**Predictive Modeling of ALS Progression**

**Predictive Modeling of ALS Progression**

**Introduction to Data science**

**PROJECT REPORT**

***Submitted by***

Pullagura Tejaswini

**Predictive Modeling of ALS Progression**

**Abstract:**

This documentation presents a detailed workflow for analyzing the Minsk2020 ALS dataset using machine learning techniques. The preprocessing steps included converting categorical data to numeric values through ordinal encoding and performing correlation analysis. Exploratory Data Analysis (EDA) was conducted to understand data distribution and detect anomalies. Recursive Feature Elimination (RFE) was employed to select the most relevant features. Various machine learning models, including Random Forest, SVM, Decision Tree, Gaussian Naive Bayes, K-Nearest Neighbors, Logistic Regression, Gradient Boosting, AdaBoost, and Bagging Classifier, were implemented. Model performance was evaluated using cross-validation techniques, ensuring robust and reliable results. The comprehensive approach leveraged multiple preprocessing, feature selection, and evaluation strategies to develop effective machine learning models for the dataset.

**INTRODUCTION**

The Minsk2020 ALS dataset presents a valuable opportunity for leveraging machine learning (ML) techniques to improve ALS diagnosis and prognosis. In this documentation, we delve into the process of applying ML algorithms to this dataset, with a focus on feature engineering, model selection, and evaluation.

The initial step involves converting categorical columns into numeric format using ordinal encoding. This enables us to utilize these variables effectively in ML models, ensuring a comprehensive analysis. Subsequently, we perform correlation analysis to understand the relationships between different features, aiding in feature selection and model performance optimization.

Exploratory Data Analysis (EDA) plays a crucial role in uncovering insights into the data distribution, identifying patterns, and detecting anomalies. These insights guide our decision-making throughout the ML process, enhancing the quality and relevance of our models.

Feature selection is streamlined through Recursive Feature Elimination (RFE), which helps identify the most influential features for predictive modeling. This step contributes to model efficiency and interpretability, ensuring that our models are both accurate and understandable.

A wide range of ML algorithms is employed, including Random Forest classifiers, Support Vector Machines (SVM), Decision Tree classifiers, Gaussian Naive Bayes, K-Nearest Neighbors (KNN), Logistic Regression, Gradient Boosting, AdaBoost, and Bagging classifiers. Each algorithm brings unique strengths to the analysis, allowing us to explore different approaches and optimize model performance.

To evaluate the efficacy of our models and ensure their generalizability, cross-validation techniques are implemented. This process validates the robustness of our models and helps avoid overfitting, enhancing their reliability in real-world applications.

By leveraging these ML techniques and methodologies, we aim to develop predictive models that can accurately diagnose ALS and contribute to advancements in medical diagnosis and treatment strategies. This documentation serves as a comprehensive guide to our ML analysis on the Minsk2020 ALS dataset, highlighting the key steps and strategies employed to achieve meaningful insights and impactful results.

**Methodology**

+-------------------+

| Start |

+-------------------+

|

V

+-------------------+

| Data Preprocessing|

+-------------------+

|

V

+-------------------+

| Convert Categorical|

| Columns to Numeric |

+-------------------+

|

V

+-------------------+

| EDA (Exploratory |

| Data Analysis) |

+-------------------+

|

V

+-------------------+

| Feature Selection |

+-------------------+

|

V

+-------------------+

| Model Selection |

+-------------------+

|

V

+-------------------+

| Model Evaluation |

+-------------------+

|

V

+-------------------+

| Output / End |

+-------------------+

**Data Source:**

The dataset used in this machine learning analysis was sourced from Kaggle, a platform known for hosting diverse datasets across various domains. The specific dataset related to Amyotrophic Lateral Sclerosis (ALS) was obtained by searching for keywords such as "ALS dataset", "Amyotrophic Lateral Sclerosis dataset", and "medical dataset" on Kaggle.

**Dataset Description :**

The dataset encompasses a wide array of features crucial to ALS diagnosis and prognosis. These features include:

**Demographic Information:**

**Age:** The age of the individuals in the dataset.

**Gender:** The gender of the individuals, potentially providing insights into any gender-specific patterns related to ALS.

**Clinical Measurements**:

J1\_a, J3\_a, S1\_a: Clinical measurements related to ALS, although the exact nature of these measurements may require domain-specific knowledge for interpretation.

**Diagnostic Outcomes:**

Diagnosis (ALS): This likely serves as the target variable, indicating the presence or absence of ALS diagnosis.

