# Exploring Factors Influencing Student Succes in School/University

جامعة مصر للمعلوماتية
EGYPT UNIVERSITY OF INFORMATICS


Presented by:

Ahmed Shebl 21-101005
Omar Malek 21-101109
Mostafa Khairy 21-101162
Ahmed Alaroussy 21-101168
Kareem Abdelhameed 21-101015
Mostafa Abdelkawy 21-101185

Course Instructors:

Prof. Fatma Elshahaby
Eng. Nervana Abdullah


Course:

Data Science Methodology (CSE 271)

# Project Abstract

Institutions dedicated to education continually seek to enhance student success by understanding the multifaceted factors that influence academic performance. Numerous studies have explored various aspects of this complex phenomenon, highlighting the importance of factors such as socio-economic background, prior academic achievement, learning styles, and extracurricular involvement. This project contributes to this body of knowledge by conducting a comprehensive analysis of a dataset comprising school and university student data. Through data visualization, hypothesis testing, and regression analysis, we aim to uncover the key variables that significantly impact academic performance. By doing so, we provide valuable insights for educational institutions to implement targeted interventions and improve overall learning outcomes. This introduction sets the stage for our investigation, emphasizing the significance of understanding the factors contributing to academic achievement and outlining the methodology employed in our analysis.

# Table of Contents

# Introduction

This project aims to analyze various datasets related to students' performance and academic life to draw meaningful conclusions and implement changes that benefit students and enhance the educational ecosystem. We initially focused on two datasets, but unfavorable results led us to expand our analysis to four datasets. These datasets were visualized using optimal techniques to clearly present the information. Hypothesis testing was conducted on all datasets except the second, due to its entirely categorical nature, which posed challenges for such analysis. Regression analysis was applied to the first and fourth datasets, while classification analysis was utilized for the second and third datasets.

# Materials

## Dataset One

### Dataset Description: Student Attitude and Behavior Dataset

The Student Attitude and Behavior Dataset comprises data collected from university students via a Google form. This dataset encompasses various parameters including certification course completion, gender, department of study, height (in cm), weight (in kg), academic performance in 10th and 12th grades, college marks, hobbies, daily study duration, preferred study setting, salary expectations, satisfaction with their degree, inclination towards pursuing a career aligned with their degree, engagement with social media and video content, commuting time, perceived stress levels, and financial background.

### Column Descriptions:

- Certification Course: Indicates if the student has completed any certification course.
- Gender: Denotes the gender of the student.
- Department: Specifies the field of study in which the student is enrolled.
- Height (CM): Represents the height of the student in centimeters.
- Weight (KG): Indicates the weight of the student in kilograms.
- 10th Mark: Reflects the student's academic performance in the 10th grade.

- 12th Mark: Reflects the student's academic performance in the 12th grade.
- College Mark: Reflects the student's academic performance in their college or university.
- Hobbies: Specifies the hobbies or interests of the student.
- Daily Studying Time: Denotes the amount of time the student spends studying on a daily basis.
- Prefer to Study in: Specifies the preferred study environment or location of the student.
- Salary Expectation: Reflects the student's expectation for their future salary.
- Do you like your degree?: Indicates the student's satisfaction with their degree.
- Willingness to pursue a career based on their degree: Denotes the student's inclination towards pursuing a career aligned with their degree.
- Social Media & Video: Reflects the student's engagement with social media and video platforms.
- Traveling Time: Denotes the time taken by the student to commute or travel to their educational institution.
- Stress Level: Indicates the perceived stress level of the student.
- Financial Status: Specifies the financial status or economic background of the student.
- Part-time Job: Indicates whether the student is engaged in a part-time job or not.

## Dataset One Wrangling

After conducting a thorough examination of the dataset, it was found that while null values were absent, certain entries exhibited unrealistic values in fields such as height (CM), weight (KG), and salary expectations. The identification of these outliers was facilitated by visualizing each field against the college mark field through scatter plots. Subsequently, a robust method involving the calculation of the interquartile range was employed to ascertain the realism of the data within each field. Following this analysis, entries containing outlier values were removed from the dataset to uphold data cleanliness, integrity, and ensure the preservation of realistic trends.

