**Research Proposal Data 698 – First Draft**

**By Enid Roman**

**Predicting Migraine Types Using Symptom and Demographic Data: A Machine Learning Approach**

**Introduction**

Migraines are a debilitating neurological disorder characterized by severe headaches and a range of associated symptoms, including nausea, sensitivity to light (photophobia), and sound (phonophobia). Affecting millions of people worldwide, migraines are known to significantly reduce quality of life and productivity (Silberstein & Lipton, 2001). Proper diagnosis and classification of different types of migraines are essential for determining effective treatment strategies. However, due to the complexity and variability of symptoms, diagnosing specific migraine types can be challenging for healthcare providers, often leading to a trial-and-error approach in treatments.

My personal connection to this study is rooted in my own experience. For the past 35 years, I have suffered from migraines and have undergone numerous treatments, medications, and therapies in an effort to manage and prevent episodes. Like many patients, I’ve experienced varying degrees of success with these treatments, which are often prescribed based on general migraine classifications rather than personalized symptom profiles. This experience motivated me to explore data-driven methods that could improve migraine diagnosis and potentially lead to more tailored treatment approaches for patients like myself.

The growing field of data science presents an opportunity to apply machine learning techniques to medical data, providing new insights into migraine classification and treatment. By analyzing large datasets of patient records, it may be possible to identify patterns that can help healthcare professionals more accurately diagnose migraine types based on a combination of symptoms and demographic factors. This study aims to leverage these techniques to predict migraine types using a dataset of 400 medical records from migraine patients, identifying key symptoms and characteristics that differentiate between migraine subtypes.

This research will explore the following key question: **Can machine learning models accurately predict the type of migraine a patient experiences based on symptom and demographic data, and which features are most influential in these predictions?**

**Literature Review**

Several studies have examined the classification of migraine types, focusing on both clinical diagnosis and the application of computational methods. Traditional clinical approaches to migraine diagnosis typically classify migraines based on the presence or absence of aura, family history, and the intensity of symptoms such as pain, nausea, and sensitivity to light or sound. For example, Silberstein and Lipton's work (2001) outlines the diagnostic criteria for various types of migraines but emphasizes that misclassification is common due to overlapping symptomatology (Silberstein & Lipton, 2001).

Recent work in applying data science methods to healthcare, particularly in neurological disorders, has shown promising results in improving diagnostic accuracy. Machine learning techniques such as Decision Trees, Random Forests, and Support Vector Machines (SVMs) have been used to predict outcomes in various medical fields. For instance, Lee et al. (2019) applied machine learning to predict the likelihood of chronic migraine development based on patient symptom profiles, demonstrating the utility of data-driven models in this domain (Lee, Kang, & Shin, 2019).

However, there remains a gap in the literature regarding the application of these models to predict specific migraine types. While some studies have focused on binary classifications (e.g., migraine vs. non-migraine), fewer studies have addressed multiclass classification problems, such as distinguishing between migraine types like those with and without aura or hemiplegic migraines. This research aims to fill this gap by applying advanced machine learning models to a multiclass migraine dataset and identifying key predictors for each migraine type.

**Research Question**

The primary research question guiding this study is:

**Can we accurately predict the type of migraine a patient experiences based on symptom characteristics and demographic factors using machine learning models?**

To explore this, the study will also address the following sub-questions:

1. What are the most important features (symptoms, demographic factors) in predicting migraine type?
2. How do machine learning models like Random Forests and Support Vector Machines perform in classifying migraine types?
3. How does the accuracy of predictive models vary across different migraine subtypes?

**Data and Variables**

The dataset comprises 400 medical records of patients diagnosed with various types of migraines, collected at the Centro Materno Infantil de Soledad in 2013 (Ranzeet, 2023). The data includes 24 attributes that capture patient demographics, symptoms, and migraine diagnosis.

**Key Variables:**

* **Dependent Variable:**
  + **Type of Migraine:** Categorical variable with seven classes (Typical aura with migraine, Migraine without aura, etc.).
* **Independent Variables:**
  + **Age**: Continuous (real) variable.
  + **Duration of symptoms:** Continuous variable (days).
  + **Frequency of episodes:** Continuous variable (episodes per month).
  + **Location of pain:** Categorical (None, Unilateral, Bilateral).
  + **Character of pain:** Categorical (None, Throbbing, Constant).
  + **Pain intensity:** Categorical (None, Mild, Medium, Severe).
  + **Nausea, Vomit, Phonophobia, Photophobia:** Binary categorical variables (Yes/No).
  + **Neurological symptoms:** Binary variables for conditions such as vertigo, tinnitus, diplopia, and ataxia.

The data is clean, with no missing values, making it suitable for immediate analysis.

**Statistical Methods**

This study will use a combination of exploratory data analysis (EDA), feature selection techniques, and machine learning classification algorithms to address the research questions, all of which will be implemented using the **R programming language**. R is chosen due to its extensive range of packages tailored for statistical analysis, machine learning, and data visualization.

1. **Exploratory Data Analysis (EDA)**:
   * The R packages **ggplot2** and **dplyr** will be used to visualize the distribution of migraine types and key symptom variables. Histograms, box plots, and heatmaps will be generated to analyze correlations between symptoms and demographic factors such as age and family history.
   * The cor() function and ggcorrplot package will be utilized to visualize correlation matrices for identifying significant relationships between variables.
2. **Feature Selection**:
   * **Pearson correlation** and **chi-squared tests** will be computed using R's cor.test() and chisq.test() functions to assess relationships between the independent and dependent variables.
   * Feature importance will be assessed using **Recursive Feature Elimination (RFE)**, facilitated by the **caret** package, and **SHAP values**, which can be computed with iml or other compatible packages in R to rank the importance of various symptoms in predicting migraine types.
3. **Classification Models**:
   * **Random Forest**: This ensemble learning method will be implemented using the **randomForest** package, which provides insights into feature importance and performs well in multiclass classification problems.
   * **Support Vector Machine (SVM)**: The **e1071** package will be used to train SVM models to classify migraine types by finding the optimal hyperplane.
   * **Logistic Regression (Multinomial)**: A multinomial logistic regression model will be built using the **nnet** package to serve as a baseline for comparing more complex models.
4. **Model Evaluation**:
   * Models will be evaluated using cross-validation techniques with the **caret** package to ensure generalizability and prevent overfitting. Performance will be measured using accuracy, precision, recall, F1-score, and the area under the ROC curve (AUC), calculated using R's **pROC** and **caret** packages.
   * The **confusion matrix** will be generated using **caret::confusionMatrix** to assess classification performance for each migraine type. Additionally, **feature importance** will be evaluated and visualized using tools like randomForest::importance() and ggplot2 for clear representation of key factors.

**References Section**

1. Ranzeet. (2023). *Migraine Dataset*. Kaggle. Retrieved from <https://www.kaggle.com/datasets/ranzeet013/migraine-dataset/data>
2. Silberstein, S. D., & Lipton, R. B. (2001). Chronic daily headache including transformed migraine, chronic tension-type headache, and medication overuse. *The Chronic Pain Handbook*, 2(4), 1-14. Retrieved from <https://www.aafp.org/pubs/afp/issues/2011/0201/p271.html>
3. Lee, S. H., Kang, J. W., & Shin, Y. H. (2019). Machine learning approaches for predicting migraine development: A comparison of decision trees and support vector machines. *Journal of Headache and Pain*, 20(8), 110-118. doi:10.3389/fneur.2023.1106612. Retrieved from https://www.frontiersin.org/journals/neurology/articles/10.3389/fneur.2023.1106612/full