MIGRAINE SYMPTOM CLASSIFICATION

Advanced Data Analysis

Prepared by: Group 01

Supul Wicramasinghe –s16177

Thenuka Yatawara – s16383

Vayani Kavindya - s16322

Sanduni Fonseka – s16026

ABSTRACT

This project conducts advanced data analysis on a migraine symptom classification dataset to develop predictive insights and uncover complex patterns among clinical and demographic variables. Building on prior exploratory analysis, we apply a range of supervised learning techniques to classify migraine types based on symptom profiles, pain characteristics, and patient information.

To address challenges such as class imbalance and feature redundancy, we implement resampling methods and model optimization strategies. Techniques such as cross-validation, performance evaluation (e.g., confusion matrix, precision, recall), and comparative model assessment are employed to ensure robust classification outcomes.

The results reveal distinct symptom patterns linked to specific migraine categories and demonstrate the effectiveness of data-driven models in capturing diagnostic features. These insights provide a foundation for supporting clinical decision-making and advancing personalized approaches to migraine diagnosis and treatment.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

Migraines are a complex neurological disorder characterized by recurrent headaches that typically cause severe throbbing or pulsing pain, often on one side of the head. These headaches are frequently accompanied by symptoms such as nausea, vomiting, and heightened sensitivity to light and sound. Some migraine sufferers experience an aura—a set of sensory disturbances including visual flashes, blind spots, tingling sensations, or speech difficulties—that can precede or accompany the headache. Migraines can last from several hours to days and can significantly disrupt daily life and activities.

Understanding the variations in migraine symptoms and their patterns is clinically important because it aids in accurate diagnosis and effective treatment planning. Different types of migraines manifest with distinct symptom profiles, and a detailed examination of these symptoms can help healthcare professionals tailor management strategies to individual patients.

This project presents an advanced data analysis of a migraine symptom classification dataset, aimed at uncovering meaningful patterns and relationships among demographic, clinical, and symptom-based variables. As the exploratory data analysis (EDA) has already been completed, this report focuses on applying advanced machine learning techniques—such as Support Vector Machines (SVM), Random Forests (RF), Decision Trees (DT), K-Nearest Neighbors (KNN), and XG Boost to classify migraine types based on symptom data. The objective is to assess the performance of these models and identify important features that contribute to accurate classification, thereby supporting improved clinical understanding and decision-making.

DESCRIPTION OF QUESTION

Migraine is a common yet often underdiagnosed neurological disorder due to its wide range of overlapping symptoms, including variations in intensity, location, duration, and sensory disturbances. Many individuals may not recognize their symptoms as migraines, leading to delayed or missed diagnoses. This not only affects timely treatment but also impacts quality of life.

To address this, our project seeks to answer the following key question:
Can machine learning techniques effectively classify migraine types based on symptom patterns, and thereby support earlier identification and improved clinical understanding of migraines?

To explore this, we use a migraine symptom dataset containing clinical and demographic information. We apply supervised machine learning models—SVM, Random Forest, Decision Tree, KNN, and XGBoost—to classify migraine types based on features such as pain characteristics, nausea, aura, and sensory triggers. The aim is to uncover predictive symptom patterns that can assist in more timely and accurate migraine identification, ultimately contributing to better diagnostic tools and personalized care strategies.

DESCRIPTION OF THE DATA SET

The dataset consists of detailed clinical and demographic information on patients diagnosed with different types of migraines. It was obtained from Kaggle and originally contained 400 observations and 24 variables. After identifying and removing 6 duplicate records, the final dataset includes 394 unique patient records. Initially, all variables were of type int64, but during preprocessing, appropriate variables were converted to categorical types to better reflect their clinical meanings (e.g., symptom presence or pain characteristics).

