Classifying Multiple Sclerosis Disease with Statistical Learning Methods

Authors: Jeffrey Kong, Noah Lee, Jacob Ventura

Introduction:

Statistical learning methods are utilized extensively in the healthcare industry to analyze complex datasets and make informed predictions that could identify at-risk patients of a given disease. Here, we aim to test multiple supervised learning strategies to correctly classify patients diagnosed with Multiple Sclerosis disease after going through clinically isolated syndrome (CIS), chosen with optimal accuracy and interpretability. Multiple Sclerosis (MS) is a chronic disease affecting an individual's central nervous system (Johns Hopkins Medicine, n.d.). During MS, the immune system mistakenly attacks its healthy cells, resulting in blurred vision, color distortion, trouble walking, speech problems, numbness, and other varying symptoms.

MS has affected around 1 million people in the United States in 2022 and about 2.8 million globally in 2020, with the majority in the 20-50 age group. The cause of MS is currently unknown, but scientists believe genetics and environmental factors to be contributing influences (Kolčava et al., 2020). However, the most significant method of identifying at-risk individuals with MS is by first analyzing CIS-diagnosed patients, as MS begins with CIS in approximately 85% of patients. CIS is the first episode of neurological symptoms in MS; however, not all go on to develop MS (OhioHealth, 2017).

Early identification of CIS patients at high risk for MS conversion is crucial, as it informs treatment decisions and long-term care strategies (Kolčava et al., 2020). Early identification is essential given the chronic and debilitating nature of MS, which can severely impact physical, mental, and neurological functions. The primary objective of this study is to utilize machine learning methods to identify the most significant predictors of conversion from CIS to MS in a dataset of Mexican mestizo patients. By addressing this gap, we aim to enhance predictive models, improve diagnostic accuracy, and support clinicians in making informed decisions about patient management and therapy initiation.

Data:

Researchers funded by Consejo Nacional de Ciencia y Tecnología (CONACYT) collected from a study conducted at the National Institute of Neurology in Mexico. The study focused on the progression of Clinically Isolated Syndrome (CIS) to Clinically Definitive Multiple Sclerosis (CDMS) Disease. The data contains information on Mexican Mestizo Patients who were diagnosed with CIS between 2006 and 2010. The data contains information on Mexican Mestizo Patients who were diagnosed with CIS between 2006 and 2010. The researchers followed up with the patients after 10 years to collect the data in this dataset.

The Multiple Sclerosis dataset contains 273 rows and 19 different features. Some predictors included:

- **Age:** Age of the patient in years (Numeric)
- **Schooling:** The number of years the patient spent in school in years (Numeric)
- **Gender:** 1 = male, 2 = female, (Categorical)
- **Breastfeeding:** 1 = Yes, 2 = No, 3 = Unknown, (Categorical)
- **Varicella:** 1 = Positive, 2 = Negative, 3 = Unknown, (Categorical)
- **Initial_Symptoms:** 1 = Visual, 2 = Sensory, 3 = Motor, 4 = Other, … - See data sheet (Categorical)
- **Mono _or_Polysymptomatic:** 1 = Monosymptomatic, 2 = Polysymptomatic, 3 = Unknown, (Categorical)
- **Oligoclonal_Bands:** 0 = Negative, 1 = Positive, 2 = Unknown, (Categorical)
- **LLSSEP:** 0 = Negative, 1 = Positive, (Categorical)
- **ULSSEP:** 0 = Negative, 1 = Positive, (Categorical)
- **VEP:** 0 = Negative, 1 = Positive, (Categorical)
- **BAEP:** 0 = Negative, 1 = Positive, (Categorical)
- **Periventricular_MRI:** 0 = Negative, 1 = Positive, (Categorical)
- **Cortical_MRI:** 0 = Negative, 1 = Positive, (Categorical)
- **Initial_EDSS:** 0 = Negative, 1 = Positive, (Categorical)
- **Final_EDSS:** 0 = Negative, 1 = Positive, (Categorical)
- **Infratentorial_MRI:** 0 = Negative, 1 = Positive, (Categorical)
- **Spinal_Cord_MRI:** 0 = Negative, 1 = Positive, (Categorical)
- **group:** 1 = CDMS, 2 = Non-CDMS, (Categorical)

The response variable we decided to use for our model building and analysis was the **group** variable, which was a categorical factor variable that indicated whether or not a patient converted to Clinically Definite Multiple Sclerosis (CDMS) or remained Non-CDMS. If **group** was equal to 1 in the dataset, a patient had successfully converted to CDMS. If **group** was equal to 0 in the dataset, a patient failed to convert to CDMS successfully.