# Dataset Two

## Dataset Description: Student Performance Data

A comprehensive Google Form survey was conducted to gather insights into the factors influencing student academic performance across various universities in Pakistan. The survey encompassed a well-considered selection of attributes. The dataset has no numerical values and contained only categorical values.

## Column Descriptions:

- University Name: Identifying the institution with which the participant is affiliated.
- Degree Name: Capturing the academic program in which the participant is enrolled.
- Average Attendance: Quantifying the regularity of attendance, an essential factor in academic success.
- Average Study Time per Day: Exploring the daily allocation of time to study activities.
- Extracurricular Activities: Investigating participation in activities beyond the academic curriculum.
- Average Sleep Time: Understanding the role of sleep patterns in academic performance.
- Family and Personal Factors:
    - Household Size: Exploring the potential influence of family structure.

    - Parents' Highest Qualification: Gauging the educational background of the participant's parents.
- Workout: Investigating the impact of physical activity on academic outcomes.
- Free Time Activity: Capturing the nature of activities during leisure time.
- University Society Membership: Identifying whether the participant is a member of any university society.

The primary outcome variable targeted in this survey is the participant's GPA, representing their academic performance.

## Dataset Two Wrangling

The initial phase of data wrangling involved standardizing field names such as 'University', 'Degree', 'Field', and 'Extracurricular Activities'. This standardization ensured consistency across datasets, mitigating the risk of identical values being perceived differently due to variations in wording or nomenclature.

Subsequently, attention was directed towards addressing null values in the 'Study Hours' field. Notably, these null entries shared a common GPA range. To remedy this, a methodological approach was adopted wherein the mean study hours corresponding to the specific GPA range were calculated. These mean values were then utilized to replace the null entries, ensuring data completeness and integrity.

# Dataset Three

## Dataset Description:

The "Campus Placement Prediction" dataset provides a comprehensive set of attributes aimed at predicting the outcome of candidate selection during campus placement processes. This dataset offers valuable insights into the factors influencing a candidate's success in securing placement opportunities within various academic institutions and corporate entities. This dataset is already clean and doesn't require further wrangling.

## Column Descriptions:

- Gender (Categorical): Represents the gender identity of the candidate participating in the placement process.
- Secondary Education Percentage (Numerical): Denotes the percentage score obtained by candidates in their secondary education.
- Secondary Education Board (Categorical): Indicates the educational board associated with the candidate's secondary education.
- Higher Secondary Education Percentage (Numerical): Reflects the percentage score attained by candidates in their higher secondary education.
- Higher Secondary Education Board (Categorical): Identifies the educational board governing the candidate's higher secondary education.
- Higher Secondary Education Stream (Categorical): Specifies the academic stream pursued by candidates during their higher secondary education.
- Undergraduate Degree Percentage (Numerical): Signifies the percentage score achieved by candidates in their undergraduate degree program.

- Undergraduate Degree Type (Categorical): Characterizes the type of undergraduate degree pursued by candidates.
- Work Experience (Categorical): Indicates whether candidates possess prior work experience.
- Employability Test Percentage (Numerical): Represents the percentage score obtained by candidates in employability tests.
- MBA Percentage (Numerical): Indicates the percentage score attained by candidates in their Master of Business Administration (MBA) program.
- Specialization (Categorical): Specifies the specialization area of candidates in their MBA program.
- Placement Status (Categorical): Serves as the target variable, indicating whether candidates were placed or not during the campus placement process.

# Dataset Four

## Dataset Description:

The dataset examines student achievement in secondary education across two Portuguese schools. It encompasses student grades, demographic, social, and school-related features, collected through school reports and questionnaires. Two distinct datasets are provided, focusing on performance in Mathematics (mat) and Portuguese language (por). It's important to note that the target attribute G3 exhibits a strong correlation with attributes G2 and G1. This correlation arises because G3 represents the final year grade (issued at the 3rd period), while G1 and G2 correspond to the 1st and 2nd period grades. Predicting G3 without considering G2 and G1 is challenging, yet such prediction holds significant utility.