Table 1:Description of Variables

Variable	Description	Original Type	Final Type
Age	Age of the participant	int64	int64
Duration	Duration of the migraine attack (in hours)	int64	int64
Frequency	Frequency of attack over a specific period	int64	int64
Location	Location of the pain: 1.Unilateral, 2. Bilateral, 3.Frontal	int64	Category
Character	Nature of pain: 1 - Throbbing, 2 - Pressing, 3 - Sharp	int64	Category
Intensity	Pain severity 1. No Pain, 2. Mild, 3. Moderate, 4. Serve	int64	Category
Nausea	Presence of nausea during the episode	int64	Category
Vomit	Presence of Vomiting	int64	Category
Phonophobia	Sensitivity to sound	int64	Category
Photophobia	Sensitivity to light	int64	Category
Visual	Visual disturbances such as aura or light flashes	int64	Category
Sensory	Sensory issues such as numbness or tingling	int64	Category
Motor	Impaired motor function	int64	Category
Language	Language disturbances	int64	Category
Vertigo	Presence of dizziness or vertigo	int64	Category
Tinnitus	Ringing in the ears	int64	Category
Hypoacusis	Decreased hearing ability	int64	Category
Diplopia	Double vision	int64	Category
Defect	Visual field defect or scotoma	int64	Category
Ataxia	Balance or coordination problems	int64	Category
Conscience	Episodes of reduced consciousness	int64	Category
Paresthesia	Abnormal skin sensations such as tingling	int64	Category
DPF	Dysfunctional physical factor	int64	Category
Type	Type of migraine diagnosed, e.g., "Typical aura with migraine". *(Target variable)*	Object	Category

FEATURE ENGINEERING & DATA PREPROCESSING

To prepare the dataset for machine learning analysis, several preprocessing and feature engineering steps were applied. First, the response variable representing migraine type was transformed by regrouping the original seven highly specific clinical migraine diagnoses into four broader and clinically meaningful categories: **Migraine with Typical Aura**, **Migraine without Aura**, **Hemiplegic Migraine Variants**, and **Other Migraine Subtypes**. This mapping strategy simplified the classification task, reduced label noise from rare or overlapping subtypes, and emphasized clinically actionable groupings to support more interpretable model outputs.

Next, the dataset was checked for data quality issues. No missing values or outliers were detected. However, six duplicate records were identified and removed to prevent redundancy and potential model bias. Given that the target classes remained imbalanced after grouping, the **Synthetic Minority Over-Sampling Technique (SMOTE)** was applied to balance the class distribution in the training data. This helped to improve model learning and generalization across all migraine types.

Categorical variables were encoded using appropriate encoding techniques (e.g., one-hot encoding or label encoding, depending on the model requirements), and numerical variables were standardized to ensure consistent scaling, particularly for algorithms sensitive to feature magnitude such as SVM and KNN.

IMPORTANT RESULTS OF THE DESCRIPTIVE ANALYSIS

Preliminary Cluster Identification Using FAMD Scores

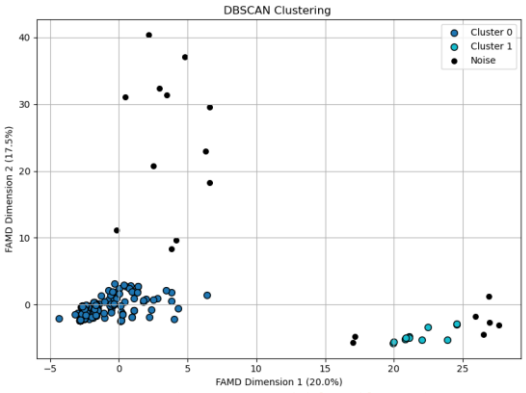

Figure 1: DBSCAN Clustering

To explore potential subgroup structures, we applied DBSCAN clustering on the FAMD dimensions. The algorithm identified two clusters, one with 284 observations and one smaller, homogeneous group with 12 observations, along with 19 noise points. Cluster 0 exhibited a diverse distribution of migraine subtypes, while Cluster 1 consisted entirely of 'Migraine with Typical Aura' cases. The noise group mainly included 'Other Migraine Subtypes' and some instances of 'Migraine with Typical Aura'. However, due to the small sample sizes in the minor clusters and the limited interpretability of the variance captured by the first two FAMD dimensions, we chose not to build separate models per cluster. Instead, we proceeded by treating the entire dataset as a single group to maintain model robustness and interpretability.

