CSC3432
Report 2 – Biomedical ML

Introduction

In this report, I embark on a comprehensive analysis of a biomedical tabular dataset, emphasizing longitudinal data related to dementia and Alzheimer's disease. This dataset, rich in complexity, chronicles patient histories across multiple visits and incorporates diverse features, including demographic, socio-economic, and medical variables.

I primarily aim to utilize machine learning techniques—specifically, Random Forests and Support Vector Classifiers—to classify patients' dementia status based on the dataset's features. A crucial element of our investigation is identifying the most influential features within these machine learning models that determine dementia classification. Moreover, I intend to delve into the characteristics of patient groups, seeking insights into trends in dementia status. This endeavour is an exercise in data analysis and a step toward a deeper understanding of dementia and Alzheimer's disease, potentially contributing to improved patient care and outcomes.

Exploratory Data Analysis of Dementia Patient Records

1. Introduction to EDA

The initial phase of my investigation entailed a rigorous examination of the dementia patient records dataset. This EDA is crucial for comprehending the data's structure and uncovering insights pivotal for dementia research. By dissecting the dataset's nuances, I aimed to lay a solid groundwork for subsequent machine learning modelling, directly linking EDA findings to predictive analysis.

2. Methodological Approach

I employed statistical techniques and visual tools to delve into the dataset. This involved assessing distributions of critical variables, exploring correlations, and understanding patterns and outliers. Each step in this process was geared towards unveiling insights relevant to dementia and Alzheimer's, informing my model selection and feature engineering strategy.

3. Results and Interpretations

3.1 Correlation Matrix

A detailed examination of correlations revealed intriguing associations. For instance, the negative correlation between Socioeconomic Status (SES) and Years of Education (YOE) may indicate socioeconomic determinants in dementia progression, an aspect worth exploring in predictive modelling. Similarly, the relationship between eTIV and nWBV could suggest physiological markers pertinent to dementia research. The minimal correlation of Clinical Dementia Rating (CDR) with numerical variables underscores the complexity of dementia, hinting at the need for sophisticated modelling techniques.

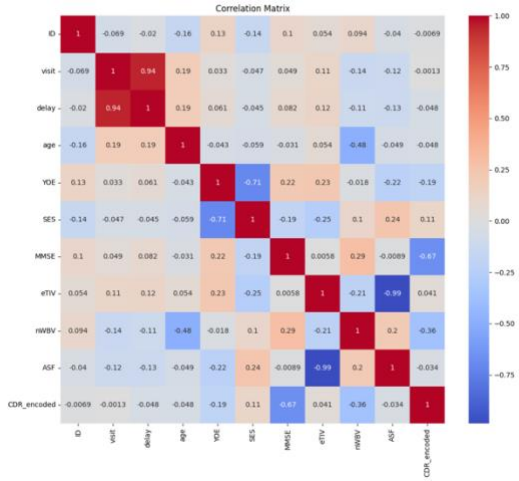

Figure 1: Correlation Matrix illustrates the pairwise correlations between variables. Shades of blue represent positive correlations, whereas shades of red indicate negative correlations. The colour intensity corresponds to the strength of the correlation.

3.2 Distribution of Clinical Dementia Rating (CDR)

The distribution highlighted a significant imbalance, with a predominance of 'none' dementia cases. This imbalance poses a challenge for predictive modelling, as it can bias the model towards the majority class. It necessitates strategies like SMOTE for oversampling or using performance metrics like the F1-score that are less sensitive to class imbalance.

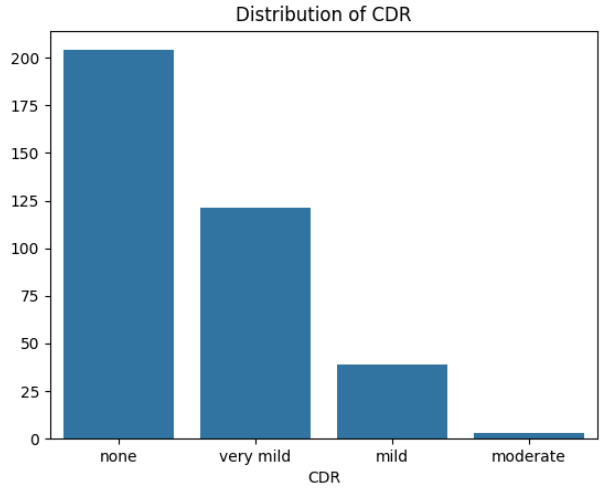

Figure 2: Distribution of CDR. A bar chart showing the frequency of each CDR category, revealing the prevalence of the 'none' category and the rarity of 'moderate' dementia within the dataset.

