pred Value
## Class: Basilar-type aura              0.7500000   0.9871795      0.7500000
## Class: Familial hemiplegic migraine   1.0000000   0.9870130      0.8333333
## Class: Migraine without aura          1.0000000   0.9857143      0.9230769
## Class: Other                          0.7500000   1.0000000      1.0000000
## Class: Sporadic hemiplegic migraine   0.3333333   0.9873418      0.5000000
## Class: Typical aura with migraine     0.9400000   0.9062500      0.9400000
## Class: Typical aura without migraine  1.0000000   1.0000000      1.0000000
##                                     Neg Pred Value Precision    Recall
## Class: Basilar-type aura                 0.9871795 0.7500000 0.7500000
```

```
## Class: Familial hemiplegic migraine       1.0000000 0.8333333 1.0000000
## Class: Migraine without aura              1.0000000 0.9230769 1.0000000
## Class: Other                              0.9873418 1.0000000 0.7500000
## Class: Sporadic hemiplegic migraine       0.9750000 0.5000000 0.3333333
## Class: Typical aura with migraine         0.9062500 0.9400000 0.9400000
## Class: Typical aura without migraine      1.0000000 1.0000000 1.0000000
##                                                 F1 Prevalence Detection Rate
## Class: Basilar-type aura             0.7500000 0.04878049     0.03658537
## Class: Familial hemiplegic migraine  0.9090909 0.06097561     0.06097561
## Class: Migraine without aura         0.9600000 0.14634146     0.14634146
## Class: Other                         0.8571429 0.04878049     0.03658537
## Class: Sporadic hemiplegic migraine  0.4000000 0.03658537     0.01219512
## Class: Typical aura with migraine    0.9400000 0.60975610     0.57317073
## Class: Typical aura without migraine 1.0000000 0.04878049     0.04878049
##                                      Detection Prevalence Balanced Accuracy
## Class: Basilar-type aura                        0.04878049         0.8685897
## Class: Familial hemiplegic migraine             0.07317073         0.9935065
## Class: Migraine without aura                    0.15853659         0.9928571
## Class: Other                                    0.03658537         0.8750000
## Class: Sporadic hemiplegic migraine             0.02439024         0.6603376
## Class: Typical aura with migraine               0.60975610         0.9231250
## Class: Typical aura without migraine            0.04878049         1.0000000
```

2.2 *KNN*



Test Set Accuracy and Overall F1 Score:

```
## [1] 0.6463415
```

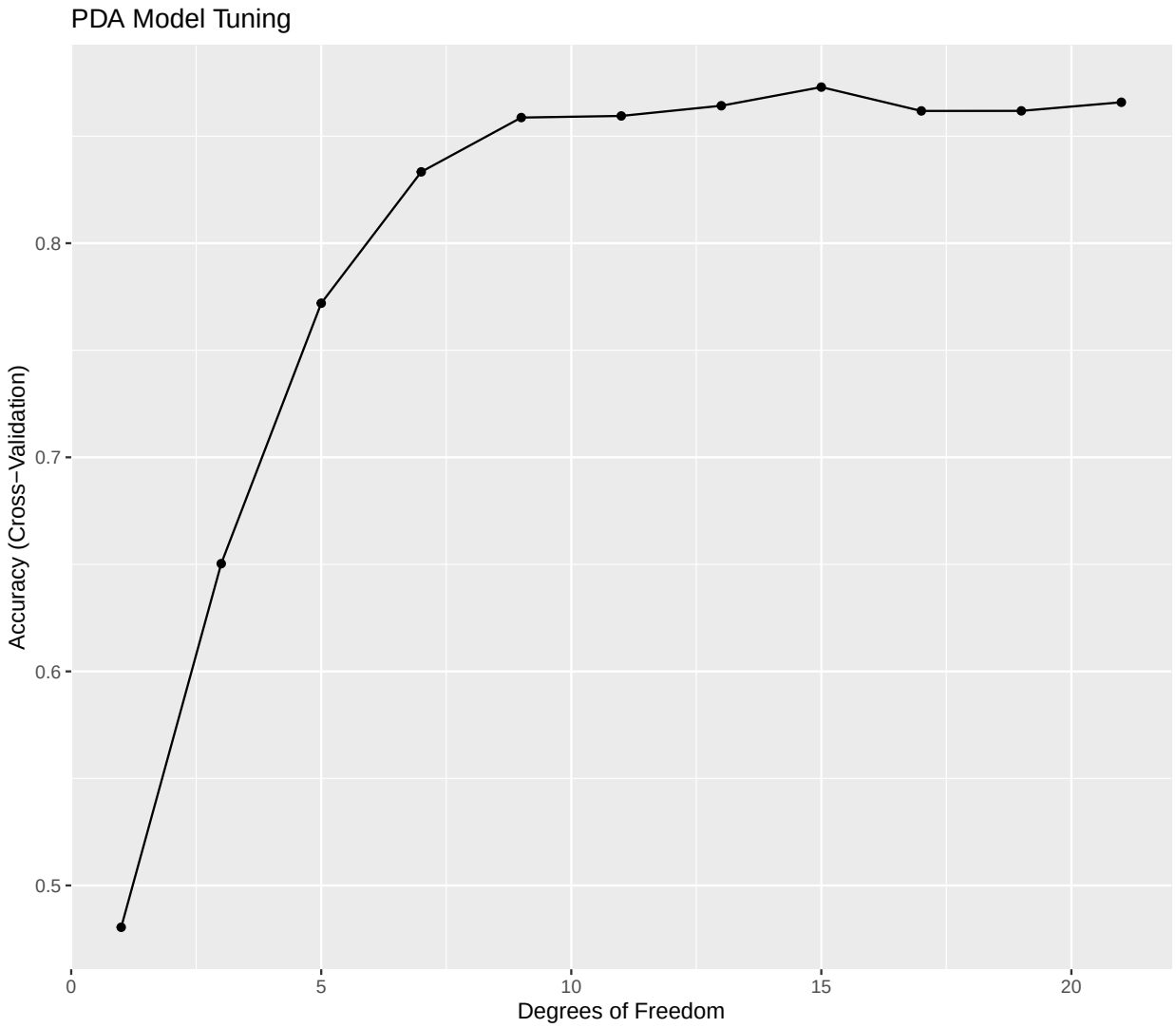
```
## [1] 0.4926888
```

Detailed Test Set Results

```
##                                    Sensitivity Specificity Pos Pred Value
## Class: Basilar-type aura             0.2500000   1.0000000      1.0000000
## Class: Familial hemiplegic migraine  0.6000000   0.9090909      0.3000000
## Class: Migraine without aura         0.5833333   0.9142857      0.5384615
## Class: Other                         0.2500000   0.9487179      0.2000000
## Class: Sporadic hemiplegic migraine  0.3333333   0.9367089      0.1666667
## Class: Typical aura with migraine    0.7400000   0.7812500      0.8409091
## Class: Typical aura without migraine 0.7500000   1.0000000      1.0000000
```

```
##                                      Neg Pred Value Precision    Recall
## Class: Basilar-type aura                  0.9629630 1.0000000 0.2500000
## Class: Familial hemiplegic migraine       0.9722222 0.3000000 0.6000000
## Class: Migraine without aura              0.9275362 0.5384615 0.5833333
## Class: Other                              0.9610390 0.2000000 0.2500000
## Class: Sporadic hemiplegic migraine       0.9736842 0.1666667 0.3333333
## Class: Typical aura with migraine         0.6578947 0.8409091 0.7400000
## Class: Typical aura without migraine      0.9873418 1.0000000 0.7500000
##                                                F1 Prevalence Detection Rate
## Class: Basilar-type aura                0.4000000 0.04878049     0.01219512
## Class: Familial hemiplegic migraine     0.4000000 0.06097561     0.03658537
## Class: Migraine without aura            0.5600000 0.14634146     0.08536585
## Class: Other                            0.2222222 0.04878049     0.01219512
## Class: Sporadic hemiplegic migraine     0.2222222 0.03658537     0.01219512
## Class: Typical aura with migraine       0.7872340 0.60975610     0.45121951
## Class: Typical aura without migraine    0.8571429 0.04878049     0.03658537
##                                      Detection Prevalence Balanced Accuracy
## Class: Basilar-type aura                        0.01219512         0.6250000
## Class: Familial hemiplegic migraine             0.12195122         0.7545455
## Class: Migraine without aura                    0.15853659         0.7488095
## Class: Other                                    0.06097561         0.5993590
## Class: Sporadic hemiplegic migraine             0.07317073         0.6350211
## Class: Typical aura with migraine               0.53658537         0.7606250
## Class: Typical aura without migraine            0.03658537         0.8750000
```

2.3 *PDA*



Test Set Accuracy and Overall F1 Score:

```
## [1] 0.8414634
```

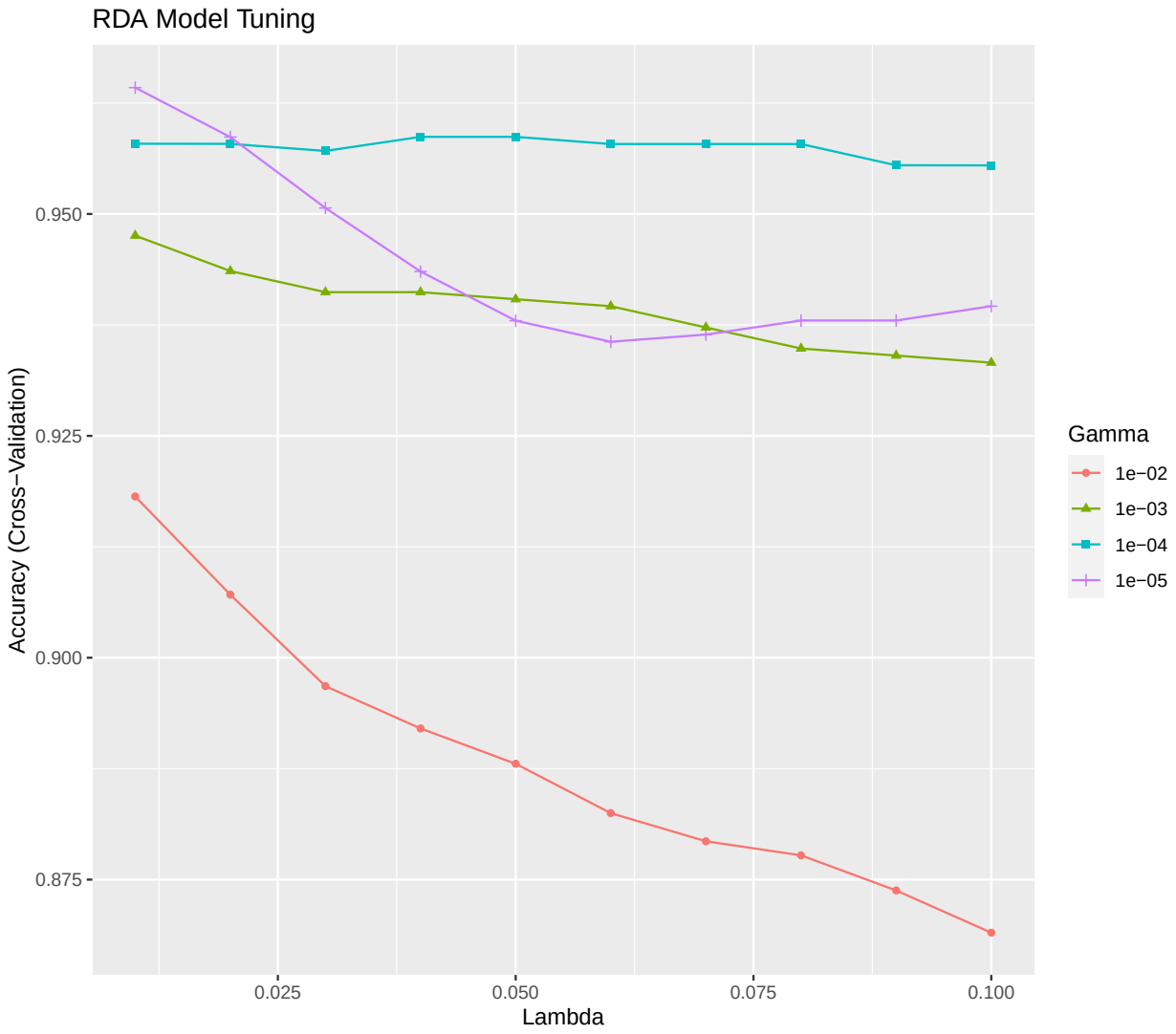
```
## [1] 0.7863106
```

Detailed Test Set Results

