A Data Mining Approach to Predicting the Likelihood of Multiple Sclerosis Progression Using Clinical and MRI Data

Evelyn C Tadlock[1]

Texas Woman's University, Denton TX 76204, USA elister@twu.edu
http://www.twu.edu

Abstract. Multiple Sclerosis (MS) is a chronic autoimmune disease that affects the human central nervous system. Its progression varies and is often unpredictable. Understanding the medical factors that contribute to MS progression from Clinically Isolated Syndrome (CIS) to Clinically Definitive Multiple Sclerosis (CDMS) is crucial for patient quality-of-life management. This project explores the use of data mining methodologies to analyze the clinical and MRI data used to identify the patterns and factors used to diagnose MS progression. The project focuses on data cleaning and preprocessing steps, scaling, and feature selection to ensure data integrity. Correlation analysis and exploratory data visualizations reveal significant relationships between clinical and imaging features. Using algorithms such as Support Vector Machines (`SVMs`) and K-nearest neighbors allows the classification of CIS patients based on their likelihood to progress to CDMS. Key findings demonstrate the importance of features such as spinal cord MRIs, oligoclonal bands, and initial Expanded Disability Status Scale (EDSS) scores in predicting the progression of MS. This study underscores the potential of data mining techniques to complement traditional diagnostic methods and provides a foundation for further research into early prediction and management of MS progression.

1 Introduction

1.1 What is Multiple Sclerosis (MS)?

Multiple Sclerosis (MS) is an autoimmune disease that affects the central nervous system. MS has occurred in nearly 1 million people in the United States alone since 2019. Throughout the globe, international MS organizations estimate that 2.8 million people have MS [5].

Damage to the protective coverings of nerve fibers, called myelin, in the brain and spinal cord are classical MS characteristics [12]. The clinical term for this is called inflammatory de-myelination with axonal transection. MS manifests itself in several ways, each with distinct patterns of disease progression in each and every patient. Also, as time progresses, the MS progresses into different stages. As shown in figure 1, the progression patterns of MS subtypes have distinct and progressive trajectories.

Clinically Isolated Syndrome (CIS) Clinically Isolated Syndrome (CIS) is the earliest stage and presentation of MS symptoms. CIS involves a single neurological episode that may progress to MS or not. Between 30 and 70% of patients who have CIS will develop MS and will develop it within 3 months of CIS symptoms [7]. CIS presents itself with the following features described in Table 1.

Table 1. Predicting Features According to McDonald et al. [4].

Feature	**Description**
Periventricular Lesions	Lesions in the periventricular region of the brain.
Infratentorial Lesions	Lesions in the brainstem or cerebellum.
Juxtacortical Lesions	Lesions near the cortex
Spinal Cord Lesions	Lesions on the spinal cord
Gadolinium-Enhancing Lesions	Lesions with active inflammation seen on MRI.
Oligoclonal Bands in CSF	Immune activity indicated by unique IgG bands in the CSF.

Relapsing-Remitting MS (RRMS) RRMS is the most common form of Multiple Sclerosis, for which 85% of all MS cases account to occur. With RRMS, a patient will have episodes of various symptom relapses, then followed up by periods of remission. During an MS patient's remission period, symptoms may only partially or wholly disappear [7].

Secondary Progressive MS (SPMS) Secondary progressive MS (SPMS) is when the disease has surpassed the relapsing-remitting stage and now involves a steady worsening of a patient's neurological function. SPMS occurs in about 50-80% of individuals with MS [7].

Primary Progressive MS (PPMS) Primary Progressive MS (PPMS) is a less common form that affects approximately 10% of individuals with MS. It is characterized by a gradual progression of symptoms without distinct relapses [7].

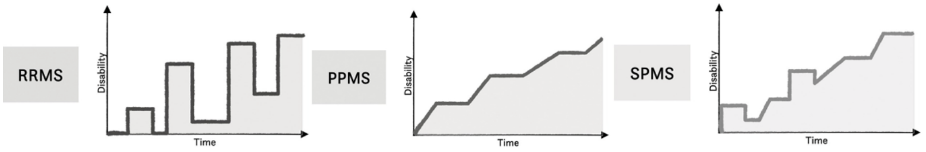


Fig. 1. Progression patterns for MS subtypes: RRMS, SPMS, and PPMS (adapted from Pérez et al) [10]

1.2 MS Features and Definitions

To understand this project's domain-specific problem, some of the clinical terms used in this project and the following dataset analysis need to be defined. Definitions are limited to the clinical features of note during exploratory data analysis. Below are the key features in the dataset, along with the explanations of their medical significance:

- **Oligoclonal Bands**: The group of proteins known as immunoglobulins detected in a patient cerebrospinal fluid (CSF). CSF ,obtained through Lumbar Puncture (a Spinal Tap), will exhibit Oligoclonal Bands. These bands can indicate inflammation in the Central Nervous System (CNS) and are used to diagnose MS [6].
- **Spinal Cord MRI**: This is an MRI of the patient's spinal cord that indicates whether legions are present. Spinal cord lesions are critical indicators of MS progression, specifically when there are motor control and sensory impairments.
- **Periventricular MRI**: An MRI of the patient's brain. The test will show the indication of the presence of lesions around the brain;s ventricles. Brain lesions are a hallmark trait of MS. Brain lesions are associated with demyelination and inflammation, the key characteristics of MS.
- **Infratentorial MRI**: An MRI of a patient;s brain stem or cerebellum in order to reveal lesions.
- **Cortical MRI**: An MRI of a patient;s cerebral cortex. Legion exhibiting on the cortical are less common.
- **Initial EDSS and Final EDSS**: The Expanded Disability Status Scale (EDSS) is widely used to quantify disability progression in MS patients. The initial score reflects the patient's disability level during the clinical evaluation. It evaluates several functional systems, emphasizing mobility and walking ability. The final EDSS score measures a patient's MS progression over time. The comparison of the initial and final EDSS scores provides insight into the rate of MS progression [8].

2 Related Work

The progression of MS from CIS has previously been studied extensively using clinical and imaging markers. Traditional diagnoses, such as the Poser Criteria [9], relied heavily on clinical and CSF findings. These criteria were revolutionized by the McDonald criteria, which incorporate MRI evidence to demonstrate dissemination [4]. The McDonald criteria were revised in 2017, which allowed for earlier diagnoses and highlighted key predictive features, such as periventricular and spinal cord lesions [5, 11].