**Data Pre-processing**

1. Converting Categorical columns to numeric

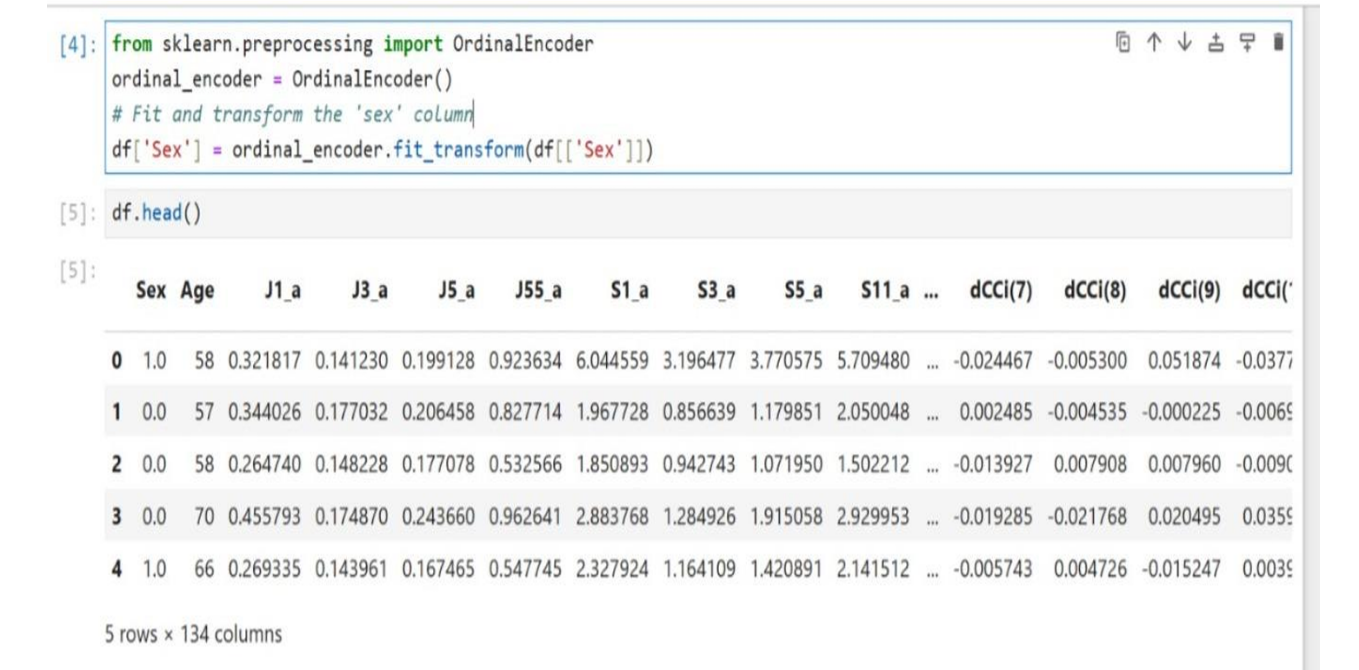
• Ordinary Encoding

Ordinal encoding is a technique used in data pre-processing to convert categorical variables into numerical format. It's particularly useful when dealing with ordinal categorical variables, where the categories have a natural order or hierarchy.

**Original Categorical Values: 'Male' and 'Female':**

• 'Male' is assigned the value 0 or 1 (depending on the encoding convention used).

• 'Female' is assigned the value 1 or 2.



1. Correlation Analysis:

• A correlation matrix was created to examine the relationships between different features in the dataset. Features with high correlation were carefully reviewed to avoid redundancy and multicollinearity in the models.