We removed **Initial_EDSS** and **Final_EDSS** from the dataset due to a high proportion of missing values and the fact that they seemed to be missing only when CDMS = 2. We replaced missing values with the mode (most frequent value) for other variables to preserve categorical consistency. This approach prevents the creation of additional categories, such as "unknown", which could negatively impact classifier performance. Given the small size of our dataset, retaining as much data as possible is critical for building a more accurate model. For instance, variables like **Breastfeeding** have 25% missing values, and removing these records would further reduce the dataset, potentially worsening model reliability. Replacing missing values with the mode minimizes data loss while ensuring the dataset remains interpretable for techniques like PCA and machine learning models.

A feature of the dataset that caused us some trouble was that if a predictor were categorical, between “1” and “2”, the data would replace missing or null values with a “3”. We removed these values and replaced them with the variable's mode to prevent future models from classifying any values as a “3” when they are not an actual class. We converted all variables into factors before proceeding with our analysis.

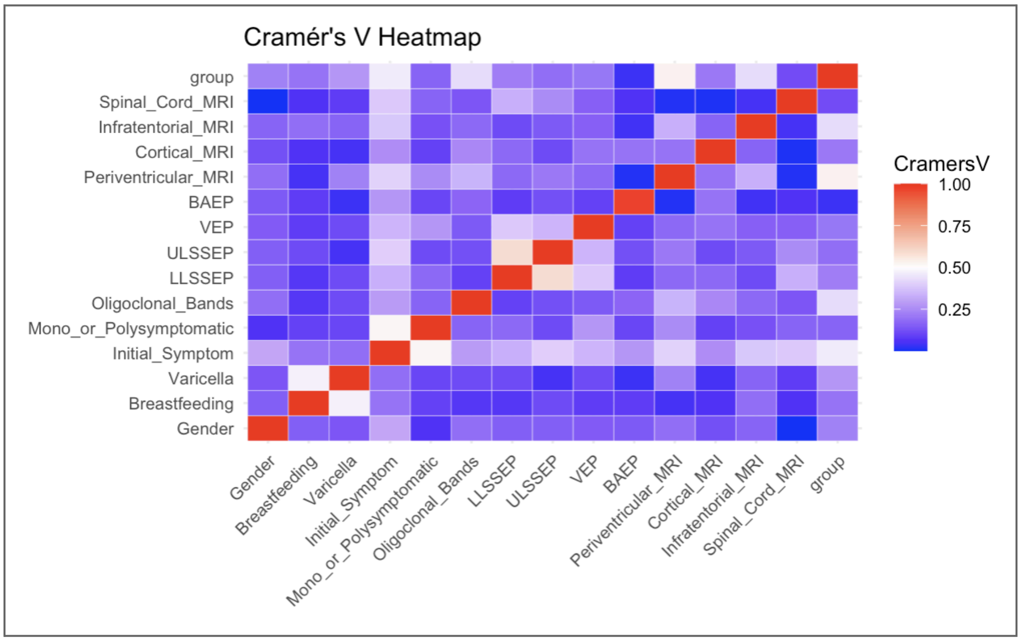


Figure 1 - Cramer's V Heat Map displaying the Association Between Categorical Variables. A red square indicates a high correlation between the two categorical variables, and a blue square indicates a low correlation. Gray squares indicate many missing values.

Figure 1 displays a Cramer’s V Heatmap and shows the strength of association between pairs of categorical variables from our dataset. We can see a high correlation between **ULSSEP** and **LLSSEP** due to their shared focus on sensory-evoked potentials. This may indicate that both of these variables could be helpful in predicting outcomes, but including both in our model may not be necessary. We can also see that certain variables, such as the MRI variable, have a low correlation with **group**, showing that they could be interesting predictors. This is expected as those variables are also listed in the McDonald criteria (Multiple Sclerosis Trust, 2022). We also see that many variables, such as **Spinal_Cord_MRI,** show low association with many of the other predictor variables, meaning they might not be strongly connected or related to most of the other factors in the dataset.

To further look into the associations between variables, we decided to proceed with Principle Component Analysis (PCA) to see if we could identify underlying patterns and reduce the dimensionality of the data.