Traditional scientific studies, including the prospective study [1] that was the basis for the dataset used in this project, have relied heavily on statistical modeling approaches such as the Cox proportional hazards models to identify significant predictors, including T2-hyperintense lesions, oligoclonal bands, and spinal cord lesions. These methods proved critical to glean insights into MS progression.

3 Problem Statement

The goal of this project is to investigate if data mining methodologies and classification algorithms can complement traditional approaches by uncovering patterns in the clinical and MRI data. The following questions aimed to answer are:

- Can data mining techniques identify the relationship between clinical features and the progression of MS, enhancing traditional statistical methods?
- What features are the best predictions of the progression of MS (CIS to CDMS)?
- Do algorithms, like `KNN` and `SVM`, compare to traditional statistical models in predicting the progression of MS?

By answering these questions, this project aims to highlight the potential of data mining techniques to supplement and enhance MS diagnosis, resulting in better patient care. The goal is to support early decision-making, which is critical for MS patients and their families.

4 Methodology

4.1 Overview of Methodology

The project uses a step-by-step data mining methods to study and analyze how Multiple Sclerosis (MS) progresses from Clinically Isolated Syndrome (CIS) to Clinically Definite Multiple Sclerosis (CDMS). The methodology techniques used include important steps such as preparing the data through preprocessing, choosing important features and using classification methods. The classification algorithms chosen are SVM and `k-NN` models. The steps were carefully chosen and designed to extract meaningful patterns from the clinical and MRI data to improve prediction accuracy, and to evaluate the contribution of individual features.

4.2 Data Preprocessing

The efficient preproccesing of data is crucial to ensure the reliability of data quality and the resulting models. the follow steps below were preformed:

Missing Values Missing data values were handled based on the types of features. If a feature was categorical, such as Initial Symptom, a group-wise mode imputation was applied. Numerical features, such as Initial EDSS, the missing values were replaced using group-wise median imputation.

Scaling Methodology In order to standardized the numerical features and improve overall model performance, a Min-Max scaling was applied manually. This method normalized feature values to a [0,1] range and ensured that compatibility with distance-based algorithms, such as `k-NN`,and enhanced the numerical stability of `SVM`.

Feature Selection The selection of features focused on identifying clinically significant variables based on the exploratory data analysis and insights from medical journals. The key features include: Initial Symptoms, Initial EDSS, Preiventricualr MRI, Spinal Cord MRI, and Oligoclonal Bands, which are known to be associated with MS Progression.

4.3 Classification Algorithms

Support Vector Machines (SVM) Support Vector Machines (`SVM`s) are a nonlinear and linear data classification method. The SVM algorithm uses nonlinear mapping to transform the training data into higher dimensions. It then searches for the decision boundary separating the tuple's different classes by using a hyperplane defined by support vectors and margins[3]. `SVM`s are used for numerical predictions and classifications, making them ideal for predicting MS progression from CIS to CDMS. The `SVM` classifier is effective in handling high-dimensional spaces. It seeks to find an optimal hyperplane that separates data points into their respective classes while maximizing the margin between them. The hyperparameters `C` (regularization parameter) and `gamma` (kernel coefficient for RBF kernels) were tuned to improve model performance.

```
Input: Training data {(x_1, y_1), (x_2, y_2),
    ..., (x_n, y_n)},
        where x_i is a feature vector and y_i
            is the class label {-1, +1}.
1. Choose a kernel function (e.g., linear,
    polynomial, RBF).
2. Initialize hyperparameters: C (
    regularization parameter) and gamma (if
    using RBF).
```

```
3. Solve the optimization problem to find the
    hyperplane that maximizes the margin.
4. Use the support vectors to define the
    decision boundary.
5. Classify new data points based on their
    position relative to the hyperplane.
Output: Predicted class label for new data.
```

K-Nearest Neighbors (k-NN) The k-Nearest-Neighbor (k-NN) method of classification works by comparing a test tuple to a similar training tuple. Each training tuple is then defined by multiple features, placing it as a point in a multi-dimensional space [3]. To classify an unseen tuple, the algorithm identifies the k nearest training tuples in that space, which are its "nearest neighbors." For classification, the unseen tuple is assigned to the class that is the most common relationship among the other k-nearest neighbors. When the number of neighbors equals 1, the tuple takes on the class of its single nearest neighbor. This method is also able to predict numeric values by averaging the labels of the k-nearest neighbors [3].The simplistic and adaptable nature of k-NN classification makes it an effective tool for analyzing clinical and MRI data.

```
Input: Training data {(x_1, y_1), (x_2, y_2),
    ..., (x_n, y_n)},
        test point x, and parameter k.
1. Calculate the distance between the test
    point x and all training points.
2. Sort all training points by ascending
    distance.
3. Select the k nearest neighbors.
4. Count the class labels of the k neighbors.
5. Assign the majority class label to the test
     point.
Output: Predicted class label for the test
    point.
```

4.4 General Pseudocode for Project Entirety

```
1. Load and preprocess the dataset.
2. Select clinically significant features.
3. Split the data into training and testing sets.
4. Scale features using Min-Max scaling.
5. Train the SVM and KNN classifiers:
a. SVM: Tune hyperparameters (C, gamma).
```

```
   b. KNN: Test with varying values of k.
6. Evaluate models using accuracy, ROC-AUC, and confusion matrices.
7. Interpret results to identify key predictors of MS progression.
```

5 Experiment

5.1 Dataset Description

The dataset was downloaded from Kaggle [2]. It originated from collected data, a part of a study by the National Institute of Neurology and Neurosurgery in Mexico City. This study was conducted with Mexican patients newly diagnosed with CIS between 2006 and 2010. It followed all 273 patients who had met the study enrollment criteria over 10 years and observed their progression from CIS to CDMS [1].

The tables below provide an overview of the dataset's columns, which include clinical and MRI being used to analyze MS progression. Table 3 details the numerical encoding for the Initial Symptoms column, which describes the symptom/s observed in the patients during their diagnosis.

Table 2. Dataset Column Descriptions.