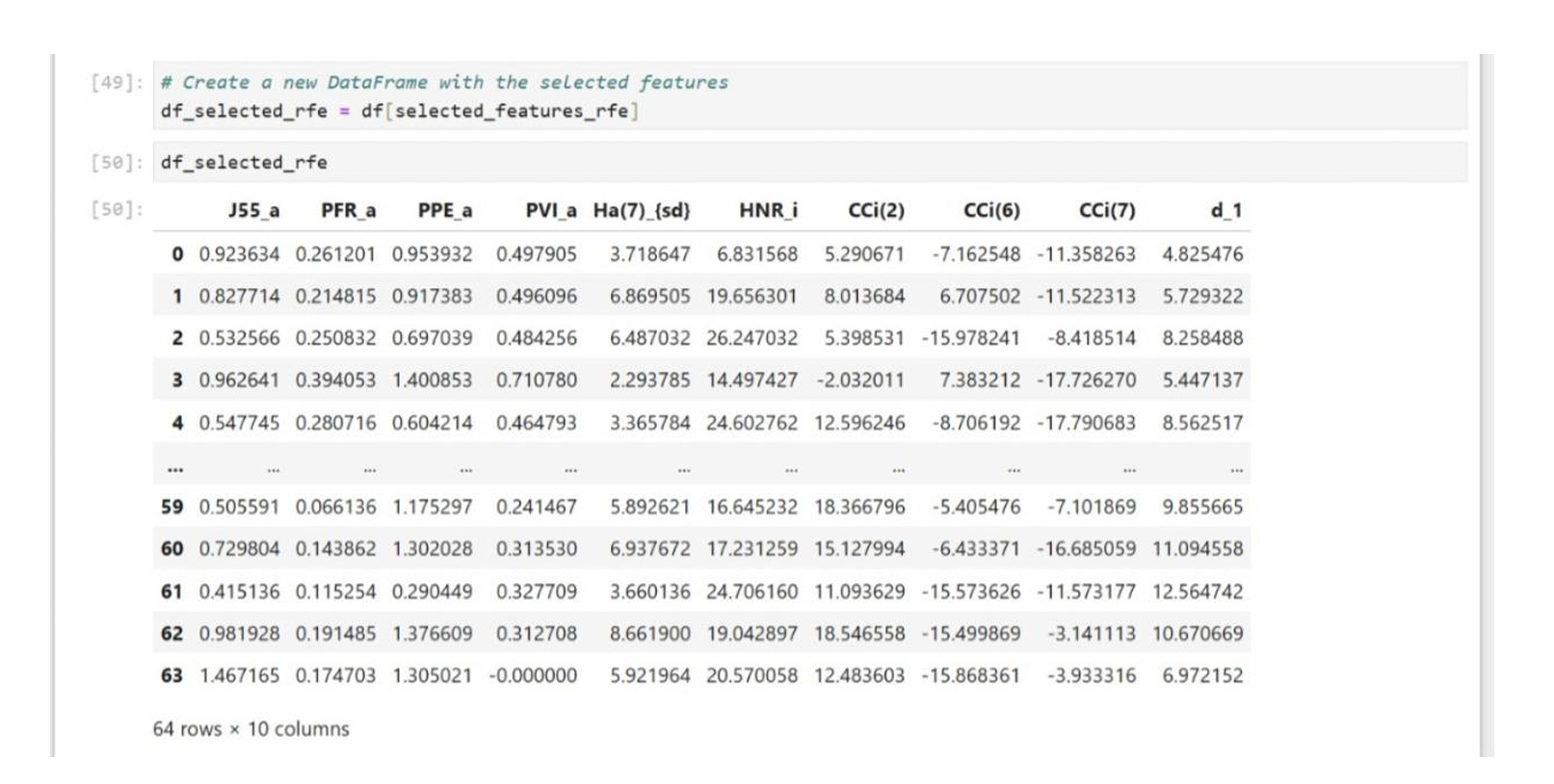
**Feature Selection**

1. Recursive Feature Elimination (RFE)

**Feature Ranking:**

RFE was applied to rank and select the most relevant features for the models. RFE recursively removes the least important features based on the model’s performance until the optimal subset of features is reached.