```
##                                     Sensitivity Specificity Pos Pred Value
## Class: Basilar-type aura              1.0000000   1.0000000      1.0000000
## Class: Familial hemiplegic migraine   0.8000000   0.9610390      0.5714286
## Class: Migraine without aura          0.9166667   0.9571429      0.7857143
## Class: Other                          0.7500000   1.0000000      1.0000000
## Class: Sporadic hemiplegic migraine   0.3333333   0.9493671      0.2000000
## Class: Typical aura with migraine     0.8400000   0.9062500      0.9333333
## Class: Typical aura without migraine  1.0000000   1.0000000      1.0000000
```

```
##                                      Neg Pred Value Precision    Recall
## Class: Basilar-type aura                  1.0000000 1.0000000 1.0000000
## Class: Familial hemiplegic migraine       0.9866667 0.5714286 0.8000000
## Class: Migraine without aura              0.9852941 0.7857143 0.9166667
## Class: Other                              0.9873418 1.0000000 0.7500000
## Class: Sporadic hemiplegic migraine       0.9740260 0.2000000 0.3333333
## Class: Typical aura with migraine         0.7837838 0.9333333 0.8400000
## Class: Typical aura without migraine      1.0000000 1.0000000 1.0000000
##                                            F1 Prevalence Detection Rate
## Class: Basilar-type aura             1.0000000 0.04878049     0.04878049
## Class: Familial hemiplegic migraine  0.6666667 0.06097561     0.04878049
## Class: Migraine without aura         0.8461538 0.14634146     0.13414634
## Class: Other                         0.8571429 0.04878049     0.03658537
## Class: Sporadic hemiplegic migraine  0.2500000 0.03658537     0.01219512
## Class: Typical aura with migraine    0.8842105 0.60975610     0.51219512
## Class: Typical aura without migraine 1.0000000 0.04878049     0.04878049
##                                      Detection Prevalence Balanced Accuracy
## Class: Basilar-type aura                        0.04878049         1.0000000
## Class: Familial hemiplegic migraine             0.08536585         0.8805195
## Class: Migraine without aura                    0.17073171         0.9369048
## Class: Other                                    0.03658537         0.8750000
## Class: Sporadic hemiplegic migraine             0.06097561         0.6413502
## Class: Typical aura with migraine               0.54878049         0.8731250
## Class: Typical aura without migraine            0.04878049         1.0000000
```

2.4 *RDA*




```
## [1] 0.8362678
```

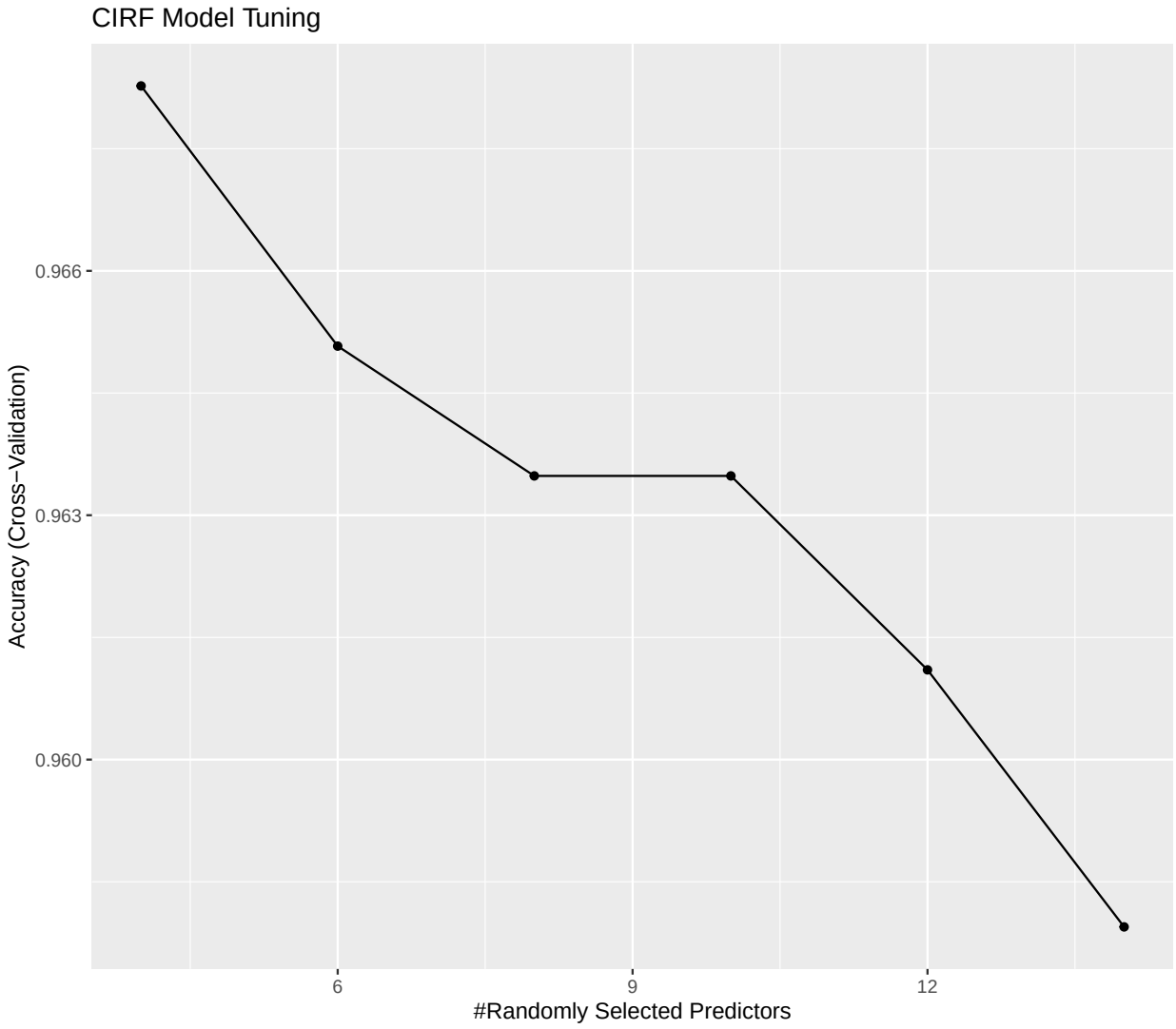
Detailed Test Set Results