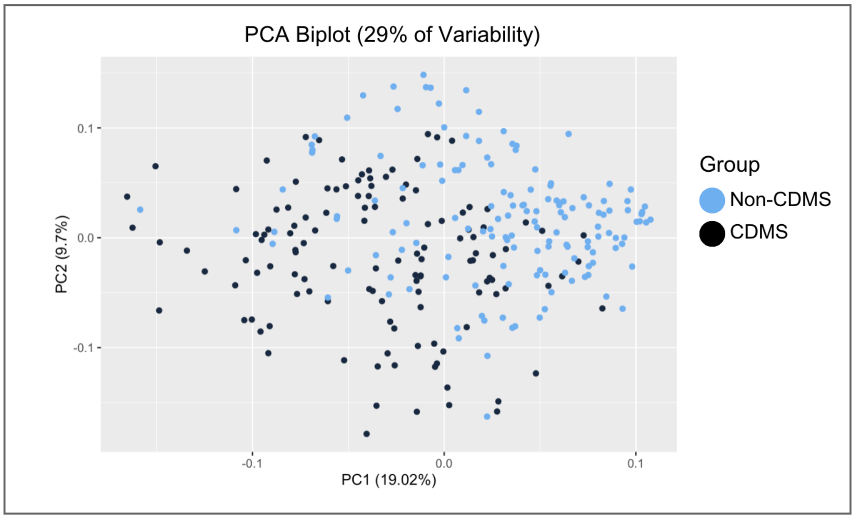


Figure 2 - PCA Biplot showing the Separation of CDMS and Non-CDMS groups (29% of Total Variability Explained)

Figure 2 shows the result of PCA to find patterns and create a lower-dimensional plot. Our primary objective with PCA was to determine whether the factors in our dataset are predictive of CDMS. **Figure 2** shows a slight difference as the Non-CDMS points have larger Principle Component One (PC1) and Principle Component Two (PC2) values. However, the separation only partially separates the groups. In addition, PCA is not typically used with both qualitative and quantitative variables, which may have impacted the strength of the observations.

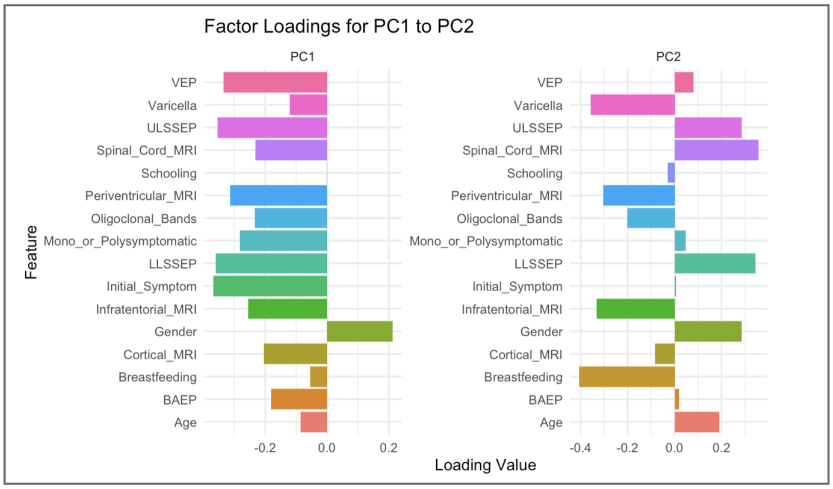


Figure 3 - Plot of Factor Loadings for PC1 and PC2

Looking at the factor loadings, which identify the most important factors that result in the biggest change in PC1 and PC2 (**Figure 3**), we can also see that PC1 is influenced by clinical and

MRI-based features, capturing variability related to lesion distribution and immune abnormalities. We can also see that higher PC1 and PC2 are linked to non-CDMS, meaning that the variables with a larger presence on **Figure 3** could indicate that they help separate PC1 and PC2 and, thus, CDMS from non-CDMS.

The observation of the data above highlights that many factors influence CDMS in the dataset. EDA suggests using a non-linear model approach when fitting a predictive model to the data. Lastly, we randomly split the data into 70% training, 15% validation, and 15% testing, stratifying by **group**. With the data cleaned and a good understanding of the data, we moved forward with building a predictive model.