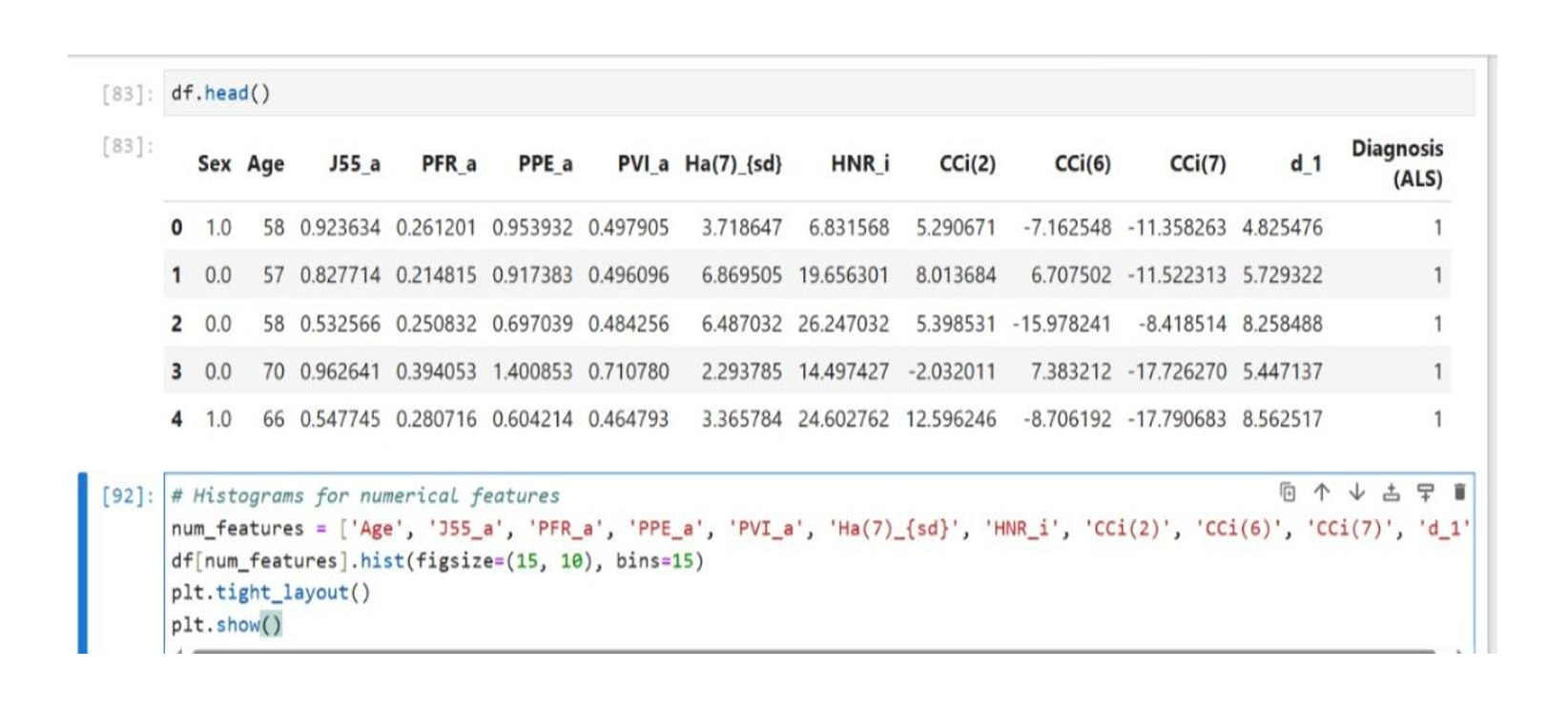


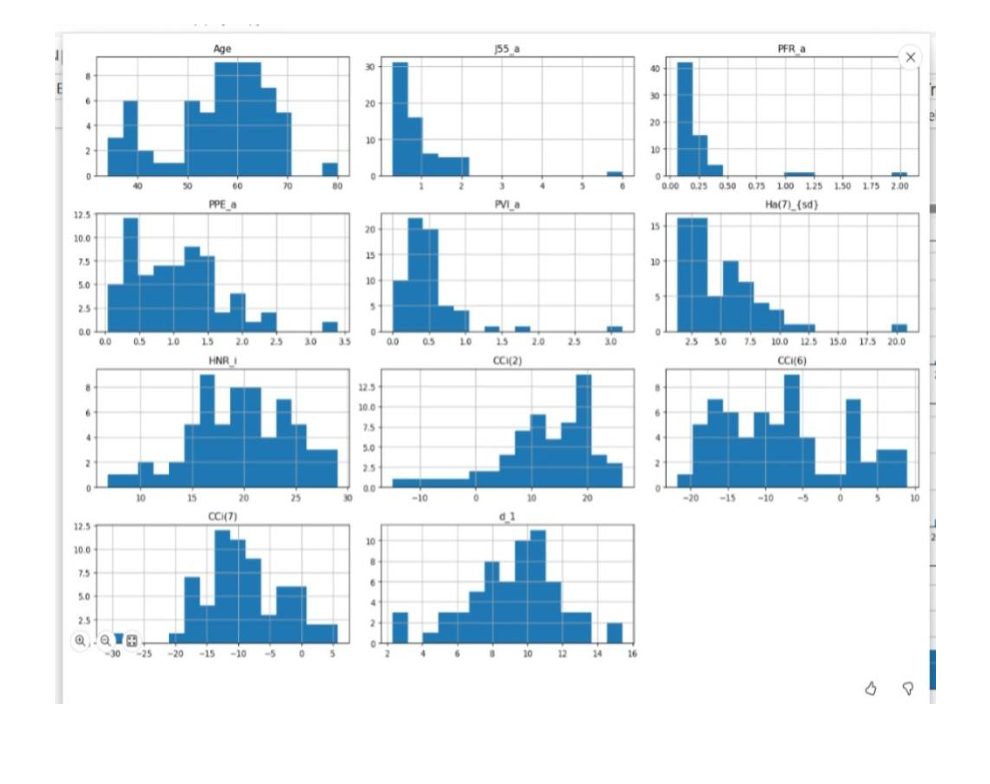


**Exploratory Data Analysis (EDA)**

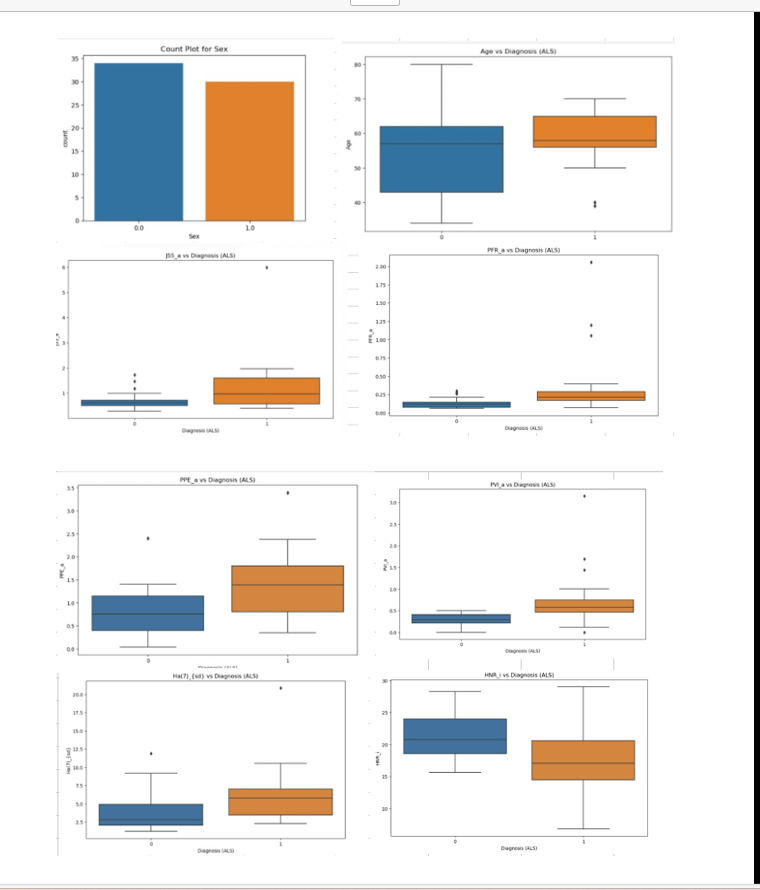