```
##                                      Sensitivity Specificity Pos Pred Value
## Class: Basilar-type aura               1.0000000   1.0000000      1.0000000
## Class: Familial hemiplegic migraine    0.4000000   0.9870130      0.6666667
## Class: Migraine without aura           1.0000000   1.0000000      1.0000000
## Class: Other                           0.7500000   1.0000000      1.0000000
## Class: Sporadic hemiplegic migraine    0.6666667   0.9873418      0.6666667
## Class: Typical aura with migraine      0.9600000   0.8750000      0.9230769
## Class: Typical aura without migraine   1.0000000   0.9871795      0.8000000
##                                      Neg Pred Value Precision    Recall
## Class: Basilar-type aura                  1.0000000 1.0000000 1.0000000
## Class: Familial hemiplegic migraine       0.9620253 0.6666667 0.4000000
## Class: Migraine without aura              1.0000000 1.0000000 1.0000000
```

```
## Class: Other                                0.9873418 1.0000000 0.7500000
## Class: Sporadic hemiplegic migraine         0.9873418 0.6666667 0.6666667
## Class: Typical aura with migraine           0.9333333 0.9230769 0.9600000
## Class: Typical aura without migraine        1.0000000 0.8000000 1.0000000
##                                                   F1 Prevalence Detection Rate
## Class: Basilar-type aura                   1.0000000 0.04878049     0.04878049
## Class: Familial hemiplegic migraine        0.5000000 0.06097561     0.02439024
## Class: Migraine without aura               1.0000000 0.14634146     0.14634146
## Class: Other                               0.8571429 0.04878049     0.03658537
## Class: Sporadic hemiplegic migraine        0.6666667 0.03658537     0.02439024
## Class: Typical aura with migraine          0.9411765 0.60975610     0.58536585
## Class: Typical aura without migraine       0.8888889 0.04878049     0.04878049
##                                           Detection Prevalence Balanced Accuracy
## Class: Basilar-type aura                                0.04878049         1.0000000
## Class: Familial hemiplegic migraine                     0.03658537         0.6935065
## Class: Migraine without aura                            0.14634146         1.0000000
## Class: Other                                            0.03658537         0.8750000
## Class: Sporadic hemiplegic migraine                     0.03658537         0.8270042
## Class: Typical aura with migraine                       0.63414634         0.9175000
## Class: Typical aura without migraine                    0.06097561         0.9935897
```

2.5 *CIRF*

