Classifying Different Types of Migraines

Robert Thompson
College of Computing and Informatics
Drexel University
Philadelphia, Pennsylvania 19014
March 19, 2023

***Abstract*—This paper focuses on migraine classification prediction using the Migraine data set developed by the Centro Materno Infantil de Soledad medical professionals in 2013 and hosted on Kaggle. The data set contains data from patients who have experienced migraines. My goal is to understand the different variables involved in classifying different types of migraines. In this study, I have used three machine-learning classification models developed by Pyspark: logistic regression, decision trees, and random forest. Of the three models, the logistic regression classifier performed the best with overall 88% accuracy when validated using the testing data.**

***Index Terms*—Migraines, Classification, Random Forest Classifier, Decision Tree Classifier, Logistic Regression Classifier, Pyspark, Python, Health**

I. Introduction

According to the American Migraine Foundation, "[a] migraine is not just a bad headache. It's a disabling neurological disease 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 migraines" [9]. Based on this estimate, roughly 11.7% of the United States estimated population of 334 million [11] in July 2022 live with migraines. Although the American Migraine Foundation estimated 39 million Americans live with migraines, this number could be drastically increased as many Americans do not realize they live with them.

A problem with migraines is that it is not clear what they are caused by but a combination of genetics and a person's environment plays a role. People who suffer from migraines may have overlapping symptoms that align with different types of headaches. Symptoms of migraines include nausea, increased sensitivity to light, sound, or smells, dizziness, and extreme fatigue [10]. Furthermore, "there's no blood test or scan (ultrasound, CT scan or MRI) that will tell your doctor if your head is in pain. The only real way for your doctor to know is to talk to you" [9]. Therefore, frequent communication between patients and their doctors must occur to accurately diagnose a migraine. The purpose of this study will not be to predict when a migraine may occur but to classify types of migraine based on the patient's data.

II. Data Set

This paper utilizes the Migraine data set developed by the Centro Materno Infantil de Soledad medical professionals in 2013 and hosted on Kaggle. The data set contains numeric and structured data spanning twenty-four variables. The sample size includes four hundred patients with various diagnoses of migraines. Fig. 1 shows the list of variables, their description, and the range of values. The data set contains twenty-three feature variables and one prediction variable. The prediction variable is the *Type* column that includes seven different migraine diagnoses all of which are actual values based on the twenty-three feature variables.

III. Data Analysis

Now that the variables of the data set are understood, we must begin our data analysis to understand the results of the data set we will use to predict different types of migraines. To begin the data analysis, I first focused on understanding the sample size, the variables (or column data), and their range. Once that was understood, choosing the important data to visualize was the easy part.

Variable	Description
Age	Patient's age
Duration	Duration of Symptoms in the last migraine episode in days: N/A
Frequency	Frequency of migraines per month: N/A
Location	Unilateral or bilateral pain: (None - 0, Unilateral - 1, Bilateral - 2)
Character	Throbbing or constant pain: (None - 0, Throbbing - 1, Constant - 2)
Intensity	Pain intensity: (None - 0, Mild - 1, Medium - 2, Severe - 3)
Nausea	Nauseous feeling: (No - 0, Yes - 1)
Vomit	Vomiting: (No - 0, Yes - 1)
Phonophobia	Noise sensitivity: (No - 0, Yes - 1)
Photophobia	Light sensitivity: (No - 0, Yes - 1)
Visual	Number of reversible visual symptoms: N/A
Sensory	Number of reversible sensory symptoms: N/A
Dysphasia	Lack of speech coordination: (No - 0, Yes - 1)
Dysarthria	Disarticulated sounds and words: (No - 0, Yes - 1)
Vertigo	Dizziness: (No - 0, Yes - 1)
Tinnitus	Ringing in the ears: (No - 0, Yes - 1)
Hypoacusis	Hearing loss: (No - 0, Yes - 1)
Diplopia	Double vision: (No - 0, Yes - 1)
Visual defect	Simultaneous frontal eye field and nasal field defect and in both eyes: (No - 0, Yes - 1)
Ataxia	Lack of muscle control: (No - 0, Yes - 1)
Conscience	Jeopardized conscience: (No - 0, Yes - 1)
Paresthesia	Simultaneous bilateral paresthesia: (No - 0, Yes - 1)
DPF	Family background: (No - 0, Yes - 1)
Type	Sporadic Hemiplegic Migraine, Basilar-Type Aura, Familial Hemiplegic Migraine, Typical Aura with Migraine, Typical Aura without Migraine, Migraine without Aura, Other

Fig. 1. Migraine Data Set Variables and Description

A. Types of Migraines

As seen in Fig. 2, 62% of the sample size has been categorized as *Typical aura with migraine*. The second highest percentage of the sample size is 15% which has been categorized as *Migraine without aura*. After the two types, the following 23% of the sample size is split across the migraine types from most to least percentage of the sample size: *Familial hemiplegic migraine*, *Typical aura without migraine*, *Basilar-type aura*, *Other*, and *Sporadic hemiplegic migraine*.