**• Data Visualization:**

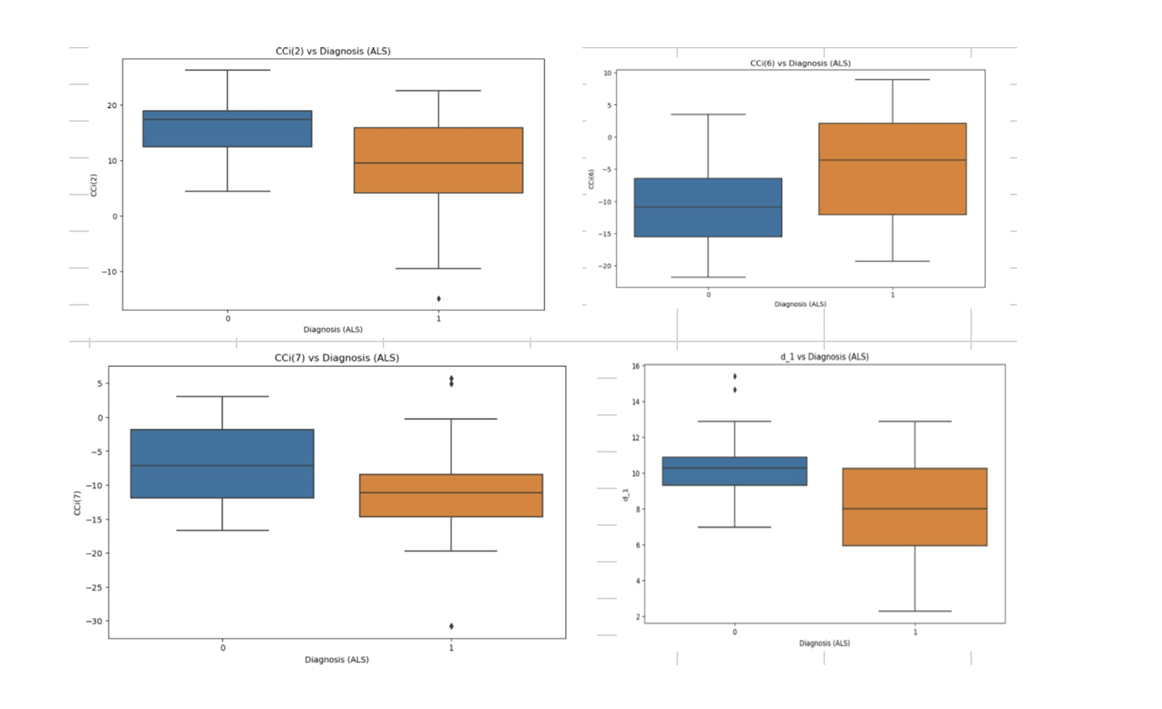
Various plots and charts (e.g., histograms, box plots, scatter plots) were used to visualize the distribution and relationships of the data.