```
## cforest variable importance
##
##   only 20 most important variables shown (out of 23)
##
##             Overall
## Visual      100.000
## Intensity    76.856
## Character    64.821
## DPF          64.667
## Location     61.198
## Vertigo      54.747
## Age          42.046
## Frequency    29.870
## Dysphasia    28.733
## Photophobia  22.651
## Phonophobia  20.121
## Tinnitus     20.035
## Hypoacusis   18.064
## Defect       17.070
## Sensory      16.332
## Duration     15.413
## Vomit        13.182
## Conscience   10.814
## Paresthesia   7.322
## Nausea        7.070
```



Test Set Accuracy and Overall F1 Score:

```
## [1] 0.902439
```

```
## [1] 0.8400878
```

Detailed Test Set Results

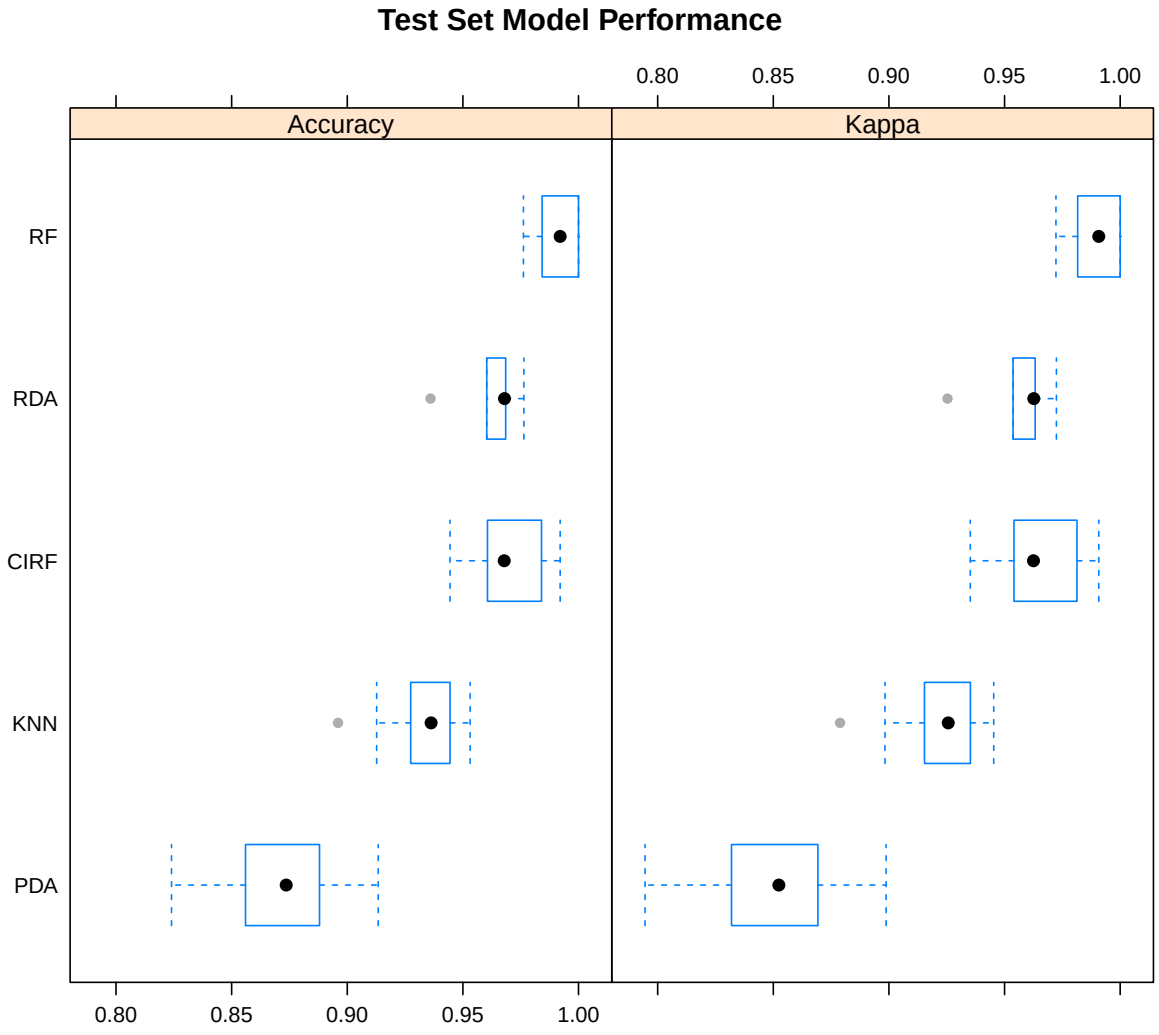
```
##                                     Sensitivity Specificity Pos Pred Value
## Class: Basilar-type aura              1.0000000   0.9871795      0.8000000
## Class: Familial hemiplegic migraine   1.0000000   0.9740260      0.7142857
## Class: Migraine without aura          1.0000000   0.9571429      0.8000000
## Class: Other                          1.0000000   1.0000000      1.0000000
## Class: Sporadic hemiplegic migraine   0.3333333   0.9746835      0.3333333
## Class: Typical aura with migraine     0.8800000   1.0000000      1.0000000
## Class: Typical aura without migraine  1.0000000   1.0000000      1.0000000
##                                     Neg Pred Value Precision    Recall
## Class: Basilar-type aura                 1.0000000 0.8000000 1.0000000
```

