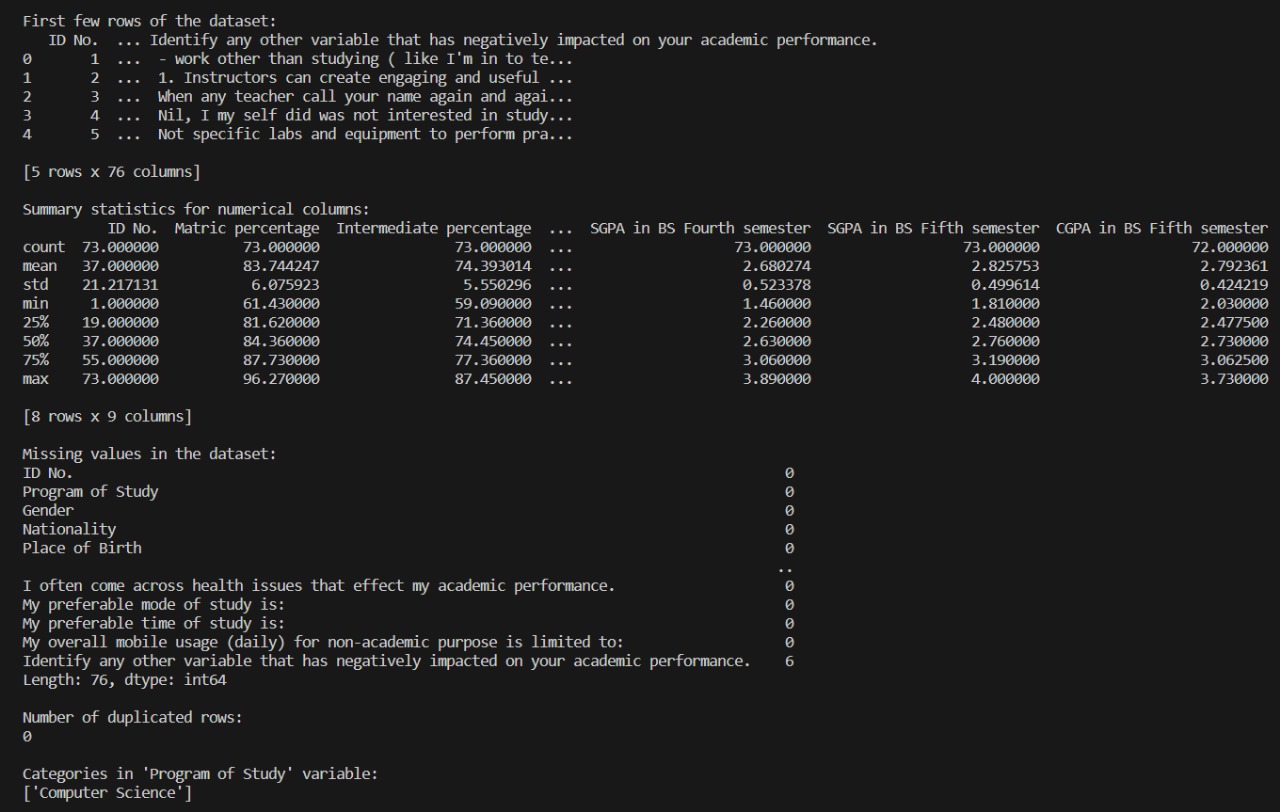
****

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Project Management**  **Project Assignment 3:**  **Submit To: Mam Saima**  Submitted by:   |  |  | | --- | --- | | **Name** | **Registration numbs** | | **Muhammad Haider** | **200901056** | | **Hamaad Ali** | **200901044** | | **Syed Abdullah Ashar** | **200901074** | | **Muhammad Ali** | **200901075** | | **Hasnat Shahid** | **200901036** |   **Introduction**  The project focuses on predicting students' Grade Point Average (GPA) for the fifth semester of their Bachelor's degree program using machine learning models. The dataset used for this analysis is stored in an Excel file named 'DataSet.xlsx.' The project is divided into three main parts: Data Preprocessing, Exploratory Data Analysis (EDA), and GPA Prediction through a Graphical User Interface (GUI). Below is a detailed report on each section of the project.  **Data Preprocessing**  Loading and Cleaning the Dataset  The project begins by importing necessary libraries and loading the dataset from the 'DataSet.xlsx' file. The dataset is read into a Pandas DataFrame, and any leading or trailing whitespaces in column names are removed for consistency.  **Dataset Overview**  The basic information about the dataset is displayed using df.info(). It provides details such as the number of non-null values, data types, and memory usage. This summary helps in understanding the structure of the dataset.  **Cleaning and Imputing Missing Values**  The dataset is checked for missing values, and it is observed that some values are missing. A SimpleImputer from scikit-learn is used to replace missing values with the mean of the respective columns. This ensures a complete dataset for training machine learning models.  **Feature Selection**  Relevant features for GPA prediction are selected, including 'Matric percentage,' 'Intermediate percentage,' and SGPA values for the first four semesters. The target variables are 'SGPA in BS Fifth semester' and 'CGPA in BS Fifth semester.'  **Standardization**  To ensure consistency in the scale of features, the data is standardized using the StandardScaler from scikit-learn. This step is crucial for algorithms sensitive to the scale of input features, such as linear regression and support vector machines.  **Train-Test Split**  The dataset is split into training and testing sets using train\_test\_split from scikit-learn. A 70-30 split is used, with 70% of the data used for training the machine learning models.  **Exploratory Data Analysis (EDA)**  **Initial Data Exploration**  The EDA section begins with a visualization of the basic statistics of numerical columns using df.describe(). It provides insights into the central tendency, dispersion, and shape of the distribution of each numerical feature.  **Visualization of Categorical Variables**  Categorical variables such as 'Gender,' 'Nationality,' and 'Intermediate Stream' are explored using count plots. These visualizations help understand the distribution of categories within each variable.  **Histogram and Scatter Plot**  A histogram is created for the 'Parental Income' variable, providing a visual representation of the distribution of parental income. Additionally, a scatter plot is generated to explore the relationship between 'Matric percentage' and 'Intermediate percentage.'  **Box Plot**  A box plot is created for the 'SGPA in BS First semester' variable to visualize the distribution and identify potential outliers.  **Additional EDA**  Several additional categorical variables related to student preferences and experiences are explored through count plots. This provides further insights into students' opinions and behaviors.  **GPA Prediction through GUI**  **GUI Setup**  The project includes a graphical user interface (GUI) built using the Tkinter library for GPA prediction. The GUI allows users to input values for relevant features, and upon clicking the "Predict GPA" button, the predicted SGPA and CGPA are displayed, along with their respective performance categories.  **GPA Prediction Function**  A function named predict\_gpa takes the input data, imputes missing values, scales the data, and predicts both SGPA and CGPA using previously trained machine learning models.  **Performance Categories**  A function named get\_performance\_category assigns performance categories based on GPA ranges. These categories provide additional insights into students' academic performance.  **Evaluation and Comparison**  The predicted GPAs are compared with actual values from the test set. Mean Squared Error (MSE) and R^2 Score are calculated for each machine learning model (Linear Regression, Random Forest, and Support Vector Regressor). The results are printed in the terminal.  **Visualization**  Finally, the project includes visualizations such as scatter plots for actual vs. predicted GPAs and residual plots for model evaluation.  **Work Breakdown Structure:**    **Activity On Node:**    Critical path is 15 days:  **Screenshots Of The Project:**    **Exploratory Data Analysis (EDA):** |

**Preprocessing:**

**A screenshot of a computer

Description automatically generated**

****

**Accuracy Measurements**

A screenshot of a computer

Description automatically generated