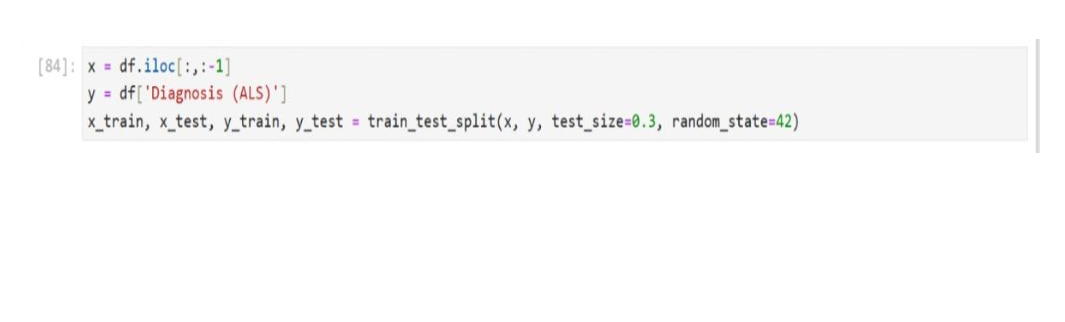
**Training Machine Learning Models**

1. **Training Dataset**:

The training dataset is a crucial component in machine learning. It comprises labeled data used to train the model, where each data point consists of input features and their corresponding output labels or target values.

1. **Splitting Data:**

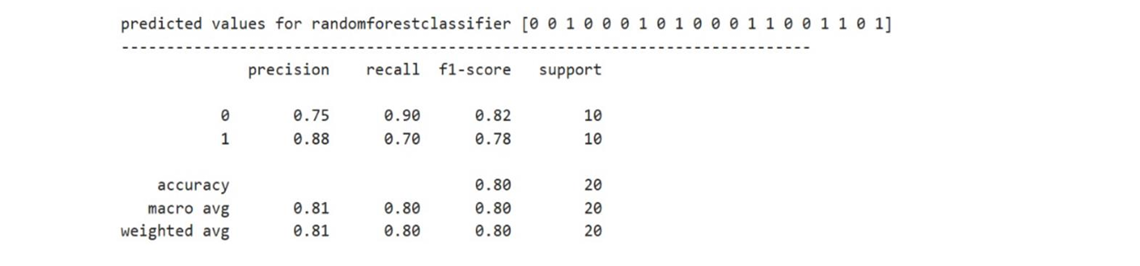
Before training, the dataset is typically split into training, validation, and testing sets. The training set is used to train the model, the validation set helps tune hyperparameters and monitor model performance during training, and the testing set evaluates the final model's performance on unseen data.