```
## Class: Familial hemiplegic migraine       1.0000000 0.7142857 1.0000000
## Class: Migraine without aura              1.0000000 0.8000000 1.0000000
## Class: Other                              1.0000000 1.0000000 1.0000000
## Class: Sporadic hemiplegic migraine       0.9746835 0.3333333 0.3333333
## Class: Typical aura with migraine         0.8421053 1.0000000 0.8800000
## Class: Typical aura without migraine      1.0000000 1.0000000 1.0000000
##                                                 F1 Prevalence Detection Rate
## Class: Basilar-type aura             0.8888889 0.04878049     0.04878049
## Class: Familial hemiplegic migraine  0.8333333 0.06097561     0.06097561
## Class: Migraine without aura         0.8888889 0.14634146     0.14634146
## Class: Other                         1.0000000 0.04878049     0.04878049
## Class: Sporadic hemiplegic migraine  0.3333333 0.03658537     0.01219512
## Class: Typical aura with migraine    0.9361702 0.60975610     0.53658537
## Class: Typical aura without migraine 1.0000000 0.04878049     0.04878049
##                                      Detection Prevalence Balanced Accuracy
## Class: Basilar-type aura                        0.06097561         0.9935897
## Class: Familial hemiplegic migraine             0.08536585         0.9870130
## Class: Migraine without aura                    0.18292683         0.9785714
## Class: Other                                    0.04878049         1.0000000
## Class: Sporadic hemiplegic migraine             0.03658537         0.6540084
## Class: Typical aura with migraine               0.53658537         0.9400000
## Class: Typical aura without migraine            0.04878049         1.0000000
```

3. Results

After fitting several different classification models, and tuning the parameters of each model. The best accuracies of each model are summarized below. It is a virtual tie between RDA and RF. RF generally was able to attain a higher overall accuracy with RDA attaining a higher average F1 score.

The RF and CIRF models each had different variable importance rankings. This may help to explain the difference between the model results.




```
##       Accuracy  Macro_F1
## RF   0.9146341 0.8308905
## KNN  0.6463415 0.4926888
## PDA  0.8414634 0.7863106
## RDA  0.9146341 0.8362678
## CIRF 0.9024390 0.8400878
```

4. Conclusion

This report detailed the fitting and results of several classification models for a migraine type classification dataset. The dataset was neither large, or balanced which presented a challenge to fit an accurate model which achieved both a high accuracy and high F1 class scores on an unseen test set. There were only a small number of some of the labels in the test, which means one or two misclassifications can have a large effect on the report F1 score. To help with training the dataset was resampled to help balance the classes prior to fitting the final models. To help mitigate potential overfitting cross validation was used for all models during training.

It is not known what the relative risks are for false positives and false negatives for each of the classes in terms of patient outcome so I am not able to draw a conclusion on which model is 'best'. The RF model achieved the highest overall accuracy, but the F1 scores were more balanced, and the average F1 score was higher for the RDA model. Comparing the F1 scores for each of the classes it can be seen that each of these two models attains a higher F1 score for different classes. An ensemble model could be built using the predicted probabilities, or other criteria from each of the models to make a final classification prediction.