Column Name	Description
ID	Patient identifier.
Age	Patient age in years.
Schooling	Years of education.
Gender	Patient gender (1 = male, 2 = female).
Breastfeeding	Breastfeeding status (1 = yes, 2 = no, 3 = unknown).
Varicella	Varicella status (1 = positive, 2 = negative, 3 = unknown).
Initial_Symptoms	Initial symptoms (see Table 3).
Mono_or_Polysymptomatic	Symptom type (1 = monosymptomatic, 2 = polysymptomatic, 3 = unknown).
Oligoclonal_Bands	Oligoclonal bands in CSF (0 = negative, 1 = positive, 2 = unknown).
LLSSEP	Lower limb evoked potentials (0 = negative, 1 = positive).
ULSSEP	Upper limb evoked potentials (0 = negative, 1 = positive).
VEP	Visual evoked potentials (0 = negative, 1 = positive).
BAEP	Brainstem auditory evoked potentials (0 = negative, 1 = positive).
Periventricular_MRI	Periventricular MRI findings (0 = negative, 1 = positive).
Cortical_MRI	Cortical MRI findings (0 = negative, 1 = positive).
Infratentorial_MRI	Infratentorial MRI findings (0 = negative, 1 = positive).
Spinal_Cord_MRI	Spinal cord MRI findings (0 = negative, 1 = positive).
initial_EDSS	Initial EDSS score.
final_EDSS	Final EDSS score.
Group	Patient classification (1 = CDMS, 2 = non-CDMS).

Table 3. Mapping for Initial_Symptoms Column.

Value	Symptoms
1	Visual
2	Sensory
3	Motor
4	Other
5	Visual and Sensory
6	Visual and Motor
7	Visual and Other
8	Sensory and Motor
9	Sensory and Other
10	Motor and Other
11	Visual, Sensory, and Motor
12	Visual, Sensory, and Other
13	Visual, Motor, and Other
14	Sensory, Motor, and Other
15	Visual, Sensory, Motor, and Other

5.2 Implementation Platforms

This project was implemented with the scripting language Python within the Juypter Notebook environment, utilizing the following environments, libraries, and systems.

- **pandas**: For data loading, manipulation, and cleaning.
- **numpy**: For numerical computations and array handling.
- **scikit-learn**: For implementing classification algorithms (`SVM` and `k-NN`), preprocessing steps (scaling and imputation), and evaluation metrics such as accuracy, F1-score, and ROC-AUC.
- **matplotlib**: For creating visualizations, including histograms and correlation heatmaps.
- **seaborn**: For enhanced data visualization, such as correlation matrices.
- **scipy**: For advanced statistical computations.
- **statsmodels**: For statistical summaries (used in feature selection or correlation analysis).

5.3 Data Preprocessing

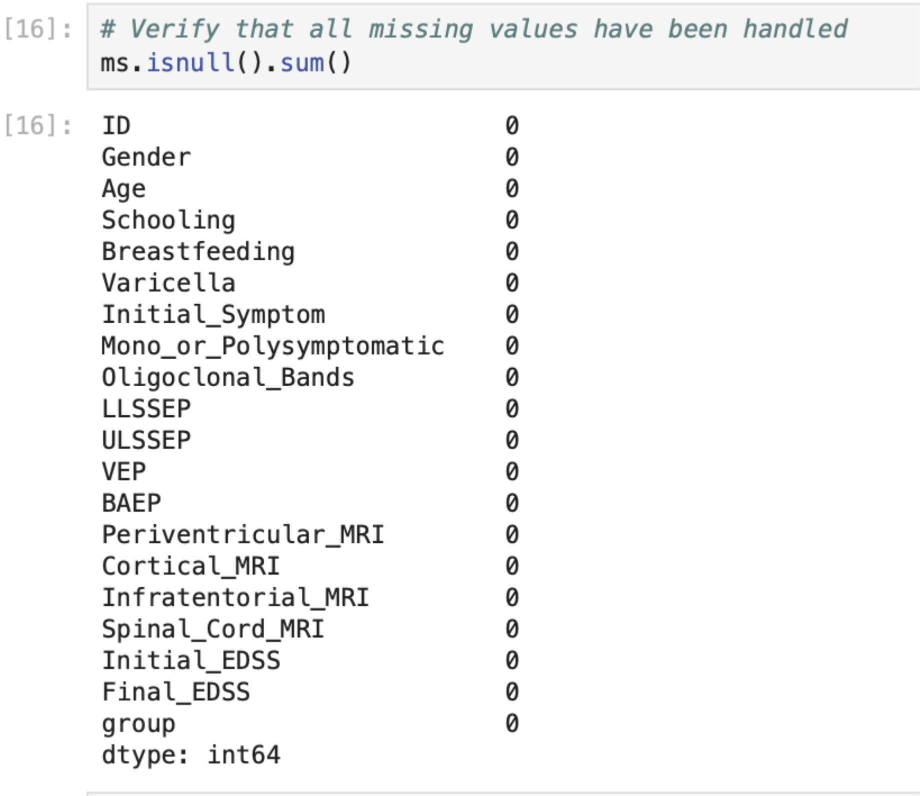
Missing Values Before preprocessing, the dataset had missing values in the following features:

- **Schooling:** 1 missing value
- **Initial Symptom:** 1 missing value
- **Initial EDSS:** 148 missing value
- **Final EDSS:** 148 missing value

In order to address these missing values the following methods were utilized as discussed in the methodology section:

- **Mode Imputation:**: Used for the categorical features.
- **Median Imputation:** Used for the numerical features.

After the imputation, the missing values were successfully handled as show in figure 2 below.