Models:

We evaluated several supervised learning models to classify patients transitioning from Clinically Isolated Syndrome (CIS) to Clinically Definite Multiple Sclerosis (CDMS), focusing on their accuracy, sensitivity, and specificity. The models included K-Nearest Neighbors (KNN), Naive Bayes, and Random Forest.

We started by considering a KNN classification model, which makes classifications based on how ‘close’ patients are to other labeled patients. We started with KNN due to its simplicity and ability to handle non-linear relationships. We chose which predictors to consider by examining the variables that impacted PC1 and PC2 in Principle Component Analysis. We then tuned for k using 10-fold cross-validation. We selected seven nearest neighbors (k = 7) to reduce the chance of overfitting with higher values of k. However, the cross-validated KNN model recorded a 68% accuracy on the validation dataset and a sensitivity of 52%, indicating that the model struggled to identify patients in the CDMS group. Despite its simplicity, KNN struggled to handle the high dimensional data effectively, so we moved forward with different models.

Given that our dataset consisted of quantitative and qualitative variables, we considered a Naive Bayes approach. Naive Bayes assumes feature independence and uses Gaussian density functions to classify patients. Despite its strong assumptions, the Naive Bayes model performed better on the validation set than KNN. The model achieved an accuracy of 80% and a sensitivity of 74%, indicating that the model was identifying fewer false positives and correctly identifying CDMS at a higher rate. The Naive Bayes model results were promising, but didn’t offer the interpretability of the predictors that would aid in investigating MS further in the future. Thus, we decided to investigate a more interpretable model.

Balancing performance and interpretability, we implemented a Random Forest Classifier. Random Forest is an ensemble method using multiple trees and considering a subset of the predictors to average a final prediction. Using grid search (t = [20, 80, 120, 160, 200], m = [3,4,5]) and 10-fold cross-validation, we optimized the number of trees and predictors per split at

160 trees and 4 predictors per split (t = 160, m = 4) . Random Forest achieved a validation accuracy of 78%, comparable to Naive Bayes, and a sensitivity of 79%. While its sensitivity and specificity metrics were slightly worse than those of Naive Bayes, Random Forest offered the advantage of interpretability through feature importance rankings.

		KNN	Naive Bayes	Random Forest
Metric	Accuracy	0.76	**0.80**	0.78
	Sensitivity (True positive rate)	0.68	0.74	**0.79**
	Specificity (True negative rate)	0.82	**0.86**	0.77
	ROC-AUC	0.85	**0.87**	**0.87**

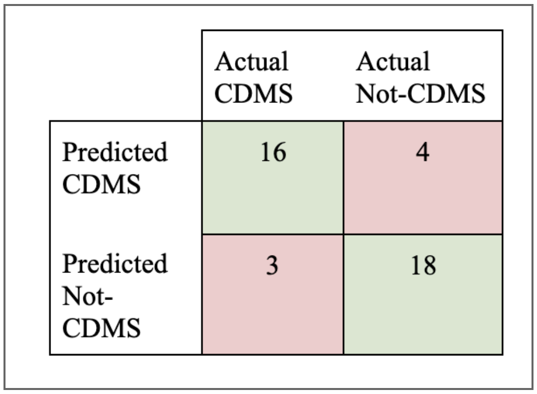
Table 1 - Table of Model Performance Metrics on Validation Set

Table 1 summarizes the model performance metrics and shows that Random Forest and Naive Bayes were close and scored the highest in most metrics. It should also be noted that the validation set was small (41 observations), so the differences in performance between the two models amounted to one more true positive and two less true negative classifications for Random Forest compared to Naive Bayes. These differences were hard to gauge given the small size of the validation set.

Given its balance of accuracy and interpretability, we selected the Random Forest model as the final method for classifying CDMS.

<u>Results:</u>

We evaluated the Random Forest model on the test set and obtained an accuracy of 83%, specificity of 82%, and sensitivity of 84%. These results improve upon the validation set metrics and provide evidence that this model generalizes well and has no evidence of overfitting. These results demonstrate the model's ability to correctly classify CDMS patients, as shown by sensitivity, while minimizing false positives, as shown by specificity. In contrast to the validation set, the model had higher metrics than Naive Bayes. **Figure 4** displays the confusion matrix of the Random Forest model on the test set. We observed that the model was roughly balanced between sensitivity and specificity, so it was not too conservative when making predictions.