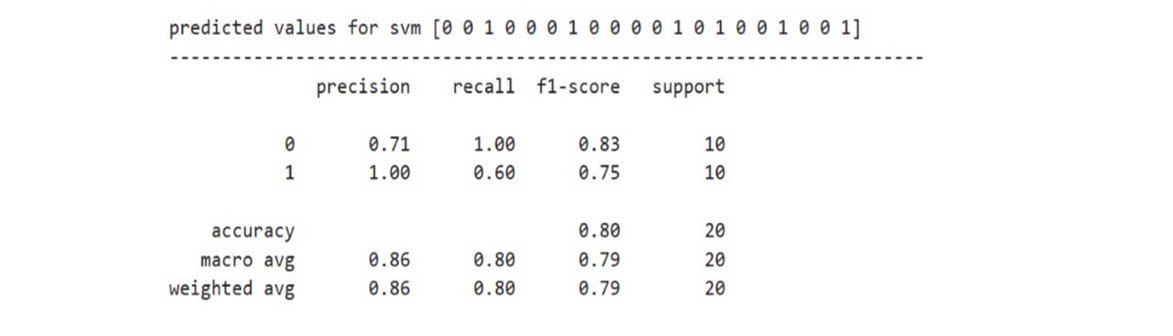
**Machine Learning Models**

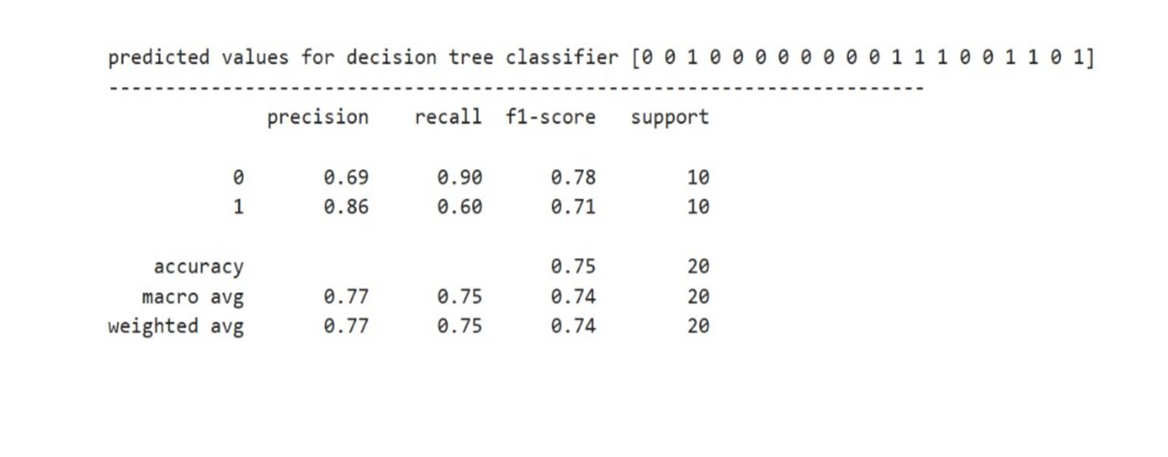
**1.Random Forest Classifier:**

An ensemble method that creates multiple decision trees and merges their results to improve accuracy and control overfitting.

**2.Support Vector Machine (SVM):**

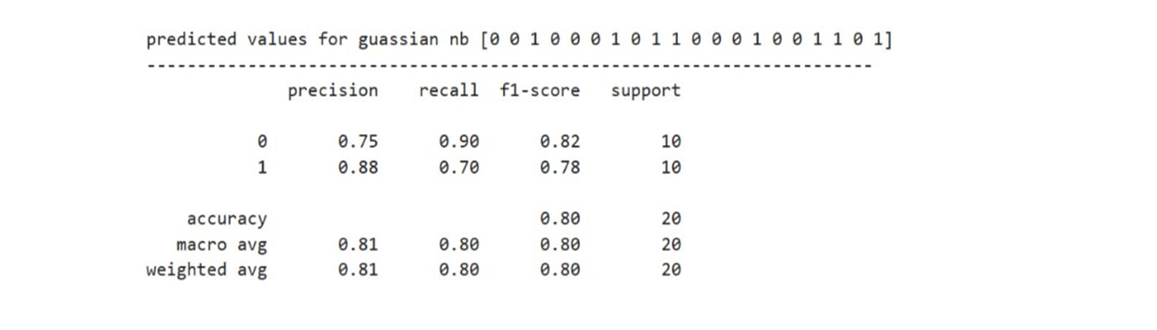
A classification technique that finds the optimal hyperplane to separate different classes in the dataset.

**3.Decision Tree Classifier:**

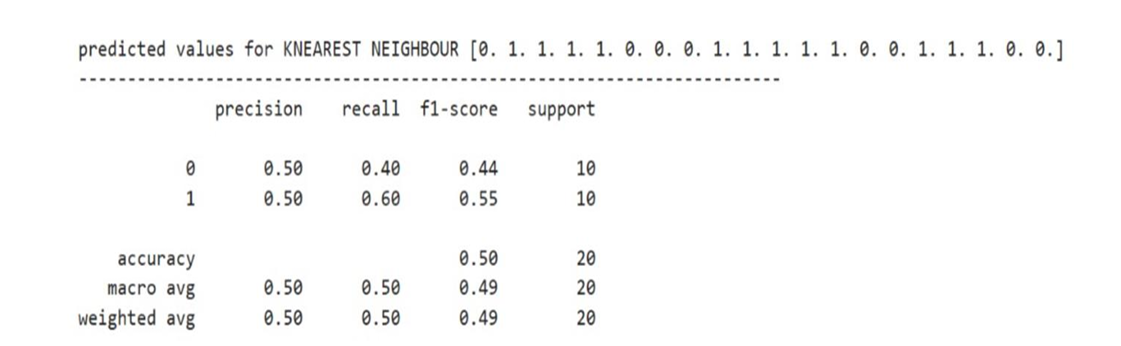
A tree-based model that splits the data into subsets based on the feature values, making decisions at each node.