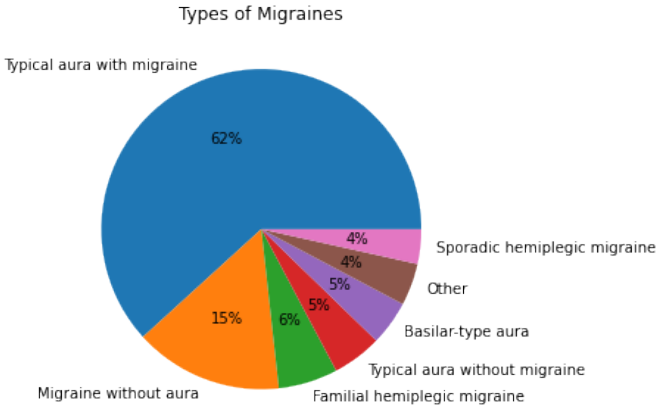

Fig. 2. Types of Migraines

B. Age Histogram

Fig. 3 shows the age distribution of the data set. The age range of the data set is between fifteen and seventy-seven with the average age of patients being thirty-two (precisely 31.705). The highest distribution begins in the early teens to late twenties where the distribution size starts at a count of fifty-five and reaches a max sample size of eighty. The sample size of patients from the data set then seems to reduce as the patient's age increases.

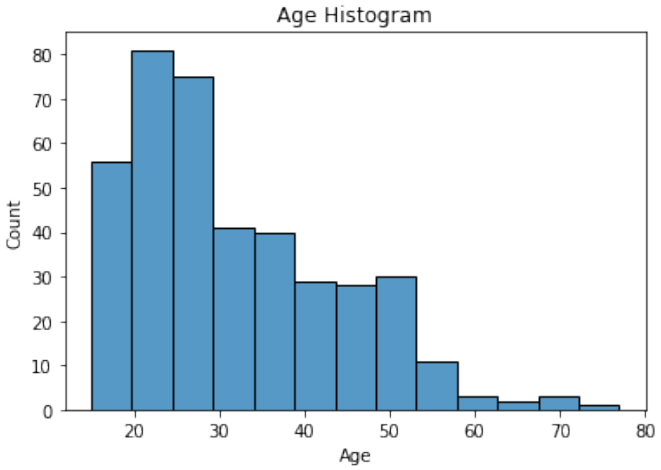

Fig. 3. Age Histogram

C. Duration of Migraines

The Fig. 4 bar chart shows the duration of migraine symptoms based on the days they persist. The data ranges between one day and three days

of migraine symptoms with the average duration coming in at 1.61 days. The data set generally consists of patients who have migraines subside within one day after receiving their first migraine symptoms.

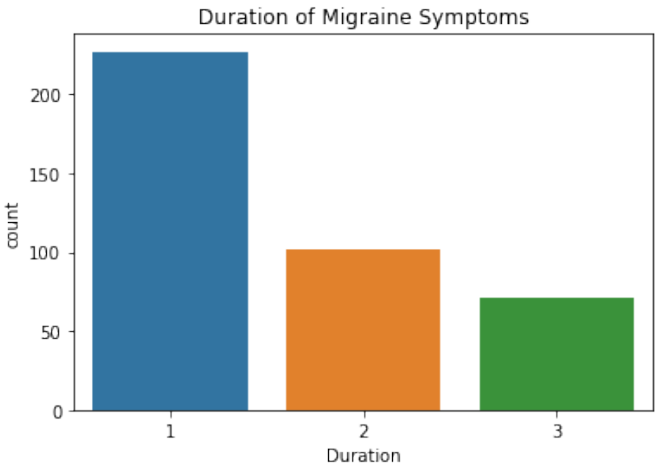

Fig. 4. Duration of Migraine Symptoms in Days

D. Frequency of Migraines

Fig. 5 shows the frequency of migraines that patients have experienced per month. The frequency of migraines ranges from one to eight migraines per month with most patients experiencing at most one migraine per month. The average migraines per month based on the sample size is 2.365 migraines per month.