3.3 Distribution of Handedness and Distribution of Sex

The overwhelming majority of right-handed patients might indicate limited variability and predictive value for this attribute. However, the balanced gender distribution is beneficial for mitigating gender bias, ensuring that the findings represent both sexes in dementia.

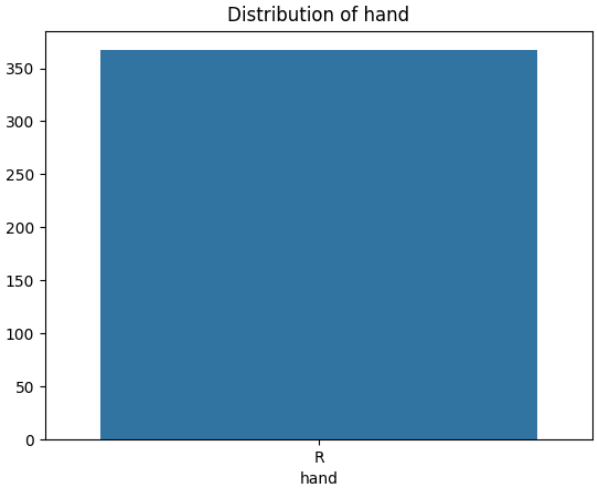

Figure 3: Distribution of Hand. The bar chart demonstrates the frequency of left-handed versus right-handed patients, with right-handed patients significantly outnumbering their left-handed counterparts, which are none in this dataset.

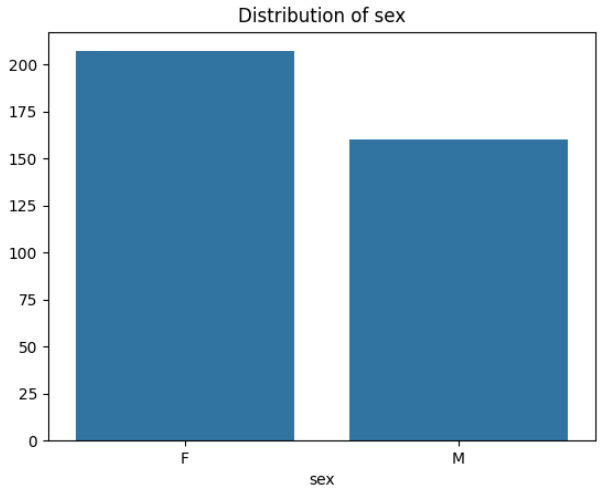

Figure 4: Distribution of Sex. This bar chart depicts a balanced distribution between female and male patients, a desirable attribute for modelling to ensure gender-neutral predictions.

3.4 Distribution of MRI_ID

The uniformity across unique MRI identifiers confirms the non-informative nature of this variable, justifying its exclusion from predictive modelling.

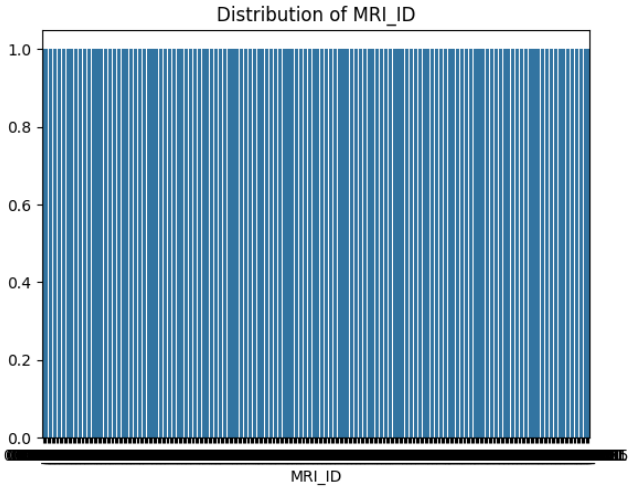

Figure 5: Distribution of MRI_ID. A uniform distribution across unique MRI identifiers emphasizes the non-informative nature of this variable for the machine learning analysis.