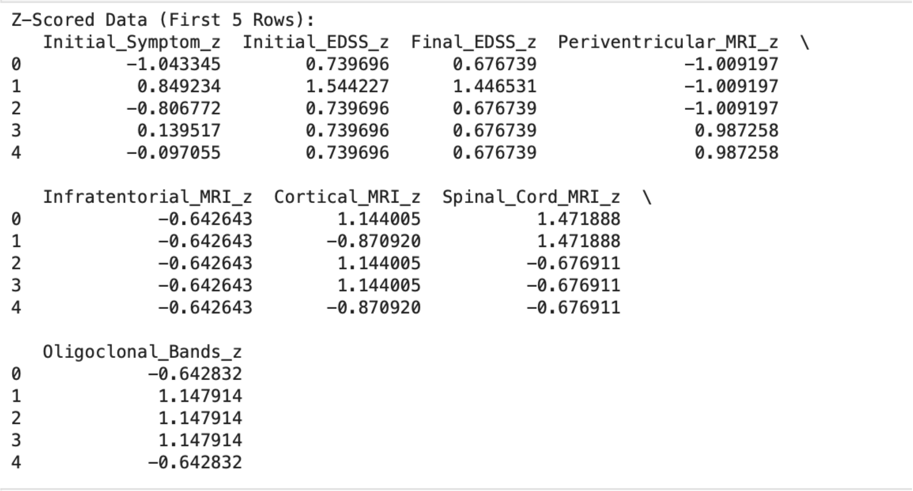


```
[16]: # Verify that all missing values have been handled
      ms.isnull().sum()
```

```
[16]: ID                          0
      Gender                      0
      Age                         0
      Schooling                   0
      Breastfeeding               0
      Varicella                   0
      Initial_Symptom             0
      Mono_or_Polysymptomatic     0
      Oligoclonal_Bands           0
      LLSSEP                      0
      ULSSEP                      0
      VEP                         0
      BAEP                        0
      Periventricular_MRI         0
      Cortical_MRI                0
      Infratentorial_MRI          0
      Spinal_Cord_MRI             0
      Initial_EDSS                0
      Final_EDSS                  0
      group                       0
      dtype: int64
```

Fig. 2. Dataset after handling missing values, showing all features are complete.

Z-Score Normalization Z-score normalization was performed as part of the data preprocessing pipeline for all features measured quantitatively. This ensured that the mean of each feature was 0 with a standard deviation of 1. Normalization significantly detected the outliers that could sway the overall modeling technique. Z-score normalization was merely a diagnostic measure, while the final normalization was done using Min-Max scaling to make it compatible with distance-based algorithms like `k-NN`. Also, Min-Max helps prevent numerical instability when training the `SVM`.

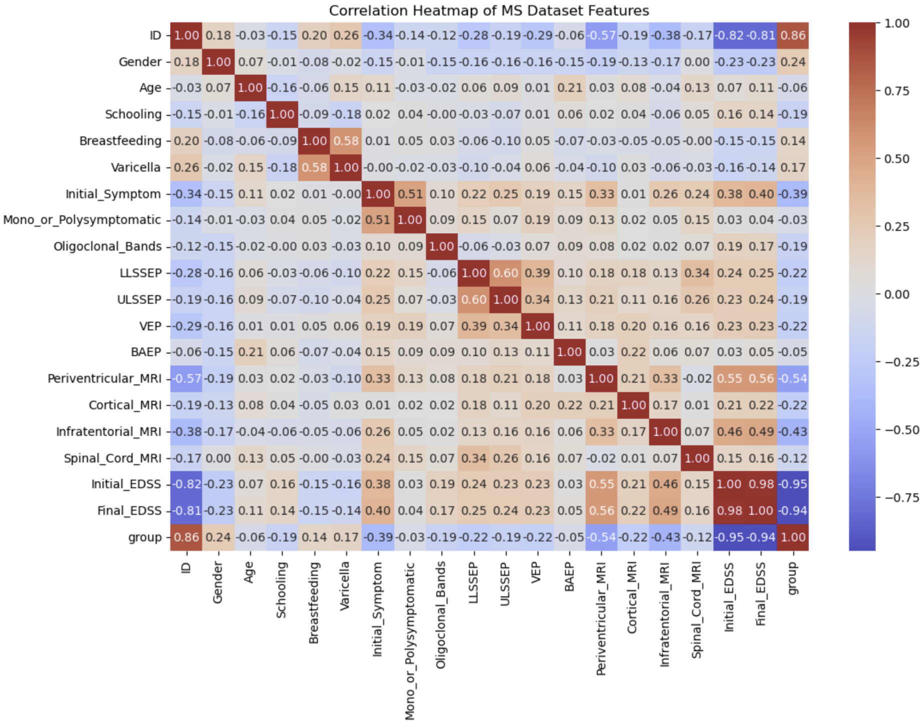


```
Z-Scored Data (First 5 Rows):
   Initial_Symptom_z  Initial_EDSS_z  Final_EDSS_z  Periventricular_MRI_z  \
0          -1.043345        0.739696      0.676739              -1.009197
1           0.849234        1.544227      1.446531              -1.009197
2          -0.806772        0.739696      0.676739              -1.009197
3           0.139517        0.739696      0.676739               0.987258
4          -0.097055        0.739696      0.676739               0.987258

   Infratentorial_MRI_z  Cortical_MRI_z  Spinal_Cord_MRI_z  \
0             -0.642643        1.144005           1.471888
1             -0.642643       -0.870920           1.471888
2             -0.642643        1.144005          -0.676911
3             -0.642643        1.144005          -0.676911
4             -0.642643       -0.870920          -0.676911

   Oligoclonal_Bands_z
0            -0.642832
1             1.147914
2             1.147914
3             1.147914
4            -0.642832
```

Fig. 3. Z-Score Testing

5.4 Exploratory Data Analysis (EDA)

Exploration of the data was performed to better understand the dataset and guide feature selection. A correlation matrix was created to visualize the relationship between features and the target variable, which was chosen to be the feature groups. The groups consist of two different groups of patients: those with CIS and those with CDMS. The analysis result here was that there is the following correlations as seen in figure 3.