## Column Description:

- school (Categorical): Student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira).
- sex (Binary): Student's sex (binary: 'F' - female or 'M' - male).
- age (Integer): Student's age (numeric: from 15 to 22).
- address (Categorical): Student's home address type (binary: 'U' - urban or 'R' - rural).
- famsize (Categorical): Family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3).
- Pstatus (Categorical): Parent's cohabitation status (binary: 'T' - living together or 'A' - apart).
- Medu (Integer): Mother's education level (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education, or 4 - higher education).

- Fedu (Integer): Father's education level (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education, or 4 – higher education).
- Mjob (Categorical): Mother's occupation (nominal: 'teacher', 'health' care related, civil 'services' (e.g., administrative or police), 'at_home', or 'other').
- Fjob (Categorical): Father's occupation (nominal: 'teacher', 'health' care related, civil 'services' (e.g., administrative or police), 'at_home', or 'other').
- reason (Categorical): Reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference, or 'other').
- guardian (Categorical): Student's guardian (nominal: 'mother', 'father', or 'other').
- traveltime (Integer): Home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour).
- studytime (Integer): Weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours).
- failures (Integer): Number of past class failures (numeric: n if 1<=n<3, else 4).
- schoolsup (Binary): Extra educational support (binary: yes or no).
- famsup (Binary): Family educational support (binary: yes or no).
- paid (Binary): Extra paid classes within the course subject (Math or Portuguese) (binary: yes or no).
- activities (Binary): Extra-curricular activities (binary: yes or no).
- nursery (Binary): Attended nursery school (binary: yes or no).
- higher (Binary): Wants to take higher education (binary: yes or no).
- internet (Binary): Internet access at home (binary: yes or no).
- romantic (Binary): With a romantic relationship (binary: yes or no).
- famrel (Integer): Quality of family relationships (numeric: from 1 - very bad to 5 - excellent).
- freetime (Integer): Free time after school (numeric: from 1 - very low to 5 - very high).
- goout (Integer): Going out with friends (numeric: from 1 - very low to 5 - very high).

- Dalc (Integer): Workday alcohol consumption (numeric: from 1 - very low to 5 - very high).
- Walc (Integer): Weekend alcohol consumption (numeric: from 1 - very low to 5 - very high).
- health (Integer): Current health status (numeric: from 1 - very bad to 5 - very good).
- absences (Integer): Number of school absences (numeric: from 0 to 93).
- G1 (Integer): First period grade (numeric: from 0 to 20).
- G2 (Integer): Second period grade (numeric: from 0 to 20).
- G3 (Integer): Final grade (numeric: from 0 to 20, output target).

# Methods

## Data Visualization

Visualizing data is a crucial step in understanding its underlying patterns, trends, and relationships. In this report, we explore the methods used for visualizing four datasets, employing various types of visualizations, and leveraging popular libraries and functions. The datasets were visualized using Python programming language along with specific libraries such as Plotly and Dash for interactive visualizations.

### Types of Visualizations

- Scatter Plots: Scatter plots were used to visualize the relationship between numerical variables. They help in identifying correlations and patterns between different features.
- Histograms: Histograms were employed to represent the distribution of numerical variables. They provide insights into the spread and shape of data distributions.
- Bar Charts: Bar charts were utilized to visualize categorical variables and compare different categories. They are effective in displaying frequency or count data.
- Box Plots: Box plots were employed to visualize the distribution of numerical data and identify outliers, quartiles, and median values. They provide a concise summary of data distribution.
- Violin Plots: Violin plots were used to visualize the distribution of numerical data similar to box plots, but they also display the probability density of the data at different values.
- Pie Charts: Pie charts were employed to display the proportion of categories within a categorical variable. They are useful for understanding the distribution of categorical data.