Distribution of New Migraine Types

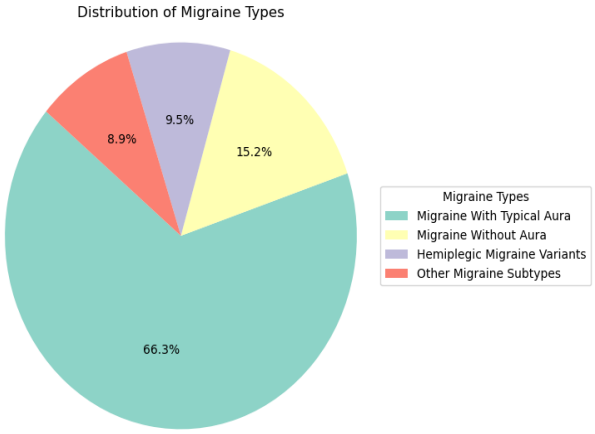

Figure 2: Distribution of New Migraine Types

The regrouped target variable revealed a moderate class imbalance. Migraine With Typical Aura was the most prevalent category, while Hemiplegic Migraine Variants and Other Subtypes were underrepresented. Migraine Without Aura occupied a middle ground. This distribution reflects real-world clinical patterns but also underscores the need to address class imbalance for fair and accurate model training and evaluation. To mitigate this imbalance, SMOTE (Synthetic Minority Over-sampling Technique) was applied before model fitting.

Exploring Nonlinear Separability of Migraine Classes.

The decision boundaries shown in the plot clearly indicate that the class separation is nonlinear. The curves separating the different migraine subtypes (e.g., Hemiplegic, Typical Aura, Without Aura, and Other Subtypes)

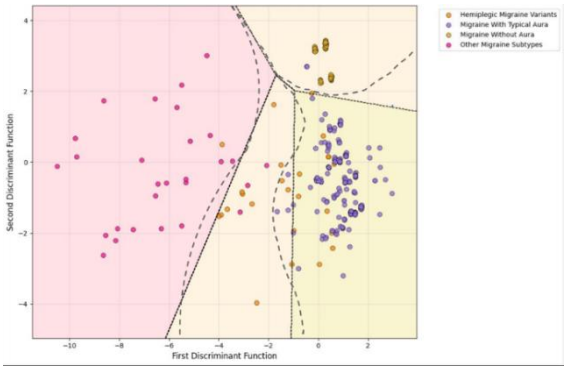

Figure 3: Decision Boundary Plot

are complex and not aligned along straight lines. This complexity suggests that linear models like Linear Discriminant Analysis (LDA) may not capture the true structure of the data effectively. Therefore, it is necessary to use nonlinear classification methods—such as Support Vector Machines with RBF kernel, Random Forests, or Neural Networks—to better model these intricate boundaries and improve classification accuracy.

index	Shapiro-Wilk Statistic	P-value
0	0.4092501599317775	8.488112375353192e-31
1	0.3218851822741686	1.8894445240122925e-32
2	0.6111908706059268	4.602133295524e-26
3	0.8769088750181223	3.407920445835331e-15
4	0.8703151716542197	1.2804729524990764e-15
5	0.9205382526332005	6.989124004653041e-12
6	0.7869053662743142	5.47292008887451e-20
7	0.852572410455529	1.0846619701116757e-16
8	0.8323714080263345	8.367725608787952e-18
9	0.8049367961239219	3.621453830958372e-19

Figure 4: Shapiro- Wilks Test Results

The results of the Shapiro-Wilk test clearly indicate that all p-values are extremely small (much less than 0.05), leading to the rejection of the null hypothesis of normality for all variables. This confirms that the data significantly deviates from a normal distribution. As a result, it is not suitable to apply linear methods that assume normality, such as LDA or QDA, without appropriate transformation or adaptation. This emphasizes the need to explore alternative approaches, including nonlinear or distribution-free classification techniques.

IMPORTANT RESULTS OF THE ADVANCED ANALYSIS

Decision Tree classifier

The Decision Tree classifier yielded a training accuracy of 73.0% and a test accuracy of 60.8%, showing signs of overfitting. It achieved a training F1-score of 0.738, but this dropped to 0.598 on the test set. This gap suggests that the model effectively captured training patterns but struggled to generalize to unseen data. Precision and recall were lower on the test set, particularly for minority classes, suggesting sensitivity to class imbalance.