	ID	Gender	Age	Schooling	Breastfeeding	Varicella	Initial_Symptom	Mono_or_Polysymptomatic	Oligoclonal_Bands	LLSSEP	ULSSEP	VEP	BAEP	Periventricular_MRI	Cortical_MRI	Infratentorial_MRI	Spinal_Cord_MRI	Initial_EDSS	Final_EDSS	group
ID	1.00	0.18	-0.03	-0.15	0.20	0.26	-0.34	-0.14	-0.12	-0.28	-0.19	-0.29	-0.06	-0.57	-0.19	-0.38	-0.17	-0.82	-0.81	0.86
Gender	0.18	1.00	0.07	-0.01	-0.08	-0.02	-0.15	-0.01	-0.15	-0.16	-0.16	-0.16	-0.15	-0.19	-0.13	-0.17	0.00	-0.23	-0.23	0.24
Age	-0.03	0.07	1.00	-0.16	-0.06	0.15	0.11	-0.03	-0.02	0.06	0.09	0.01	0.21	0.03	0.08	-0.04	0.13	0.07	0.11	-0.06
Schooling	-0.15	-0.01	-0.16	1.00	-0.09	-0.18	0.02	0.04	-0.00	-0.03	-0.07	0.01	0.06	0.02	0.04	-0.06	0.05	0.16	0.14	-0.19
Breastfeeding	0.20	-0.08	-0.06	-0.09	1.00	0.58	0.01	0.05	0.03	-0.06	-0.10	0.05	-0.07	-0.03	-0.05	-0.05	-0.00	-0.15	-0.15	0.14
Varicella	0.26	-0.02	0.15	-0.18	0.58	1.00	-0.00	-0.02	-0.03	-0.10	-0.04	0.06	-0.04	-0.10	0.03	-0.06	-0.03	-0.16	-0.14	0.17
Initial_Symptom	-0.34	-0.15	0.11	0.02	0.01	-0.00	1.00	0.51	0.10	0.22	0.25	0.19	0.15	0.33	0.01	0.26	0.24	0.38	0.40	-0.39
Mono_or_Polysymptomatic	-0.14	-0.01	-0.03	0.04	0.05	-0.02	0.51	1.00	0.09	0.15	0.07	0.19	0.09	0.13	0.02	0.05	0.15	0.03	0.04	-0.03
Oligoclonal_Bands	-0.12	-0.15	-0.02	-0.00	0.03	-0.03	0.10	0.09	1.00	-0.06	-0.03	0.07	0.09	0.08	0.02	0.02	0.07	0.19	0.17	-0.19
LLSSEP	-0.28	-0.16	0.06	-0.03	-0.06	-0.10	0.22	0.15	-0.06	1.00	0.60	0.39	0.10	0.18	0.18	0.13	0.34	0.24	0.25	-0.22
ULSSEP	-0.19	-0.16	0.09	-0.07	-0.10	-0.04	0.25	0.07	-0.03	0.60	1.00	0.34	0.13	0.21	0.11	0.16	0.26	0.23	0.24	-0.19
VEP	-0.29	-0.16	0.01	0.01	0.05	0.06	0.19	0.19	0.07	0.39	0.34	1.00	0.11	0.18	0.20	0.16	0.16	0.23	0.23	-0.22
BAEP	-0.06	-0.15	0.21	0.06	-0.07	-0.04	0.15	0.09	0.09	0.10	0.13	0.11	1.00	0.03	0.22	0.06	0.07	0.03	0.05	-0.05
Periventricular_MRI	-0.57	-0.19	0.03	0.02	-0.03	-0.10	0.33	0.13	0.08	0.18	0.21	0.18	0.03	1.00	0.21	0.33	-0.02	0.55	0.56	-0.54
Cortical_MRI	-0.19	-0.13	0.08	0.04	-0.05	0.03	0.01	0.02	0.02	0.18	0.11	0.20	0.22	0.21	1.00	0.17	0.01	0.21	0.22	-0.22
Infratentorial_MRI	-0.38	-0.17	-0.04	-0.06	-0.05	-0.06	0.26	0.05	0.02	0.13	0.16	0.16	0.06	0.33	0.17	1.00	0.07	0.46	0.49	-0.43
Spinal_Cord_MRI	-0.17	0.00	0.13	0.05	-0.00	-0.03	0.24	0.15	0.07	0.34	0.26	0.16	0.07	-0.02	0.01	0.07	1.00	0.15	0.16	-0.12
Initial_EDSS	-0.82	-0.23	0.07	0.16	-0.15	-0.16	0.38	0.03	0.19	0.24	0.23	0.23	0.03	0.55	0.21	0.46	0.15	1.00	0.98	-0.95
Final_EDSS	-0.81	-0.23	0.11	0.14	-0.15	-0.14	0.40	0.04	0.17	0.25	0.24	0.23	0.05	0.56	0.22	0.49	0.16	0.98	1.00	-0.94
group	0.86	0.24	-0.06	-0.19	0.14	0.17	-0.39	-0.03	-0.19	-0.22	-0.19	-0.22	-0.05	-0.54	-0.22	-0.43	-0.12	-0.95	-0.94	1.00

Fig. 4. Correlation Heatmap of MS Data Features

Next, pairwise distributions between the feature groups (non-CDMS (CIS) and CDMS) and other features were also visualized. The visualizations below revealed distinct patterns that differentiated the two groups. For example, key features such as Initial symptoms and Oligoclonal Bands presented clear separations between CDMS and non-CDMS groups. CDMS cases presented with a higher prevalence of positive Oligoclonal Bands.

MRI-related features, including Periventricular and Spinal Cord MRI, also showed strong associations with CDMS progression. In particular, the CDMS cases had a significantly higher frequency of positive findings in these MRI categories compared to the non-CDMS cases. In comparison, EDSS scores displayed distinct distributions, with higher values correlating with CDMS progression.

It is essential to note that final EDSS scores may be more predictive of a patient already in CDMS since they represent a post-diagnosis measurement. Because of this, Final EDSS was removed from feature selection during the model training to prevent data leakage and ensure that the model focused on relevant features of the CIS stage.

These EDA findings underscore how important features such as Oligoclonal Bands and MRI results are in distinguishing between CDMS and non-CDMS groups, thus providing critical insights for accurate modeling.

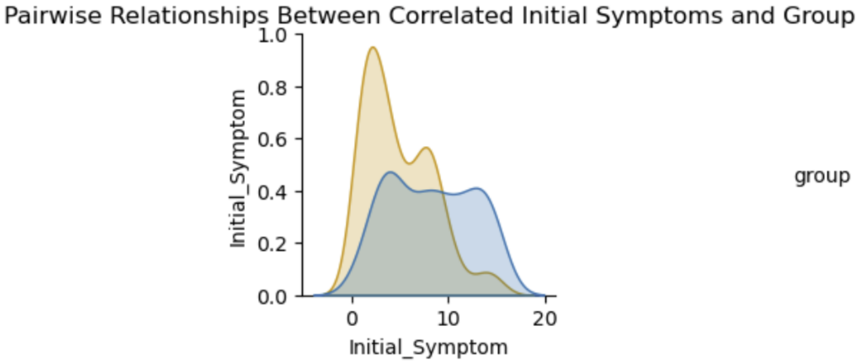


Fig. 5. Pairwise Relationships Between Correlated Initial Symptoms and Group.