- Heat Map: A heat map is a two-dimensional data representation where different colors correspond to different values. This enables users to rapidly identify the most significant data points.
- 3D Scatter Plots: 3D Scatter plots were utilized to visualize the relationship between three numerical variables simultaneously. They were used to explore the relationship between marks and stress levels, hobbies, financial status, and part-time jobs

## Libraries Used

- Plotly: Plotly is a Python library used for creating interactive and publication-quality visualizations. It offers a wide range of chart types and customization options.
- Dash: Dash is a productive Python framework for building web applications. It enables the creation of interactive dashboards and data visualization interfaces.

# Hypothesis Analysis

In this report, we conduct a hypothesis test on multiple datasets to examine to draw useful conclusions and help in creating a better learning environment. We use a significance level (α) of 0.05 to evaluate the results.

## Techniques Used

- We employ the following techniques for hypothesis testing:
- Sampling: We randomly sample 25% of the data from the population using the sample() function in Python's pandas library.
- Descriptive Statistics: We calculate the mean weight of both the sample and the population.
- Standard Error Calculation: We compute the standard error of the mean for the sample.
- Z-test: We use the Z-test statistic to determine the probability (p-value) of observing the sample mean given the population mean.
- P-value Calculation: The two-tailed p-value was computed based on the Z-score.
- Decision Making: Based on the p-value, we make a decision regarding the null hypothesis.
- Paired t-test: We use the ttest_rel() function from the scipy.stats library to perform a paired t-test, which compares the means of two related groups to determine if there is a statistically significant difference between them.
- Hypothesis Formulation: Our null hypothesis (H0) states that there is no significant difference in mean percentage scores between undergraduate degree and MBA. The alternative hypothesis (H1) suggests that there is a significant difference.
- Decision Rule: Based on the p-value obtained from the t-test, we decide whether to reject or fail to reject the null hypothesis.

# Regression Analysis

Two regression models were developed in order to help predict the dominant features that affect students' success and grades. The regression analysis models use certain techniques and methods in order to improve the accuracy and metrics provided by the model.

Methods Used:

- Data Preprocessing: The datasets were preprocessed using one-hot encoding for nominal variables and label encoding for ordinal variables. Nominal variables were one-hot encoded to convert categorical data into a format suitable for machine learning algorithms. Ordinal variables were encoded with label encoding to preserve their ordinal nature.
- Feature Selection: Correlation analysis was conducted to identify significant features correlated with the target variables. Features with an absolute correlation coefficient greater than 0.13 in dataset one and 0.15 in dataset four were considered significant.
- Normalization: Numerical features were normalized using standard scaling to ensure all features have the same scale. This prevents features with larger magnitudes from dominating the model training process.
- Model Selection: A Ridge regression model was chosen for its ability to handle multicollinearity in the dataset and mitigate overfitting by penalizing large coefficients. Ridge regression is suitable for situations where there are correlated predictors, as it can help stabilize coefficient estimates.
- Cross-Validation: K-fold cross-validation with k=5 was employed to evaluate the model's performance and generalize its accuracy. Mean squared error (MSE) and $R^2$ score were used as evaluation metrics to assess the model's predictive capability.

# Classification Analysis

The Bayesian classification model employed in this analysis is built upon the Gaussian Naive Bayes algorithm. Here's a breakdown of the techniques utilized:

Data Preprocessing:

- The dataset was split into training and testing sets using the train_test_split function from the sklearn.model_selection module.
- Categorical variables were identified using their data type as 'object', using list comprehension.
- The dataset was first divided into numerical and categorical features.
- Numerical features were identified using their data types as non-object.
- Category encoding was performed using the OneHotEncoder from the category_encoders module. This step transformed categorical variables into numerical representations suitable for machine learning algorithms.