Table 2: Performance Matrix of Decision Tree

Training set				Test set			
Precision	**Recall**	**F1**	**Accuracy**	**Precision**	**Recall**	**F1**	**Accuracy**
0.790	0.73	0.738	0.730	0.631	0.672	0.598	0.608

XG Boost classifier

Table 3: Performance Matrix of XG Boost

Training set				Test set			
Precision	**Recall**	**F1**	**Accuracy**	**Precision**	**Recall**	**F1**	**Accuracy**
0.702	0.929	0.929	0.929	0.702	0.660	0.677	0.823

The XG Boost classifier showed better overall performance, with a training accuracy of 92.9% and a test accuracy of 82.3%. It maintained high F1-scores on both the training (0.929) and test sets (0.677), reflecting strong predictive power and better generalization. While recall was slightly lower on the test set, the precision remained stable, highlighting the model's robustness even in the presence of class imbalance and complex symptom patterns.

Support Vector Machine (SVM)

Table 4: Performance Matrix of SVM

Training set				Test set			
Precision	**Recall**	**F1**	**Accuracy**	**Precision**	**Recall**	**F1**	**Accuracy**
0.940	0.940	0.940	0.938	0.850	0.820	0.830	0.823

The SVM classifier demonstrated strong performance across both training and test datasets. It achieved a training accuracy of 93.8% and a test accuracy of 82.3%, indicating effective learning and generalization. The model maintained a high F1-score of 0.94 on the training set, with a respectable score of 0.83 on the test set. Although a slight drop in recall was observed during testing likely due to class imbalance the overall precision remained high. This suggests the model is capable of making confident and reliable predictions, even in the presence of overlapping symptoms and imbalanced class distributions.

K Nearest Neighbor (KNN)

Table 5: Performance Matrix of KNN

Training set				Test set			
Precision	**Recall**	**F1**	**Accuracy**	**Precision**	**Recall**	**F1**	**Accuracy**
0.90	0.90	0.90	0.90	0.83	0.77	0.79	0.77

The K-Nearest Neighbors (KNN) classifier achieved solid performance, with a training accuracy of 90% and a test accuracy of 77.2%. The model maintained balanced precision and recall values across training and testing, with an F1-score of 0.90 on the training set and 0.79 on the test set. While performance dipped slightly on unseen data as expected with KNN, the model still demonstrated reasonable generalization. Lower recall for the "Hemiplegic Migraine Variants" and "Other" categories on the test set suggests that KNN may be sensitive to class imbalance or symptom overlap, but overall it provides a reliable baseline for classification in this context.

Random Forest

Table 6: Performance Matrix of Random Forest

Training set				Test set			
Precision	**Recall**	**F1**	**Accuracy**	**Precision**	**Recall**	**F1**	**Accuracy**
0.97	0.97	0.97	0.975	0.86	0.87	0.85	0.855

The Random Forest classifier achieved the best overall performance among the models tested, with 97.5% accuracy on the training set and 85.5% on the test set. Precision, recall, and F1-scores were consistently high and well-balanced across both datasets, indicating strong generalization to unseen data. The model handled class predictions effectively, with minimal performance drop between training and testing. While recall for certain subtypes like "Hemiplegic Migraine Variants" and "Other" was slightly lower, this likely reflects underlying class

imbalance or overlapping symptom patterns. Overall, Random Forest proved to be a robust and reliable model for this classification task.

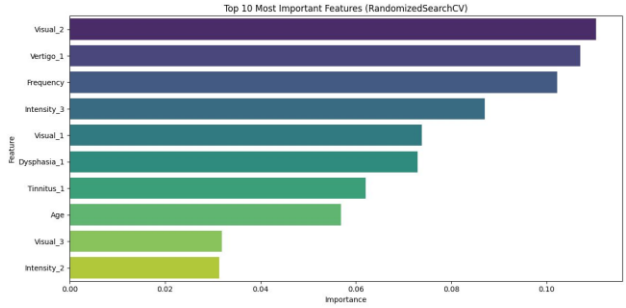

Figure 5: Top 10 Most Important Features of random Forest

Based on its strong and consistent performance, we selected the **Random Forest classifier** as our final prediction model. The bar chart below (figure 5) displays the **top 10 most important features** identified by the model using feature importance scores from the RandomizedSearchCV-optimized Random Forest. Among these, *Visual_2*, *Vertigo_1*, and *Frequency* were the most influential in classifying migraine types, indicating that visual disturbances and symptom frequency play a key role in differentiating between subtypes. Other important predictors include various intensity and sensory-related features, such as *Intensity_3*, *Dysphasia_1*, and *Tinnitus_1*, along with *Age*. These results highlight the complex interplay of symptom characteristics in migraine classification and provide valuable insights for both model interpretation and clinical understanding.