	Actual CDMS	Actual Not-CDMS
Predicted CDMS	16	4
Predicted Not-CDMS	3	18

Figure 4 - Confusion Matrix of Random Forest Model Predictions on Testing Data Set

Using the vip package in R, which records the total increase in the node purity due to each split by a predictor during training, we were able to find variables that were most important in classifying CDMS (**Figure 5**):

1. **Periventricular_MRI:** Strongly associated with CDMS conversion, aligning with existing medical research.
2. **Infratentorial_MRI:** Indicates lesions in the brainstem or cerebellum, which is crucial for MS diagnosis.
3. **Oglioclonal_Bands:** These variables represent the presence (X1) and absence (X0) of oligoclonal bands in cerebrospinal fluid, which are important biomarkers for multiple sclerosis (Multiple Sclerosis Trust, 2022).
4. **Initial_Symptoms (X1):** Highlights the role of clinical presentation in predicting MS conversion, as initial symptom type often correlates with lesion location and disease progression.
5. **Schooling:** Patients with higher levels of schooling may have better access to healthcare, earlier diagnosis, or better management of symptoms, which could affect the observed conversion rates from CIS to CDMS.

These features emphasized the importance of neurological imaging and auditory assessments in predicting MS progression, aligning with current clinical practices.

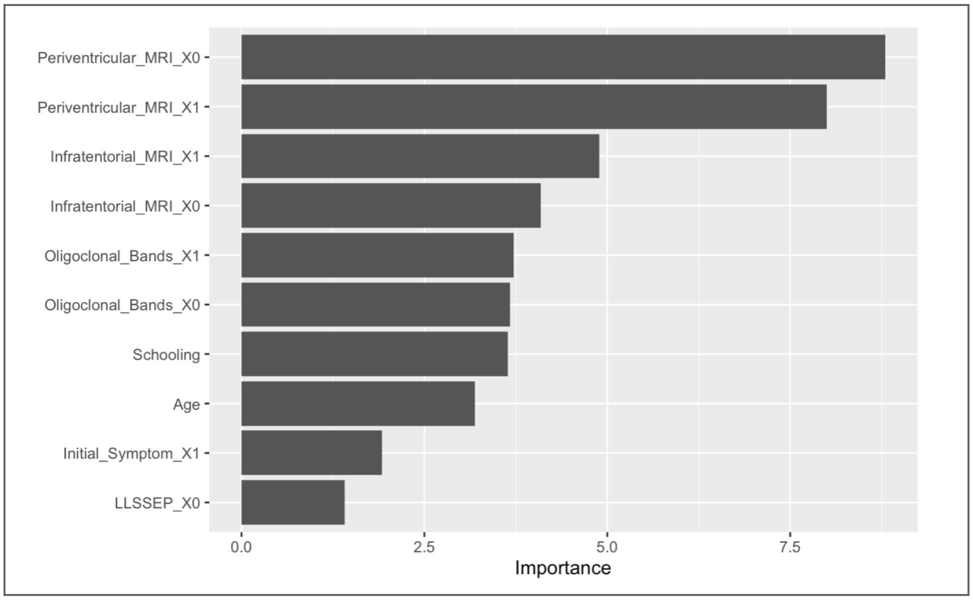


Figure 5 - Feature Importance Plot Using the vip Package on our Random Forest Model

Analyzing misclassified cases, we see that misclassifications for **group** = 1 (CDMS) often occur when MRI or diagnostic markers (e.g. Periventricular_MRI, Oligoclonal_Bands) are less prominent. In contrast, misclassifications for **group** = 2 (Non-CDMS) involve cases with mixed or uncertain features. In addition, **Initial_Symptom** = 4 and 8 (other and sensory/motor) appeared in almost all misclassified records despite only representing 25% of the dataset. The over-representation of **Initial_Symptom** = 4 and 8 in misclassified cases suggests the model may need to capture these symptom categories adequately or may represent more complex cases with overlapping characteristics. Further investigation as to what exactly **Initial_Sympyom** = 'other' meant could help clarify what went wrong with the model. All misclassified records had relatively close prediction probabilities at around 0.5, but one notable record predicted non-CDMS with a probability of 0.89. Notably, the record was positive for all MRI predictors and had **Initial_symptom** = 12 (visual, sensory). This misclassification suggests that this may have been an edge instance and that our model heavily weighted MRI predictors compared to other predictors.