Model Training:

- Gaussian Naive Bayes (GNB) was chosen as the classification algorithm due to its simplicity and efficiency, particularly for datasets with continuous features.
- The GNB model was trained on the training set after encoding the categorical variables.
- K-Nearest Neighbors (KNN) algorithm was chosen for classification.
- The algorithm was trained on the training set using the KNeighborsClassifier from the sklearn.neighbors module.
- Cross-validation was performed to select the optimal number of neighbors (k) using 7-fold cross-validation.

Evaluation Metrics:

- Confusion matrix: Used to evaluate the performance of the classifier, showing the count of true positives, true negatives, false positives, and false negatives.
- Classification Report: Provides a comprehensive summary of the model's performance including precision, recall, F1-score, and support for each class.
- Accuracy: The accuracy of the KNN model was calculated using the accuracy_score function from the sklearn.metrics module.
- Scatter Plot: A scatter plot of the dataset was created to visualize the distribution of 'degree_p' and 'mba_p' features, with color indicating class labels.

# Results and Findings

## Hypothesis

Dataset One:

After conducting the hypothesis test on dataset one, we obtained the following results:

- Sample Size: The size of the sample is 54.
- Standard Error: The standard error of the mean for the sample is approximately 1.95.
- Population Mean: The population mean weight is approximately 60.92 kg.
- Sample Mean: The sample mean weight is approximately 59.11 kg.
- P-value: The p-value for the hypothesis test is approximately 0.353.
- Decision: With a significance level of 0.05, we fail to reject the null hypothesis. There is not enough evidence to support the claim that the sample mean weight is significantly different from the population mean weight.

Dataset Three:

After conducting the paired t-test, we obtained the following results:

- P-value: The p-value for the hypothesis test is approximately 4.14e-22, which is significantly lower than the significance level of 0.05.
- Decision: With such a low p-value, we reject the null hypothesis. Therefore, there is a significant difference in mean percentage scores between undergraduate degree and MBA.

Dataset Four:

After conducting the hypothesis test, the following findings were obtained:

- Population Mean: The mean number of absences in the population is approximately 3.66.
- Sample Mean: The mean number of absences in the sample group is approximately 3.98.
- Sample Size: The sample consists of 162 observations.
- Standard Error: The standard error of the mean for the sample is approximately 0.36.
- Z-score: The calculated Z-score is approximately 0.88.
- P-value: The computed p-value for the hypothesis test is approximately 0.38.
- Decision: With a p-value greater than the significance level (0.05), the null hypothesis cannot be rejected. There is not enough evidence to support the claim that the mean number of absences in the sample group differs significantly from that of the population.

# Regression

Dataset One:

- Cross-Validation Mean Squared Error (MSE): The average MSE across the 5-fold cross-validation was found to be approximately 186.87, indicating the average squared difference between the predicted and actual college marks.
- Cross-Validation Root Mean Squared Error (RMSE): The RMSE, calculated as the square root of the MSE, was approximately 13.67. This metric represents the average deviation of the predicted college marks from the actual marks and provides a measure of the model's prediction accuracy.
- Cross-Validated $R^2$ Score: The $R^2$ score, which measures the proportion of the variance in the target variable that is predictable from the independent variables, was approximately 0.254. This indicates that the model explains around 25.4% of the variance in college marks, suggesting moderate predictive capability.

Feature Coefficients:

- 10th Mark Coefficient: For every standard deviation increase in the 10th-grade marks, the predicted college marks are expected to increase by approximately 3.62 units.
- 12th Mark Coefficient: Similarly, for every standard deviation increase in the 12th-grade marks, the predicted college marks are expected to increase by approximately 2.83 units.
- Other Coefficients: Various categorical features such as gender, department, certification courses, preferences, and attitudes towards the degree showed significant coefficients, indicating their impact on predicting college marks.

Intercept: The intercept term of the regression model was found to be approximately 69.00. This represents the expected value of the college marks when all predictor variables are set to zero.