4. Discussion of EDA Findings

The EDA phase was instrumental in highlighting critical aspects of the dataset relevant to dementia research. The interplay between SES, YOE, and brain volume measurements could offer new avenues for understanding dementia's socioeconomic and physiological dimensions. The challenge posed by a class imbalance in CDR categories was identified as a critical aspect of model building. The insights from the distributions of handedness, sex, and MRI_ID informed the feature selection and data-cleaning processes. This comprehensive analysis sets a robust foundation for the forthcoming data preprocessing and predictive modelling stages.

5. EDA Conclusion

This exploratory analysis has provided a deeper understanding of the dataset's characteristics and highlighted specific challenges and opportunities in modelling dementia-related outcomes. The insights derived will be pivotal in guiding the strategic handling of attributes during machine learning, ensuring that the models developed are nuanced and relevant to the complex nature of dementia and Alzheimer's disease.

Data Preprocessing for Dementia Patient Records

1. Introduction to Preprocessing

Data preprocessing is a pivotal step in ensuring the reliability and validity of machine learning models. This project used preprocessing to refine the dataset for a more accurate and meaningful analysis of dementia patient records.

2. Preprocessing Steps

2.1 Data Consolidation

As I chose the Tabular dataset, the multiple CSV files representing different patient visits were combined into a single dataset. This step was crucial for establishing a comprehensive dataset, allowing for an uninterrupted analysis across all visits.

2.2 Numeric Conversion and Imputation

The 'ASF' (Atlas Scaling Factor) values, recorded inconsistently, were standardized to a uniform numeric format. Additionally, missing SES values were imputed based on the mode SES value within each age group. This step ensured consistency in quantitative analysis and preserved socioeconomic context, potentially a significant factor in dementia analysis.

2.3 Missing Value Handling

Critical attributes with missing data such as 'ASF', 'eTIV', and 'MMSE' were treated by removing incomplete rows to maintain the robustness of the analysis. This approach reduced the dataset size but enhanced data quality. However, it also introduced the risk of bias, as data missingness might correlate with specific patient conditions.

2.4 Typo Correction

Standardisation of categorical values in the 'CDR' column was performed by correcting typographical errors and inconsistencies. This was vital for the reliability of the target variable, ensuring that input errors did not skew the classification models trained on this data.

3. Results of Preprocessing

3.1 CDR Category Refinement

The standardisation process led to a more accurate representation of dementia severity levels. A significant imbalance in the 'CDR' categories was noted, prompting the need for strategies to handle this imbalance in the modelling phase.

3.2 Age Feature Analysis

A statistical summary of age-related data provided insights into the patient population's age distribution. Understanding the age distribution was crucial, as age is a significant factor in dementia progression.

4. Discussion of Preprocessing Findings

The preprocessing stage effectively addressed several data quality issues, enhancing the potential predictive accuracy of the models. By aligning SES data with age groups, we retained valuable socioeconomic information, which is likely influential in predicting dementia. The treatment of the class imbalance recognised during CDR category refinement and the focus on age statistics were essential, as they directly relate to the primary objectives of understanding dementia progression.

Random Forest Model for Dementia Status Classification

1. Introduction to Random Forest Modelling

The primary goal of employing a Random Forest model was to classify patients' dementia status with high accuracy and identify significant features influencing this classification. Random Forest was chosen due to its robust handling of high-dimensional data and its inherent capability for feature importance evaluation, aligning perfectly with the dual objectives of classification and insight generation.

2. Model Training and Hyperparameter Tuning

For data preparation, irrelevant features were removed, and categorical variables were encoded, tailoring the dataset for optimal compatibility with the Random Forest model. To address the class imbalance in CDR, SMOTE was used to improve model efficiency and sensitivity towards minority classes, essential for a nuanced understanding of dementia status. Through grid search, optimal hyperparameters were identified as {'max_depth': None, 'min_samples_split': 10, 'n_estimators': 100}, chosen to capture complex data patterns effectively while preventing overfitting.