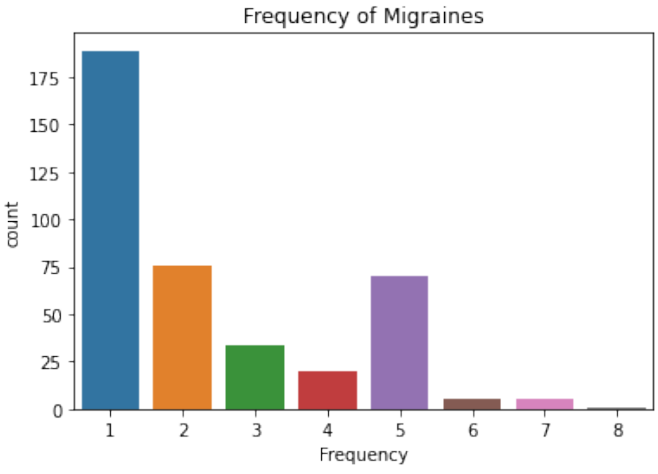

Fig. 5. Frequency of Migraines Per Month

IV. Methodology

A. Data Pre-Processing

The initial approach to handle the pre-processing of the data initially began during the *Data Analysis* phase. When understanding the different variables and their range of values, it was discovered that one of the variables had only one unique value. This was the *Ataxia* variable and because of their only one unique result, the decision was to remove the column entirely from the data set. After removing this variable, the next step was to remove any NULL or missing data. The remaining twenty-two feature variables were chosen to be used to predict the different types of migraines.

The next step was to leverage Pyspark's String Indexer function to convert the different migraine types (or *target* values) that were string values to integer values to allow them to be successfully passed to the classification models. After that, I leveraged another Pyspark function to create a feature vector that took all the feature data and merged the data into a singular vector.

The final step to be taken before providing the data to the Pyspark models was to split our data set into training and testing data. In order to avoid overfitting in the models, splitting the data as 80% (or 324 samples) training and 20% (or 76 samples) testing data were performed.

After splitting the data into train and test, the models described below were created. Finally, the target values, feature vector, and models were passed to a Pyspark transformation pipeline. The pipeline was then trained on the training data, validated against both the training and testing data, and then evaluated for their performance.

B. Models

In order to perform the classification of different types of migraines, the following multi-class classification models were leveraged from Pyspark: Logistic Regression, Decision Tree, and Random Forest. Logistic regression is generally used for binary classification (ex. Yes/no) but in this case, it is used for multi-class classification of the different types of migraines. The model uses the logistic function to obtain the probability of the different migraine types occurring based on the trained model. The Decision

Tree model was chosen for its incremental approach and the ability to actually visualize the tree as it is broken down into its individual nodes and leaves. The Random Forest classifier is essentially a 1-N *forest* of different decision trees. The classifier is built on bootstrapped data from 1-N trees and the final label is chosen based on the results of all the decision trees.

V. Evaluation

To verify the performance of each of the models by classifying a migraine type, I used accuracy, precision, recall, and f1-score to determine the overall performance. Since I used Pyspark for these classifier models, I was able to run the data through each of these models based on the Pyspark default hyper-parameters and also tuned them to increase the accuracy. Fig. 6 shows each of the models and their default hyper-parameters as well as the parameters that were tuned to increase performance.

Model	Parameters
Random Forest	Impurity = entropy, maxDepth = 5 (default), numTrees = 20 (default)
Random Forest – Tuned	Impurity = entropy, maxDepth = 10, numTrees = 25
Decision Tree	Impurity = entropy, maxDepth = 5 (default)
Decision Tree – Tuned	Impurity = entropy, maxDepth = 10
Logistic Regression	maxIter = 100 (default)
Logistic Regression – Tuned	maxIter = 1000

Fig. 6. Machine Learning Models and Parameters

A. Train Data

Fig. 7 shows the results of each of the model's performances using the training data for validation. As expected, each of the models performed well with the least accurate model being the Decision Tree classifier with default hyper-parameters with an accuracy of 89%. The Decision Tree classifier that was tuned performed the best with an accuracy of 99%.

Model	Accuracy	Recall	Precision	F1
Random Forest	0.92	0.962	0.965	0.981
Random Forest – Tuned	0.981	0.981	0.985	0.99
Decision Tree	0.898	0.962	0.943	0.981
Decision Tree – Tuned	0.994	1	0.995	1
Logistic Regression	0.978	1	0.982	1
Logistic Regression – Tuned	0.988	1	0.989	1

Fig. 7. Classifier Performance using the Training Data

B. Test Data

Fig. 8 shows the results of each of the model's performances using the testing data for validation. The testing data achieved far fewer results than the training data with the least accurate model of the Decision Tree classifier achieving a 77% accuracy for predicting a migraine type. Unlike the results of the training data, the most accurate model was the Logistic Regression classifier with default hyper-parameters that achieved an 88% accuracy for predicting a migraine type.