Dataset Four:

- Cross-Validation Mean Squared Error (MSE): The average MSE across the 5-fold cross-validation was found to be approximately 1.73, indicating the average squared difference between the predicted and actual final grades.
- Cross-Validation Root Mean Squared Error (RMSE): The RMSE, calculated as the square root of the MSE, was approximately 1.31. This metric represents the average deviation of the predicted final grades from the actual grades and provides a measure of the model's prediction accuracy.
- Cross-Validated $R^2$ Score: The $R^2$ score, approximately 0.815, suggests that the model explains around 81.5% of the variance in final grades. This indicates a strong predictive capability of the model.

Feature Coefficients:

- Medu (Mother's Education) and Fedu (Father's Education) Coefficients: Both Medu and Fedu have negative and positive coefficients, respectively, indicating that higher levels of education for parents tend to have differing impacts on the final grades of students.
- G1 (First Period Grade) and G2 (Second Period Grade) Coefficients: Both G1 and G2 have positive coefficients, with G2 having a substantially higher coefficient. This suggests that the second period grade has a more significant influence on the final grade compared to the first period grade.
- Other Coefficients: Various other features such as study time, failures, alcohol consumption, school type, address, reason for choosing school, belief in higher education, and internet availability also showed significant coefficients, indicating their impact on predicting final grades.

Intercept: The intercept term of the regression model was found to be approximately 12.12. This represents the expected value of the final grades when all predictor variables are set to zero.

# Classification

Dataset Two:

Confusion Matrix Analysis:

- True Positives (TP): 64 - This indicates the count of correctly predicted instances of students with a performance level of 0.0 - 2.9.
- True Negatives (TN): 76 - Represents the count of correctly predicted instances of students with a performance level of 3.0 - 4.0.
- False Positives (FP): 2 - Reflects the count of instances wrongly classified as having a performance level of 0.0 - 2.9 when they actually belong to the 3.0 - 4.0 category.
- False Negatives (FN): 4 - Shows the count of instances wrongly classified as having a performance level of 3.0 - 4.0 when they actually belong to the 0.0 - 2.9 category.

Classification Report:

- Precision: The model achieved a precision of 94% for class 0.0 - 2.9 and 97% for class 3.0 - 4.0. This indicates a high proportion of correctly classified instances among those predicted positive.
- Recall: The recall, or sensitivity, was 97% for class 0.0 - 2.9 and 95% for class 3.0 - 4.0. This indicates the proportion of actual positive instances that were correctly classified.
- F1-score: Both classes have F1-scores of 96%, indicating a balance between precision and recall.
- Accuracy: The overall accuracy of the model was 96%, indicating the proportion of correctly classified instances among the total instances.

Dataset Three:

Scatter Plot Analysis:

- The scatter plot illustrates the distribution of 'degree_p' (percentage of undergraduate degree) and 'mba_p' (MBA percentage) features among placed and not placed students.
- Placed students generally exhibit higher values of both 'degree_p' and 'mba_p' compared to not placed students. However, there is noticeable overlap between the two classes, indicating that these features alone may not be sufficient for accurate classification.

Optimal Number of Neighbors:

- Cross-validation was performed to determine the optimal number of neighbors (k) for the KNN algorithm.
- A range of odd values for k from 1 to 29 was evaluated, and the mean accuracy scores were calculated for each value of k using 7-fold cross-validation.
- The highest mean accuracy was achieved with k=17, indicating that considering 17 nearest neighbors provides the best balance between bias and variance in the model.

Model Performance:

- After selecting the optimal k value, the KNN model was trained on the training set and evaluated on the test set.
- The accuracy of the KNN model on the test set was found to be approximately 75.58%.
- This accuracy score suggests that the model is able to correctly classify about 75.58% of students' placement status based on their 'degree_p' and 'mba_p' features alone.

Interpretation of Accuracy:

- While achieving an accuracy of 75.58% is notable, it's essential to interpret this metric in the context of the problem domain.
- The accuracy score indicates the proportion of correctly classified instances out of the total instances in the test set.
- However, it does not provide insights into the model's ability to correctly predict specific classes or potential biases in the predictions.