3. Model Evaluation

3.1 Classification Performance

The model achieved an accuracy of 92.07%, as illustrated in the classification report (Figure 6). This report highlights the model's balanced performance across various dementia severity stages, with high precision, recall, and f1-score metrics scores. Such balanced metrics underscore the model's diagnostic accuracy in dementia classification.

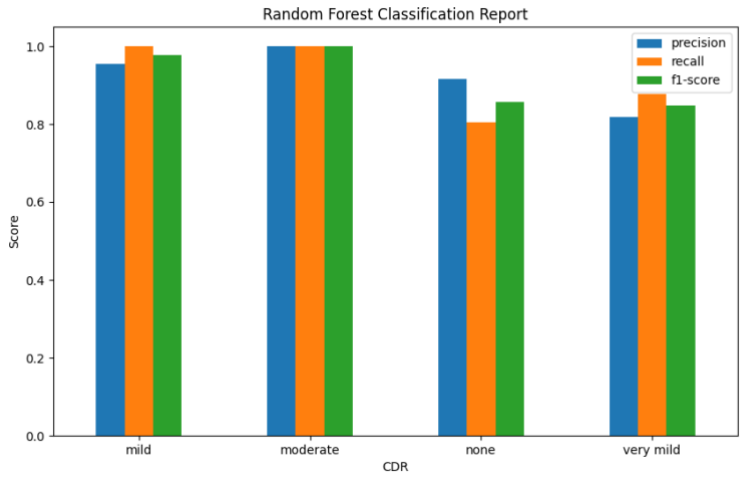

Figure 6: Random Forest Classification Report. This bar chart illustrates the model's precision, recall, and f1-score for each class of dementia severity, highlighting its diagnostic accuracy.

3.2 Feature Importance Analysis

The model's ability to evaluate feature importance directly addresses my objective to identify the most significant predictors of dementia status. The analysis shed light on which features held the most weight in the classification decision-making process, providing a pathway to understanding how different attributes contribute to dementia diagnosis.

4. Insights and Implications

The model's success in accurately classifying dementia status and revealing the importance of features provides deeper insights into the predictive attributes of dementia. This understanding is pivotal for guiding future medical assessments and informing healthcare strategies focused on dementia.

5. Conclusion on Random Forest Model

Implementing the Random Forest Classifier successfully met the core objectives by providing a high-accuracy classification of dementia status and identifying influential features in the predictive process. The model's robustness and reliability, as visualised in the classification report and confusion matrix (Figure 7), are crucial for advancing our understanding of dementia trends and improving patient care.

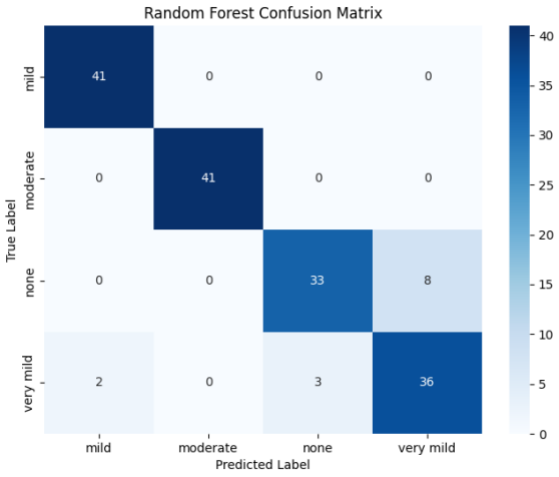

Figure 7: Confusion Matrix for Random Forest Model. This heatmap illustrates the actual vs. predicted labels, visually representing the Model's classification accuracy across the different dementia stages.

Support Vector Classifier (SVC) Model for Dementia Status Classification

1. Introduction to SVC Modeling

The SVC model was employed to classify dementia status among patients accurately, chosen for its proficiency in defining hyperplanes in a high-dimensional space, crucial for categorising varying stages of dementia severity.