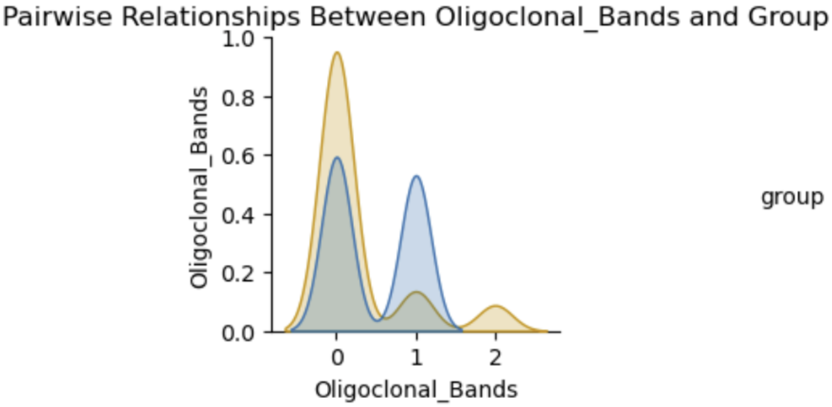


Fig. 6. Pairwise Relationships Between Oligoclonal Bands and Group.

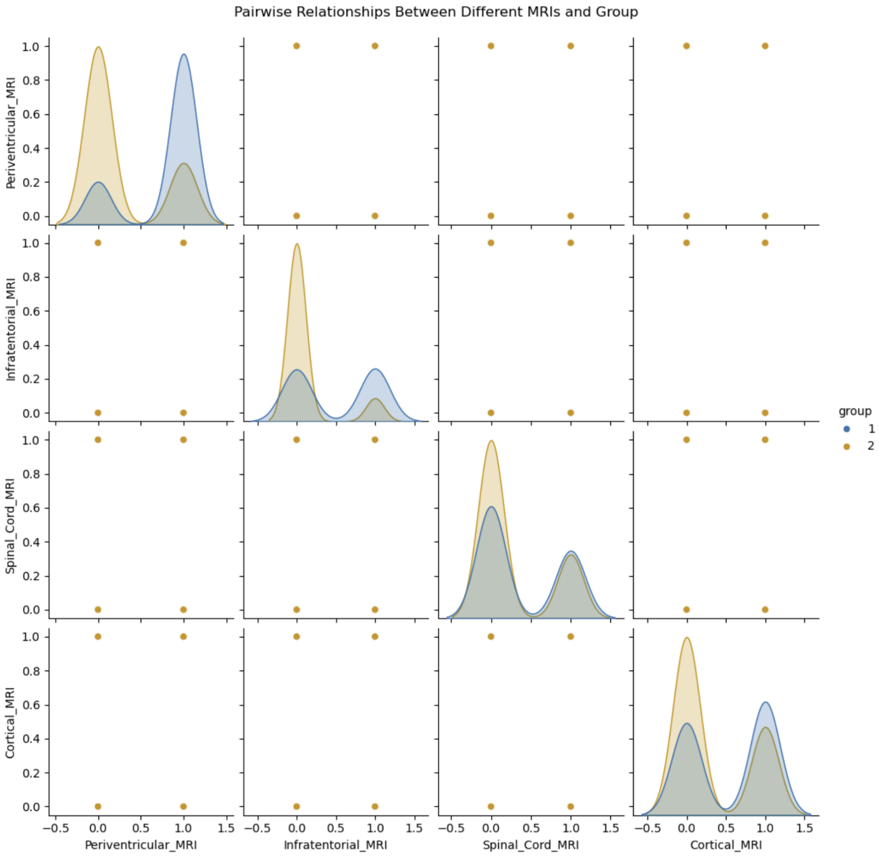


Fig. 7. Different MRI by Group.

Bar charts were also added to visualize the distribution of various category features alongside pairwise visualizations. While pairwise visualizations emphasize relationships and possible separations between CIS and CDMS, bar charts, on the other hand, provide an additional view as they look at the overall prevalence and frequency of categorical variables across groups. For instance, the bar chart of spinal cord MRI findings shows the difference in positive MRI findings between groups, immeasurably emphasizing the importance of MS progression prognosis. In these two visualizations, analyses combine subtle balance to open individual relationships and depict broader group trends. This approach provides the basis for identifying clinically relevant features, thus augmenting further feature selection and building the model step.

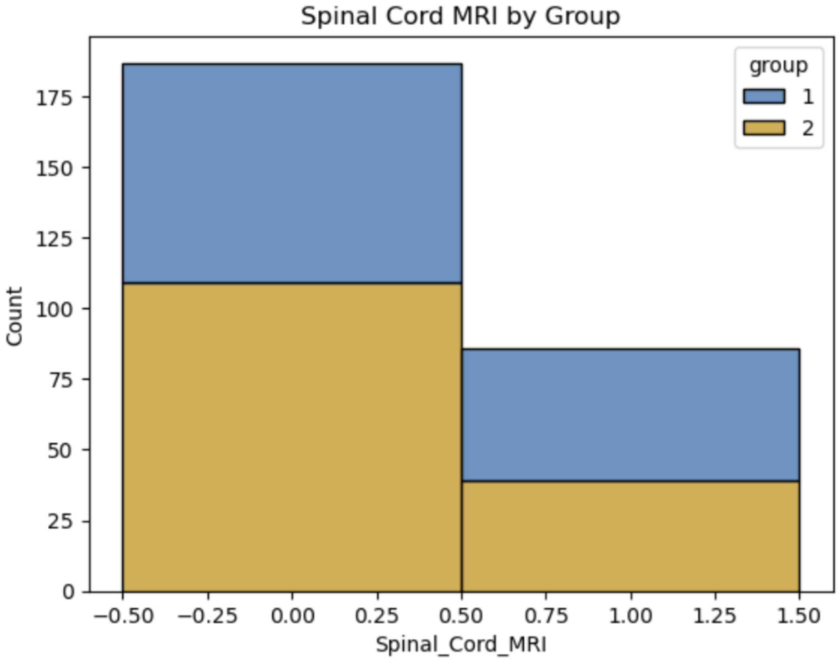


Fig. 8. Spinal Cord MRI by Group.