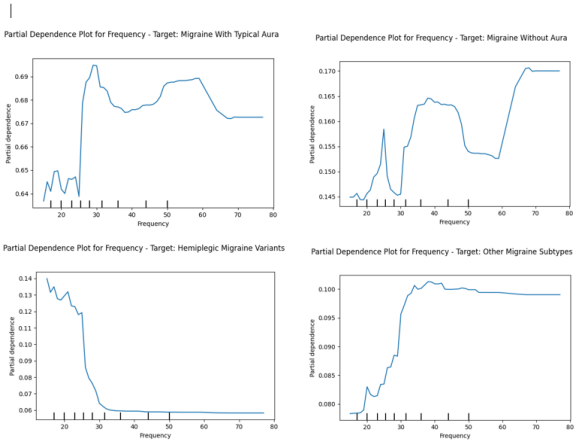

Figure 6: PD Plots of Migraine Frequency

The PDP analysis of migraine **frequency** reveals distinct patterns across different migraine types. **Migraine With Typical Aura** shows a strong positive relationship with frequency, especially between 25–30 attacks, indicating this subtype is more prevalent among patients with frequent migraines. **Migraine Without Aura** exhibits a **bimodal distribution**, with higher probabilities at both moderate (30–45) and very high (65+) frequencies, and a noticeable dip between 45–60. In contrast, **Hemiplegic Migraine Variants** demonstrate a clear **inverse relationship**, being most common among patients with infrequent attacks. Finally, the **Other Migraine Subtypes** category is most likely in patients with 30–50 attacks, indicating a moderate frequency range. These distinct frequency patterns help differentiate migraine subtypes and support more personalized predictions.

SHAP analysis (figure 7) reveals clear patterns in how **visual symptoms** and **vertigo** influence the prediction of different migraine subtypes. High visual symptoms (indicated by red dots) are strong positive predictors for *Migraine With Typical Aura*, with SHAP values ranging from approximately 0.1 to 0.2, while low visual symptoms (blue dots) show negative SHAP values, suggesting they reduce the likelihood of this subtype. In contrast, *Migraine Without Aura* shows the opposite trend—high visual symptoms are associated with negative SHAP values (around –0.1 to –0.2), indicating they reduce the probability of this diagnosis, whereas low visual symptoms are positively associated with it. For *Hemiplegic Migraine Variants*, the visual symptom impact is mixed, with both high and low values showing moderate positive influence. Vertigo shows a weaker and more variable pattern. In Hemiplegic variants, low vertigo slightly increases the probability (clustered around SHAP