2. Model Training and Hyperparameter Tuning

Careful dataset preprocessing ensured the selection of features suitable for the SVC model. Categorical variables were encoded, and numerical features were normalised, ensuring accurate data representation. SMOTE was utilised to counter the class imbalance in CDR, mirroring the approach in the Random Forest model. Hyperparameters {'C': 10, 'gamma': 'scale', 'kernel': 'linear'} were identified as optimal, enhancing the model's predictive accuracy and generalisation capabilities.

3. Model Evaluation

The SVC model showcased exceptional classification performance with an accuracy of 93.90%, as detailed in the classification report (Figure 8). This report demonstrates the model's high precision and recall in distinguishing the mild and moderate stages of dementia.

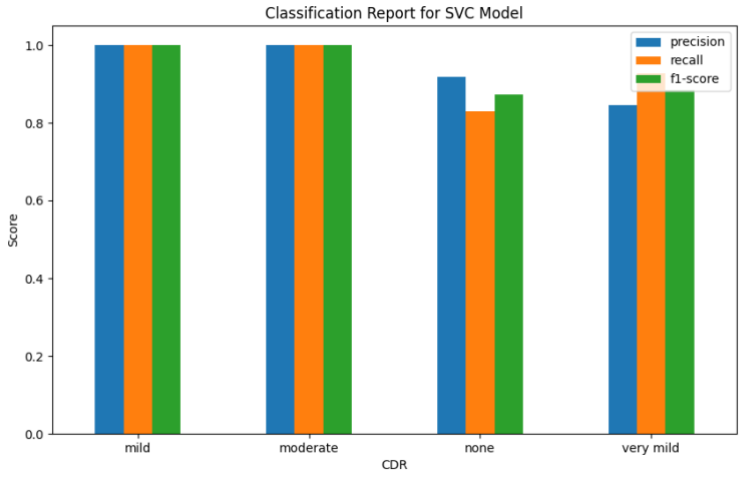

Figure 8: Classification Report for SVC Model. A bar chart detailing the precision, recall, and f1-score for each dementia category, highlighting the SVC model's diagnostic precision.

3.1 Feature Importance Analysis

Feature importance analysis revealed that MMSE scores, visit frequency, patient age, delay time, and years of education were among the most influential predictors. This underscores the Model's alignment with our objectives to identify critical features contributing to dementia status.

4. Insights and Implications

The results from the SVC model offer a deeper understanding of the attributes significantly impacting dementia classification. The model's prioritisation of features like MMSE scores in its decision-making process is instrumental in identifying patients at various dementia stages.

5. Conclusion on SVC Model

The Support Vector Classifier met and exceeded the objectives, achieving high accuracy in dementia classification and providing insights into feature relevance. The model's robustness and reliability, as visualised in the classification report and confusion matrix (Figure 9), are crucial for advancing our understanding of dementia trends and improving patient care.

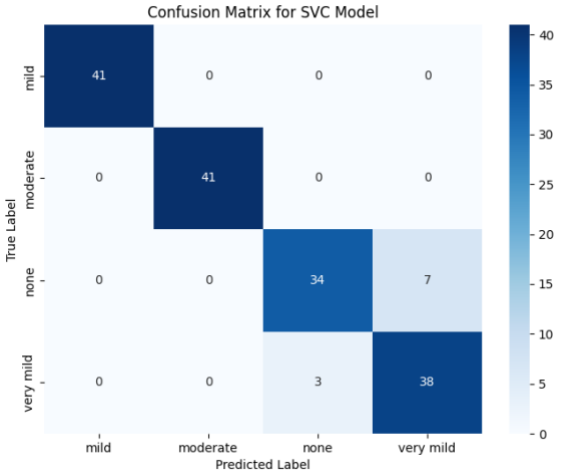

Figure 9: Confusion Matrix for SVC Model. This heatmap illustrates the actual vs. predicted labels, visually representing the Model's classification accuracy across the different dementia stages.

K-Means Clustering for Patient Group Characterization

1. Introduction to K-Means Clustering

K-Means clustering was implemented to identify and analyse distinct patient groups, aiming to uncover data structures that reveal insights into patient dementia status trends.

2. Clustering Methodology

The dataset was ready for clustering after preprocessing, including one-hot encoding of categorical variables and standardising numerical variables. The elbow method determined the optimal number of clusters, resulting in four distinct groups. The K-Means algorithm then partitioned the dataset accordingly.