5.5 Results

After the exploratory analysis was preformed it was assumed that Spinal Cord MRI, Initial EDSS, and Oligoclonal Bands were likely the most significant predictors of MS progression. While both models performed well, KNN outperformed SVM in terms of both accuracy and ROC-AUC scores.

SVM Model Results Table 4 summarizes the performance metrics of the SVM models, both with and without hyperparameter tuning combinations. Models were then evaluated using test/train accuracy, F1-score, and ROC-AUC on both the training and test sets. The highest-performing highest-performing model was observed with C = 100and gamma = 1, which achieved perfect scores across all the metrics. This kind of performance brings up concerns of overfitting. The model with C=0.1 and gamma =1.5 achieved a balanced performance with a test accuracy of 82%, F1-score of 0.82, and ROC-AUC of 0.94.

Table 4. SVM Performance Metrics for Various Hyperparameter Combinations.

C	Gamma	Train Acc	Test Acc	F1-Score	ROC-AUC
100	1.0	1.00	1.00	1.00	1.00
10	1.0	0.98	0.95	0.93	0.96
0.1	0.1	0.89	0.85	0.85	0.94
0.15	0.15	0.84	0.82	0.82	0.94
0.2	0.15	0.84	0.82	0.82	0.94

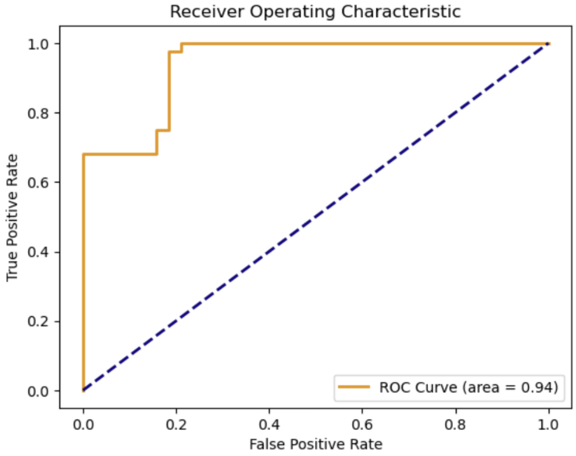


Fig. 9. ROC Curve for SVM model with $C = 0.1$ and $textttgamma = 0.1$, achieving an AUC of 0.94. The curve demonstrates the model's strong ability to distinguish between CDMS and non-CDMS classes.

k-NN Results The `k-NN` classification algorithm was evaluated using various values for `n_neighbors` to determine the impact on model performance. As seen in Table 5, the configuration with `n_neighbors` = 1 achieved perfect metrics across the board, just like the base `SVM` model. As with the perfect `SVM` model, these outcomes indicate overfitting, which may lead to limited generalization for unseen patient data. On the other hand, when `n_neighbors` was set to 5, the model performed with a strong balance between accuracy and generalization, achieving a test accuracy of 98%, an F-1 score of 0.98, and a ROC-AUC of 0.97.

Table 5. k-NN Performance Metrics for Various Values of *n_neighbors*.

n_neighbors	Train Acc	Test Acc	F1-Score	ROC-AUC
1	1.00	1.00	1.00	1.00
5	0.99	0.98	0.98	0.998
10	0.93	0.91	0.91	0.990
15	0.91	0.89	0.89	0.974
42	0.81	0.77	0.77	0.923

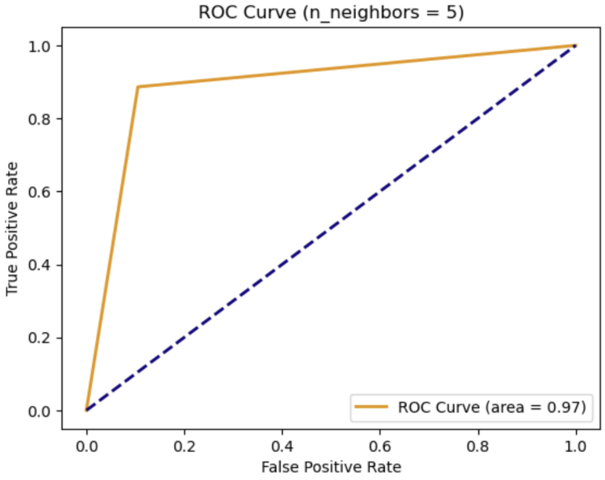


Fig. 10. ROC Curve for k-NN model with $n\_neighbors = 5$, achieving an AUC of 0.97. The curve demonstrates the model's strong ability to distinguish between CDMS and non-CDMS classes, with high sensitivity and specificity across thresholds.

5.6 Model Result Comparison

The results of the SVM and k-NN models demonstrated high accuracy and ROC-AUC scores in delivering predictive solid performances. The SVM model was the best-balanced model, achieving a test accuracy of 82% and a ROC-AUC of 0.94 under configurations C=0.1 and gamma=0.1. Likewise, the k-NN model with n=5 achieved a test accuracy of 98% and a ROC-AUC of 0.97, showing generalizability. While both models performed successfully, the k-NN yielded better accuracy than the SVM, making it an appropriate fit for this dataset. However, concerns about overfitting under specific configurations have already been raised, particularly for models achieving perfect training and test set metrics.

6 Conclusions

6.1 Can data mining techniques identify the relationship between clinical features and the progression of MS, enhancing traditional statistical methods

This project highlighted how data mining techniques can reveal significant relationships between clinical and MRI features and the transition from CIS to CDMS. Initial EDSS, Oligoclonal Bands, and Spinal Cord MRI findings were the most forceful predictors. They matched standardized medical knowledge. The advanced algorithms and pairwise visualizations allowed these investigations to illustrate those distinct patterns between the CIS and CDMS groups. These findings suggest that data mining techniques in such a case could augment conventional statistical approaches by providing information about the interaction between the features with features and the probability of progression.