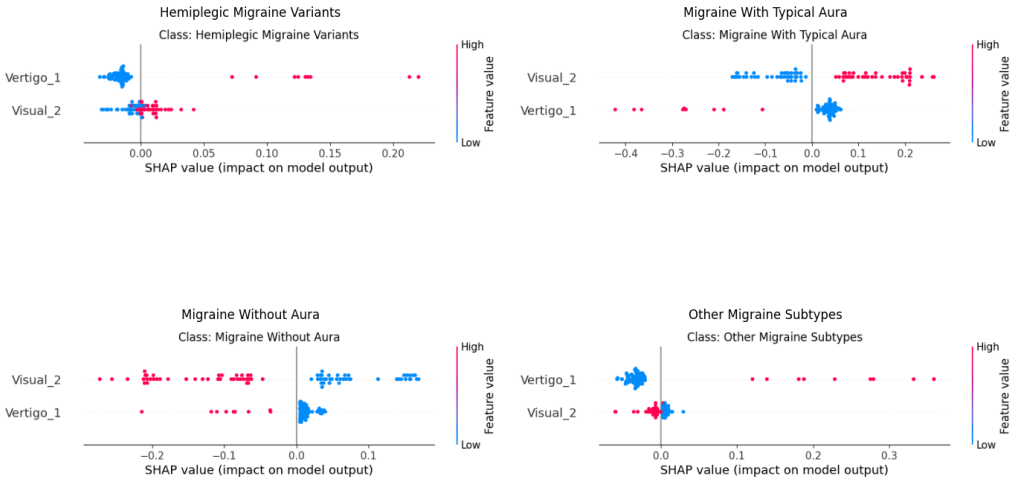

Figure 7: SHAP Value Plots of Important Features

~0.02), and high vertigo has a moderate positive impact. For *Typical Aura*, vertigo appears mostly neutral, while in *Without Aura* and *Other Subtypes*, low vertigo is associated with positive SHAP values, suggesting it may support those diagnoses. Clinically, this means high visual symptoms are strong indicators for *Typical Aura* and rule out *Without Aura*, while low visual symptoms point more toward *Without Aura* or *Other Subtypes*. Vertigo contributes useful but secondary diagnostic cue

ISSUED ENCOUNTED AND PROPOSED SOLUTIONS

During model development, a significant challenge was the **imbalance in the dataset**, with some migraine subtypes having very few samples. To address this, we initially applied **SMOTE (Synthetic Minority Over-sampling Technique)** to balance the class distribution. However, due to the extremely small sample size in some classes, SMOTE generated synthetic data that led to **overfitting**, particularly in models like XGBoost and Random Forest. The models performed well on training data but failed to generalize effectively to unseen data. To overcome this, we **re-categorized the migraine subtypes by grouping similar classes together**, reducing the total number of classes to four. This approach helped create a more balanced dataset and significantly improved model generalization by reducing overfitting, making the predictions more clinically meaningful and stable.

DISCUSSION AND CONCLUSION

Table 7: Summary of all performance Metrics

Model	Training set				Test set			
	Precision	**Recall**	**F1**	**Accuracy**	**Precision**	**Recall**	**F1**	**Accuracy**
Decision Tree	0.790	0.73	0.738	0.730	0.631	0.672	0.598	0.608
XG Boost	0.702	0.929	0.929	0.929	0.702	0.660	0.677	0.823
SVM	0.940	0.940	0.940	0.938	0.850	0.820	0.830	0.823
KNN	0.90	0.90	0.90	0.90	0.83	0.77	0.79	0.77
Random Forest	0.97	0.97	0.97	0.975	0.86	0.87	0.85	0.855

Based on the evaluation metrics across multiple algorithms, the **Random Forest classifier** emerged as the best-performing model for classifying migraine subtypes. It achieved the **highest accuracy on the test set (85.5%)**, along with strong F1-scores and balanced precision and recall, indicating robust generalization performance.

Notably, the **training accuracy (97.5%)** was slightly higher, but the relatively small gap between training and test scores suggests that **overfitting is under control**. The XGBoost classifier also demonstrated solid performance, particularly in recall (92.9%) and F1-score on the training set, and a competitive test accuracy (82.3%), though it showed a slightly larger drop between training and testing metrics. SVM and KNN models performed consistently well with F1-scores and accuracies above 82%, but slightly lower than Random Forest. The Decision Tree model, however, had the lowest test accuracy and F1-score, indicating it is less suitable for this task. Therefore, the **Random Forest model will be used as the core classification engine** for the migraine diagnostic system, offering both high accuracy and reliable generalization.

REFERENCES

https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

https://xgboost.readthedocs.io/en/stable/python/python_api.html#xgboost.XGBClassifier

https://arxiv.org/pdf/2004.05041

https://www.scirp.org/journal/paperinformation?paperid=141994

https://www.mayoclinic.org/diseases-conditions/migraine-headache/symptoms-causes/syc-20360201

https://metana.io/blog/support-vector-machine-svm-classifier-in-python-svm-classifier-python-code/

https://medium.com/data-science/preprocessing-encode-and-knn-impute-all-categorical-features-fast-b05f50b4dfaa

https://www.geeksforgeeks.org/machine-learning/random-forest-regression-in-python/

https://www.aidancooper.co.uk/a-non-technical-guide-to-interpreting-shap-analyses/

APPENDIX

Data Set Link:

- Migraine Symptom Dataset for Classification

Git Hub Link for Code:

- https://github.com/Vayani-Kavindya/ST-4052_project-02.git