3. Cluster Evaluation and Characterization

PCA was applied to visualise the clusters in a two-dimensional space (Figure 10), demonstrating clear group separation. Each cluster was characterised using descriptive statistics of various features, revealing patterns in dementia-related variables across different patient groups.

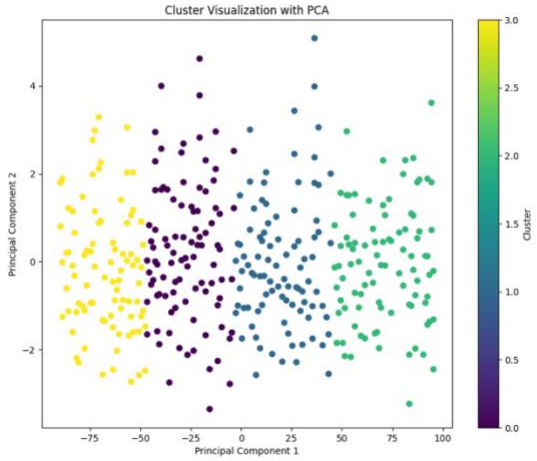

Figure 10: Cluster Visualization with PCA. This scatter plot uses PCA for dimensionality reduction, showing patients' clustering and clusters' distribution in the feature space.

Examples of the Clusters:

- Cluster 0 was typified by slightly lower MMSE scores and smaller brain volumes, hinting at a potential trend toward more severe dementia status.
- Cluster 1 showed a mean closer to the average for most variables, possibly representing a general patient demographic.
- Cluster 2 demonstrated higher brain volumes and better MMSE scores, which might correlate with milder dementia or a healthier patient segment.
- Cluster 3 presented with higher age and SES, suggesting a subgroup of patients with distinct socioeconomic and age-related characteristics.

4. Insights and Implications

The clustering provided valuable insights into patient demographics and medical profiles, aiding in understanding dementia's distribution and characteristics within the population. These characterisations can guide healthcare professionals in customising interventions and monitoring.

5. Conclusion on K-Means Clustering

K-means clustering was a powerful tool in the analysis, revealing significant patient groupings that align to deduce dementia status trends. This technique has enriched the dataset's understanding and opened avenues for further personalised patient care investigation.

Conclusion

This report undertook a detailed analysis of a biomedical dataset, with the primary goal of classifying dementia and Alzheimer's disease using advanced machine learning techniques. Through this analysis, significant strides were made in understanding the complex nature of dementia, its various stages, and the factors influencing its progression.

The Random Forest model, selected for its robustness in handling complex data, achieved a notable classification accuracy of 92.07%. This model effectively uncovered the most influential features for dementia classification, providing valuable insights into the disease's underlying mechanics. The balanced performance across different dementia severity stages, as depicted in the classification report (Figure 6) and the confusion matrix (Figure 7), highlighted its diagnostic accuracy.

Complementing the Random Forest model, the Support Vector Classifier (SVC) demonstrated a slightly higher accuracy of 93.90%. Its strength lies in distinguishing the mild and moderate stages of dementia, as shown in its detailed classification report (Figure 8) and confusion matrix (Figure 9). The SVC's ability to prioritize critical features such as MMSE scores, patient age, and years of education made it a powerful tool for identifying patients at various stages of dementia.

Furthermore, applying K-Means clustering offered a different yet equally important perspective. This approach provided a comprehensive view of the patient demographics and medical profiles by identifying distinct patient groups and characterizing them through PCA visualization (Figure 10) and descriptive statistics. This analysis was pivotal in understanding how dementia manifests differently across the patient population.

In summary, combining Random Forest and SVC models and K-Means clustering provided a multi-faceted understanding of dementia and Alzheimer's disease. While each method offered unique insights, they painted a comprehensive picture of the disease's impact on patients.

For future improvements, it's recommended to explore additional modelling techniques, incorporate more diverse datasets, and perform longitudinal studies to track dementia progression over time. Collaborating with healthcare professionals for clinical validation and focusing on patient-specific models will also be crucial. These advancements will not only deepen our understanding of dementia but also contribute significantly to developing more effective treatment strategies and improving patient care in this challenging field of medicine.