6.2 What features are the best predictors of the progression of MS (CIS to CDMS)?

Feature selection and model evaluation identified the most predictive features as Initial EDSS, Oligoclonal Bands, and MRI findings (Spinal Cord, Periventricular, and Infratentorial). These features exhibited strong separations between CDMS and non-CDMS groups in the exploratory data analysis. They significantly contributed to the predictive performance of both the `SVM` and `k-NN` models. This highlights the importance of combining clinical expertise with data-driven methods to prioritize features that are both interpretable and effective for predictive modeling.

6.3 Do algorithms like k-NN and SVM compare to traditional statistical models in predicting the progression of MS?

The `SVM` and `k-NN` models achieved a very high accuracy and robustness in predicting MS progression. The best SVM model is a well-balanced model, whereby its performance on a test set gives an accuracy of 82% and a ROC-AUC of 0.94, representing good generalization. The same is true for `k-NN` with `n_neighbors` =5. It had a very high accuracy value on the test set of 98% and an ROC-AUC equal to 0.97, showing a balance between accuracy and good generalization. These findings underscore that as long as they are adequately tuned and guided with good data preprocessing, modern machine learning algorithms can sometimes match or even outstrip the performance of classical statistical models.

6.4 Project Limitations

- The dataset was limited to a single population study on a single center in Mexico. This may, however, restrict the generalization within other population groups concerning ethnicity or healthcare systems.

- Specific models were overfitted, creating perfect metrics for the train and test sets. Nevertheless, such performance is important for the model's ability to generalize to previously unseen data.
- Missed values were handled with simple imputation techniques, such as mode and median imputation, which might not reflect the true variation in the data. More advanced imputation techniques can help heighten the robustness of the entire analysis.
- The project focused on `SVM` and `k-NN` algorithms, and even though they may work well, they might fail to capture more complex or nonlinear relationships than ensemble methods and deep learning models.
- Evaluation metrics such as accuracy and ROC-AUC were applied to assess performance but may not adequately account for class imbalance effects.

6.5 Future Work

The project described the utility of modern machine learning tools and data mining techniques in conjunction with the traditional approaches to the diagnosis of MS. These findings reinforce the importance of utilizing features such as Initial EDSS, the type of MRI abnormalities, and whether or not Oligoclonal Bands are present for accurate prediction. The methodologies can be applied in real-world settings to assist in interventions at the onset of a disease process, especially when resources are limited.

Future work may extend to adding new features, such as the use of immunological markers or longitudinal MRI data, to further enhance model performance. External validation on different datasets and applying ensembles shall shed light on the generalizability of the findings.

6.6 Closing Reflection

The author of this project is an MS patient themselves and found this research to be particularly important as a patient whose diagnosis came early. In retrospect, they found that some physicians ignored many initial symptoms found in the features of this dataset until an MRI was actually performed. Being able to put symptoms together and have software based on these data mining and classification techniques might aid someone like themselves in the future and catch MS even earlier.

References

1. Chavarria, V., Espinosa-Ramírez, G., Sotelo, J., Flores-Rivera, J., Anguiano, O., Hernández, A., Guzmán-Ríos, E., Salazar, A., Ordoñez, G., Pineda, B.: Conversion predictors of clinically isolated syndrome to multiple sclerosis in mexican patients: A prospective study. Archives of Medical Research **54**(5), 102843 (Jul 2023). https://doi.org/10.1016/j.arcmed.2023.102843

2. Gebretsadik, D.: Conversion predictors of cis to multiple sclerosis (2023), https://www.kaggle.com/datasets/desalegngeb/conversion-predictors-of-cis-to-multiple-sclerosis, accessed: 2024-10-31
3. Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques. Morgan Kaufmann, Burlington, MA, USA, 3rd edn. (2011)
4. McDonald, W.I., Compston, A., Edan, G., Goodkin, D., Hartung, H.P., Lublin, F.D., McFarland, H.F., Paty, D.W., Polman, C.H., Reingold, S.C., Sandberg-Wollheim, M., Sibley, W., Thompson, A., Van Den Noort, S., Weinshenker, B.Y., Wolinsky, J.S.: Recommended diagnostic criteria for multiple sclerosis: Guidelines from the international panel on the diagnosis of multiple sclerosis. Annals of Neurology **50**(1), 121–127 (2001). https://doi.org/https://doi.org/10.1002/ana.1032, https://onlinelibrary.wiley.com/doi/abs/10.1002/ana.1032
5. McGinley, M.P., Goldschmidt, C.H., Rae-Grant, A.D.: Diagnosis and treatment of multiple sclerosis: A review. JAMA **325**(8), 765–779 (2021). https://doi.org/10.1001/jama.2020.26858
6. MedlinePlus: Oligoclonal bands. https://medlineplus.gov/ency/article/003631.htm (2024), accessed: 24 November 2024
7. Miller, D., Barkhof, F., Montalban, X., Thompson, A., Filippi, M.: Clinically isolated syndromes suggestive of multiple sclerosis, part i: natural history, pathogenesis, diagnosis, and prognosis. The Lancet Neurology **4**(5), 281–288 (2005). https://doi.org/https://doi.org/10.1016/S1474-4422(05)70071-5, https://www.sciencedirect.com/science/article/pii/S1474442205700715
8. MS Society: Expanded disability status scale (2024), https://www.mssociety.org.uk/living-with-ms/treatments-and-therapies/getting-treatment-for-ms/expanded-disability-status-scale, accessed: 2024-11-23
9. MS Trust: Poser criteria (nd), https://mstrust.org.uk/a-z/poser-criteria, accessed: November 24, 2024
10. Pérez, C.A., Perez, C.A., Smith, A., Nelson, F.: Multiple Sclerosis Phenotypes, pp. 37–44. Cambridge Manuals in Neurology, Cambridge University Press (2021)
11. Unknown, A.: Mcdonald criteria for ms diagnosis: 2020 update and impact. International Journal of Multiple Sclerosis **18**, 2061–2070 (2020)
12. Weinstock-Guttman, B., Sormani, M.P., Repovic, P., et al.: Predicting long-term disability in multiple sclerosis: A narrative review of current evidence and future directions. International Journal of MS Care **24**(4), 181–190 (2022), https://digitalcommons.psjhealth.org/publications/6467/