Our final model faces limitations from the over-reliance on specific features such as MRI predictors and struggles with more ambiguous data like **Initial_symptom** = 4 and 8, which appear frequently in the misclassified observations. Additionally, the dataset's focus on Mexican mestizo patients limits this study's generalizability. Furthermore, missing potential predictors from other studies, like advanced imaging from the McDonald Criteria (Multiple Sclerosis Trust, 2022), could further restrict the model's accuracy. Another aspect to note was the close prediction probabilities in a few of the predictions, which may signal the model's weakness in confidently identifying CDMS despite the strong classification metric results.

To address these issues, future improvements could include feature engineering to capture latent interactions between variables, such as creating an MRI feature. We could also expand the dataset to include a more diverse group of patients and remove ambiguous feature definitions like **Initial_Symptom** = 'other'. Leveraging other classification models like stochastic gradient descent or looking into imaging with deep learning models could yield more accurate models and new interpretations on notable predictors (Scikit Learn, 2024). In addition, the dataset was relatively small, with 271 observations after preprocessing. This could bring up issues as some predictors, such as **Breastfeeding,** were missing a quarter of its values. Sparse data made it difficult for the model to find generalizable patterns and make more confident predictions. Increased data collection on CDSS from other hospital centers would help with generalizability and provide more data points to train a future model.

References:

Johns Hopkins Medicine (n.d.). *Multiple Sclerosis (MS)*. [online] John Hopkins Medicine. Available at: https://www.hopkinsmedicine.org/health/conditions-and-diseases/multiple-sclerosis-ms.

OhioHealth (2017). *High-Fat Diets Could Pose Danger to Young MS Patients*. [online] High-Fat Diets Could Pose Danger to Young MS Patients . Available at: https://newsroom.ohiohealth.com/high-fat-diets-could-pose-danger-to-young-ms-patients/ [Accessed 24 Nov. 2024].

Kolčava, J., Kočica, J., Hulová, M., Dušek, L., Horáková, M., Keřkovský, M., Stulík, J., Dostál, M., Kuhn, M., Vlčková, E., Bednařík, J. and Benešová, Y. (2020). Conversion of clinically isolated syndrome to multiple sclerosis: a prospective study. *Multiple Sclerosis and Related Disorders*, 44, p.102262. doi:https://doi.org/10.1016/j.msard.2020.102262.

Multiple Sclerosis Trust (2022). *McDonald criteria*. [online] MS Trust. Available at: https://mstrust.org.uk/a-z/mcdonald-criteria.

Scikit Learn (2024). *12. Choosing the right estimator*. [online] scikit-learn. Available at: https://scikit-learn.org/stable/machine_learning_map.html.

Appendix:

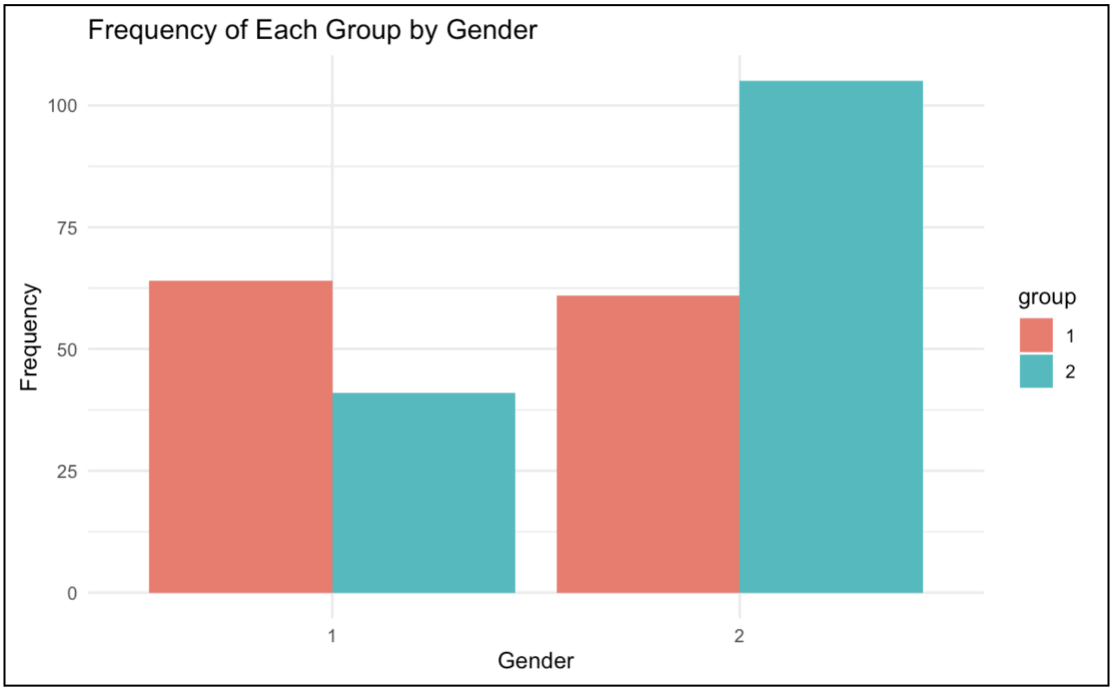
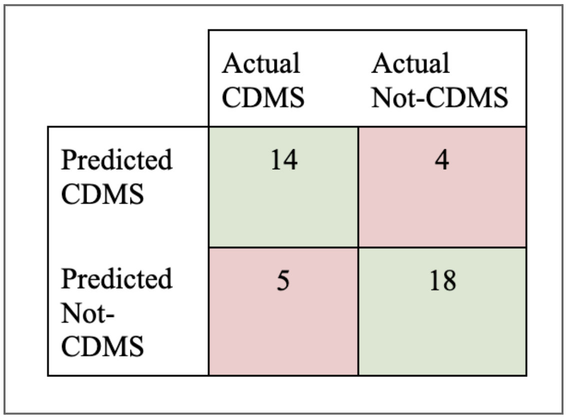
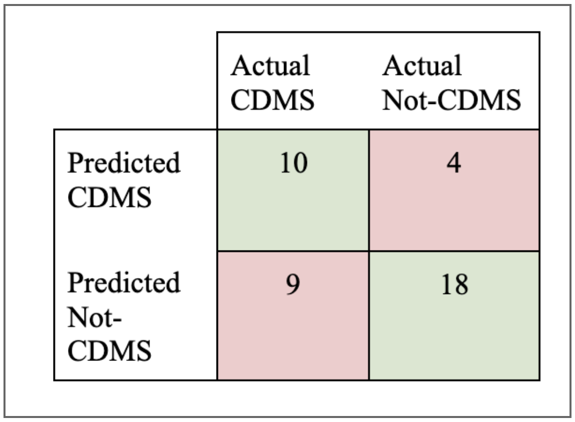


Figure 6: Frequency of CDMS by Gender – Group: 1 = CDMS (125), 2 = non-CDMS (146) Gender: 1 = Male (105), 2 = Female (166). Observation: More female participants tested negative for CDMS despite prior studies indicating that CDMS occurs twice as often in women than in men.



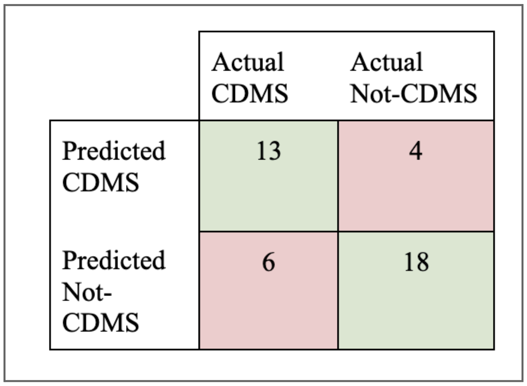
	Actual CDMS	Actual Not-CDMS
Predicted CDMS	14	4
Predicted Not-CDMS	5	18

Figure 7: Logistic Regression Confusion Matrix on the Validation set



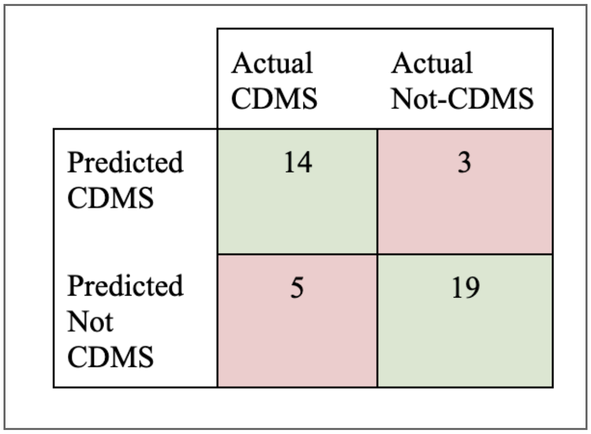
	Actual CDMS	Actual Not-CDMS
Predicted CDMS	10	4
Predicted Not-CDMS	9	18