**4.Gaussian Naive Bayes (GaussianNB):**

A probabilistic classifier based on Bayes' theorem, assuming features follow a Gaussian distribution.

**5. K-Nearest Neighbors (KNN):**

A non-parametric algorithm that classifies a data point based on the majority class among its k-nearest neighbors.

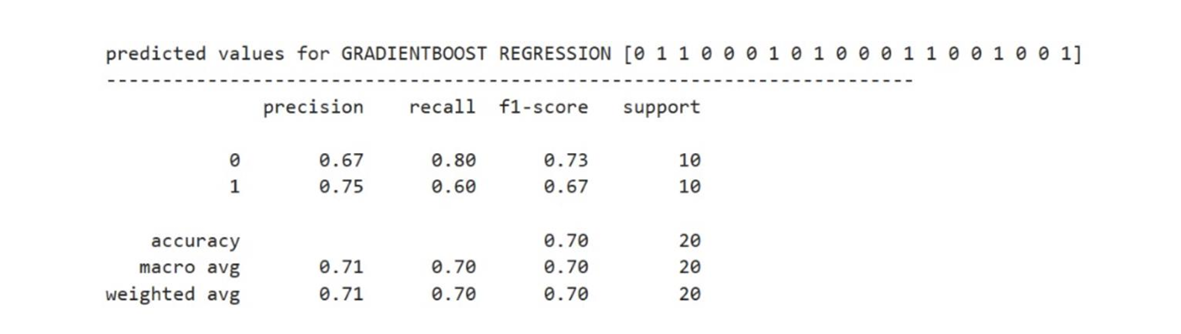
**6.Logistic Regression:**

A statistical model that uses a logistic function to model the probability of a binary outcome.



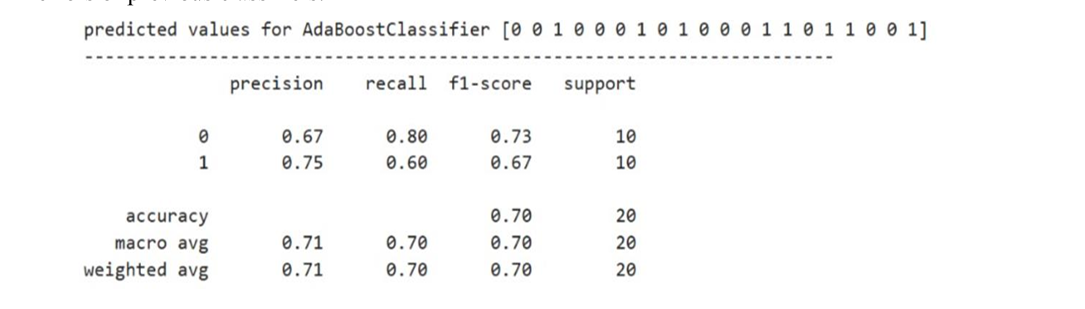
**7.Gradient Boosting:**

An ensemble technique that builds multiple weak learners (usually decision trees) sequentially, each correcting the errors of its predecessor.



**8.AdaBoost:**

An ensemble method that combines multiple weak classifiers, adjusting their weights based on the errors of previous classifiers.

**9.Bagging Classifier:**

An ensemble technique that combines the results of multiple models (usually the same type) trained on different subsets of the data.



**Conclusion**

This comprehensive approach ensured that the best possible machine learning models were developed for the minsk2020\_als\_dataset, leveraging various preprocessing, feature selection, and evaluation techniques to achieve robust and reliable results.

**References**

**1**.ALS and Medical Research References:

"Amyotrophic Lateral Sclerosis: From Diagnosis to Multidisciplinary Management" edited by Yasmine M. Argov and Michael Swash

"Understanding ALS: From Diagnosis to Treatment" by Naoki Atsuta and Akio Kimura

"Advances in Understanding the Pathophysiology of Amyotrophic Lateral Sclerosis" edited by Alvaro G. Estévez

**2**.Online Resources and Websites:

Towards Data Science (<https://towardsdatascience.com/>)

Kaggle Notebooks and Datasets (<https://www.kaggle.com/>)

GitHub repositories with machine learning projects and ALS-related datasets.