Model	Accuracy	Recall	Precision	F1
Random Forest	0.816	0.818	0.971	0.9
Random Forest – Tuned	0.868	0.9	0.933	0.947
Decision Tree	0.776	0.8	0.902	0.842
Decision Tree – Tuned	0.803	0.889	0.823	0.889
Logistic Regression	0.882	1	0.9	0.941
Logistic Regression – Tuned	0.829	1	0.845	0.941

Fig. 8. Classifier Performance using the Testing Data

VI. Conclusion

When hyper-tuning the parameters, the models performed with over 80% accuracy using the twenty-two features. Amongst the classifiers, the logistic regression classifier finished with an 88% accuracy, the random forest classifier finished with an 86% accuracy, and the decision tree classifier finished with an 80% accuracy. In the future, obtaining a data set with more than four-hundred rows of data

would be imperative to effectively train each model and perform predictions on that data set.

VII. Future Work

As someone who has experienced first-hand the struggles that migraines have in ordinary everyday life, it has become very important to learn as much as possible about them. There are many different variables that play a key role in causing even the slightest headache and this same approach could be used when predicting migraines. Although this data set was used to classify different diagnoses of migraines, I would like to explore in the future data sets that would enable the use of different machine-learning algorithms to effectively predict when and if a migraine would occur.

References

1 P. A. Sanchez-Sanchez, J. R. García-González, and J. M. Rúa Ascar, "Automatic migraine classification using artificial neural networks," F1000Research, vol. 9, no. 2, p. 618, Jul. 2020, doi: https://doi.org/10.12688/f1000research.23181.2.

2 A. Gago-Veiga et al., "To what extent are patients with migraine able to predict attacks?," Journal of Pain Research, vol. Volume 11, no. 11, pp. 2083–2094, Sep. 2018, doi: https://doi.org/10.2147/jpr.s175602.

3 M. Banoula, "Classification in Machine Learning — The Best Classification Models," Simplilearn.com, Feb. 13, 2023. https://www.simplilearn.com/tutorials/machine-learning-tutorial/classification-in-machine-learning

4 J. Brownlee, "4 Types of Classification Tasks in Machine Learning," Machine Learning Mastery, Apr. 07, 2020. https://machinelearningmastery.com/types-of-classification-in-machine-learning/

5 M. Clinic, "Migraine - Symptoms and causes," Mayo Clinic, Jul. 02, 2021. https://www.mayoclinic.org/diseases-conditions/migraine-headache/symptoms-causes/syc-20360201#:~:text=A%20migraine%20is%20a%20headache

6 A. Spark, "LogisticRegression — PySpark 3.1.3 documentation," spark.apache.org. https://spark.apache.org/docs/3.1.3/api/python/reference/api/pyspark.ml.classification.LogisticRegression.html#pyspark.ml.classification.LogisticRegression (accessed Mar. 17, 2023).

7 A. Spark, "DecisionTreeClassifier — PySpark 3.1.3 documentation," spark.apache.org. https://spark.apache.org/docs/3.1.3/api/python/reference/api/pyspark.ml.classification.DecisionTreeClassifier.html#pyspark.ml.classification.DecisionTreeClassifier (accessed Mar. 17, 2023).

8 A. Spark, "RandomForestClassifier — PySpark 3.1.3 documentation," Apache.org, 2023. https://spark.apache.org/docs/3.1.3/api/python/reference/api/pyspark.ml.classification.RandomForestClassifier.html#pyspark.ml.classification.RandomForestClassifier (accessed Mar. 17, 2023).

9 American Migraine Foundation, "What is Migraine? — American Migraine Foundation," American Migraine Foundation, Jan. 21, 2021. https://americanmigrainefoundation.org/resource-library/what-is-migraine/

10 P. Medicine, "Migraine vs. Headache: How to Tell the Difference – Penn Medicine," www.pennmedicine.org, Feb. 31, 2022. https://www.pennmedicine.org/updates/blogs/health-and-wellness/2019/november/migraines-vs-headaches

11 United States Census Bureau, "U.S. Census Bureau QuickFacts: United States," www.census.gov, Jul. 01, 2022. https://www.census.gov/quickfacts/fact/table/US/PST045222