Figure 8: KNN Confusion Matrix on the Validation Set



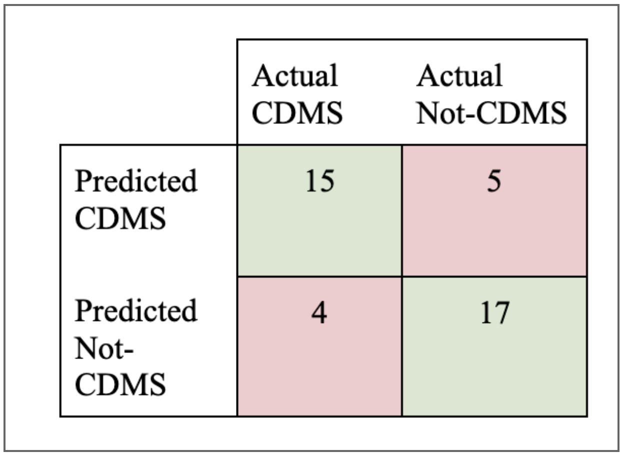
	Actual CDMS	Actual Not-CDMS
Predicted CDMS	13	4
Predicted Not-CDMS	6	18

Figure 9: LDA Confusion Matrix on the Validation Set



	Actual CDMS	Actual Not-CDMS
Predicted CDMS	14	3
Predicted Not CDMS	5	19

Figure 10: Naive Bayes Confusion Matrix on the Validation Set



	Actual CDMS	Actual Not-CDMS
Predicted CDMS	15	5
Predicted Not-CDMS	4	17

Figure 11: Random Forest Confusion Matrix on the Validation Set

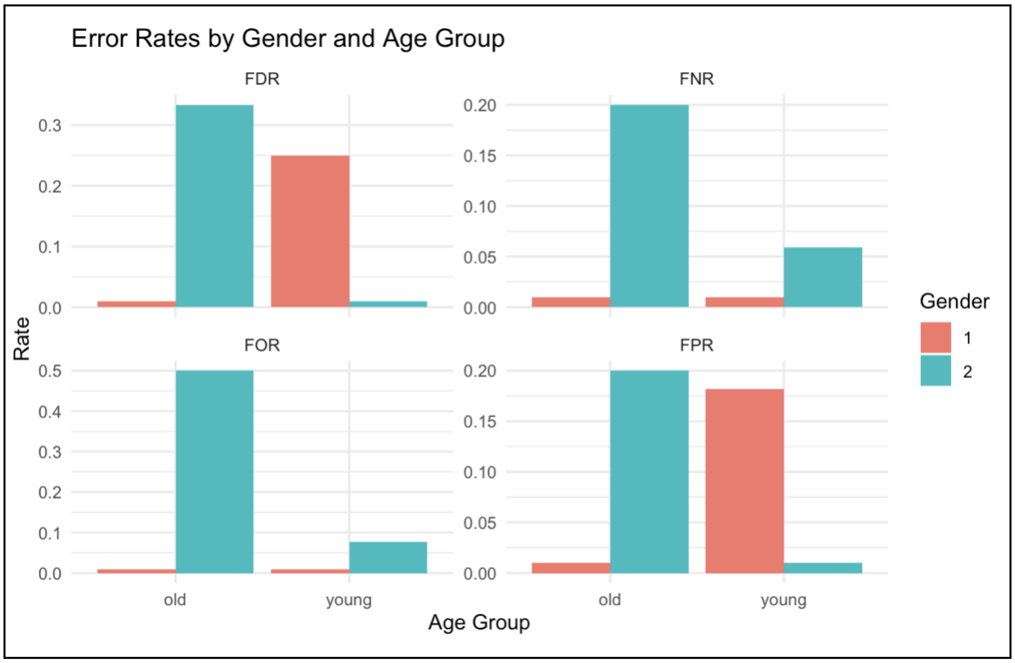


Figure 12: Error Rate Chart on Random Forest Model on Testing Set - Used in the model card. Gender = 1 represents Male, and Gender = 2 represents Female.