# Conclusions

Based on the comprehensive analysis conducted, several key findings emerge regarding factors influencing academic performance and student outcomes. Firstly, concerning student weight and absenteeism, the sampled groups appear to be representative of the population, suggesting no significant deviations in these aspects. However, further investigation or larger sample sizes could bolster confidence in these conclusions.

Regarding educational attainment, completing an MBA program demonstrates a significant positive impact on academic performance compared to undergraduate studies. This has implications for individuals considering further education and institutions designing their curriculum

Regression analyses underscore the multifaceted nature of academic success, indicating that factors such as socio-demographic background, parental education, study time, and previous academic performance significantly influence final grades. While these models demonstrate strong predictive capabilities, there's room for improvement to account for unexplained variance.

Classification models, particularly Gaussian Naive Bayes and K-Nearest Neighbors, show promise in predicting student performance levels and placement status. However, both models could benefit from further optimization, feature engineering, and exploration of additional variables to enhance their accuracy and practical applicability.

# Acknowledgements

The work in the project was distributed as follows:

- Datasets one and four visualization were done by Ahmed Alaroussy
- Datasets two and three visualization were done by Omar Malek
- Datasets one and four hypothesis were done by Kareem Abdelhameed
- Dataset three hypothesis was done by Mustafa Abdelkawy
- Datasets one and four regression analysis were done by Ahmed Shebl
- Datasets two and three classification were done by Mustafa Khairy

Each member was responsible for reporting their own part of the project.

# References


S, A. R. N. (2023, January 11). *Python Pandas Data Analysis*. Notebook by Aakash Rao N S (Aakashns) | Jovian. https://jovian.ai/aakashns/python-pandas-data-analysis

Python, R. (2023, March 17). *Develop Data Visualization Interfaces in Python With Dash*. https://realpython.com/python-dash/

L. A. J. (2021, January 24). *Python Interactive Dashboards with Plotly Dash - Quick Tutorial*. YouTube. https://www.youtube.com/watch?v=UYGwgHhazMA

Bobbitt, Z. (2021, September 27). *How to Perform One Sample & Two Sample Z-Tests in Python*. Statology. https://www.statology.org/z-test-python/

On, T. T. J. (2024, February 4). *Hypothesis Testing with Python: T-Test, Z-Test, and P-Values (+Code Examples)*. Medium. https://medium.com/@techtoy2023/hypothesis-testing-with-python-t-test-z-test-and-p-values-code-examples-fa274dc58c36#:~:text=Z%2Dtest%20is%20used%20to,population%20standard%20deviation%20is%20known.

R, L. (2022, July 26). *T-Test -Performing Hypothesis Testing With Python*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/07/t-test-performing-hypothesis-testing-with-python/

Saxena, S. (2023, September 24). *What are Categorical Data Encoding Methods | Binary Encoding*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/08/types-of-categorical-data-encoding/

Badole, M. (2024, January 17). *Multiple linear regression : Definition , Example and Applications*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/05/multiple-linear-regression-using-python-and-scikit-learn/

*3.1. Cross-validation: evaluating estimator performance*. (n.d.). Scikit-learn. https://scikit-learn.org/stable/modules/cross_validation.html

*1.9. Naive Bayes*. (n.d.). Scikit-learn. https://scikit-learn.org/stable/modules/naive_bayes.html

G. (2023, January 11). *k-nearest neighbor algorithm in Python*. GeeksforGeeks. https://www.geeksforgeeks.org/k-nearest-neighbor-algorithm-in-python/

Dataset 1: https://www.kaggle.com/datasets/susanta21/student-attitude-and-behavior

Dataset 2: https://www.kaggle.com/datasets/shaikhabdulrafay03/student-data

Dataset 3: https://www.kaggle.com/datasets/meruvulikith/campus-selection-classification-dataset/data

Dataset 4: https://www.kaggle.com/datasets/larsen0966/student-performance-data-set