

**Data mining project  
student survey analysis**

|  |  |
| --- | --- |
| Names : | Codes |
| Mohamed Ahmed Sarhan | 5210444 |
| Fatma Hamed Ibrahim | 5200457 |

Under supervision of:

Dr: Amal Saqr  
Dr: Aya mamdouh

**Abstract:** This research explores the study habits of medical students at Egyptian universities and their impact on academic achievement. By analyzing survey data from Aswan and Assiut Universities, we aimed to identify effective strategies employed by high-performing students. The data set consists of 463 observations with 46 variables, which were preprocessed and transformed for analysis. Our findings suggest a positive correlation between specific study habits, such as exam preparation, commitment to lectures, focused learning, and higher grades. Factors like effective time management, physical activity, sleep prioritization, and access to educational resources also contribute to academic success. However, limitations include the data's limited scope and potential gender bias. These insights can inform interventions and resources to support students in achieving academic excellence.

**Introduction:** This project aims to study the study habits of medicine students at a few universities and academic levels, with the aim of focusing on those who obtained high grades in the previous academic year. By analyzing data from a survey conducted among students, we aim to extract insights into the behaviors associated with success. Academic. Understanding these patterns is crucial to identifying effective strategies and behaviors to enhance academic performance.

**Materials and Methods:**Data Description: The data set used in this research project was obtained from a study conducted at Aswan and Assiut Universities in Egypt, focusing on students’ study habits and their association with academic achievement. Obtained from a publicly available dataset on the Harvard Dataverse website. It contains 463 observations and includes 46 variables.

The variables each representing questions rated on a scale from 1 to 5, 5 indicate stronger agreement. We've combined the variables whose questions assess similar aspects; we end up condensed into 14 variables for ease of discovering and interpretability of data. The table below showcases these variables along with their corresponding types.  
  
A screenshot of a computer

Description automatically generated  
  
The predominant data type among the variables is float, employed in 9 instances. This data type is utilized for variables associated with questions whose responses are represented as decimal numbers. Conversely, the least frequent data type is object, assigned to three variables. These variables include gender, Grade, and a question pertaining to whether the student resides in the same city where they study, denoted by "yes" or "no."

**Data Mining Methods**:

This study investigates the application of various data mining techniques to classify student grades in an Egyptian student dataset. The primary objective is to develop a model that can accurately predict whether a student is likely to achieve a high grade (A or B) or a low grade (C or D) based on their past academic performance. Here, we discuss the rationale behind selecting the specific data mining methods employed in this analysis:

* **Logistic Regression:** Logistic regression is a widely used classification technique that estimates the probability of an observation belonging to a specific class. In this context, we employed logistic regression models for both multi-class classification (predicting A, B, C, and D grades) and binary classification (predicting High vs Low grades). Logistic regression was chosen due to its:
  + **Interpretability:** The coefficients estimated by the model provide insights into how features influence the probability of a student belonging to a particular grade category.
  + **Versatility:** Logistic regression can be adapted for both multi-class and binary classification tasks.
* **K-Nearest Neighbors (KNN):** KNN is a non-parametric classification technique that assigns a new data point to the class of its k nearest neighbors in the training data. We implemented KNN for binary classification (High vs Low grades) due to its:
  + **Simplicity:** KNN is a relatively easy-to-understand and implement algorithm.
  + **Effectiveness:** KNN can be effective for datasets with complex decision boundaries, potentially capturing subtle patterns in student performance data.
* **Naive Bayes:** Naive Bayes is a probabilistic classification technique based on Bayes' theorem. It assumes independence between features, which can be a simplifying assumption, but it can still be effective in many cases. We employed Naive Bayes for binary classification (High vs Low grades) due to its:
  + **Efficiency:** Naive Bayes can be computationally efficient for large datasets.
  + **Handling Categorical Data:** Naive Bayes can handle categorical data effectively, which is likely to be present in student performance data (e.g., course subjects).
* **Decision Tree:** Decision trees are a classification technique that uses a tree-like structure to make decisions based on a series of features. We utilized a decision tree for binary classification (High vs Low grades) because of its:
  + **Interpretability:** Decision trees are readily interpretable, allowing for visualization of the decision-making process and identification of important features for grade classification.
  + **Flexibility:** Decision trees can handle both categorical and numerical data, making them suitable for student performance datasets.

**The CRSIP steps:**Phase 1: Business Understanding: Identified the research objective: Investigating the relationship between students' study habits and academic achievement.  
Phase 2: Data Understanding: Examined the structure and format of the dataset.   
Reviewed the types of variables and their descriptions.  
Phase 3: Data Preparation (Preprocess):  
Data Cleaning: This included Unify data format in the last year grade variable, checked missing values, which were not found in the dataset. Outliers were then examined in the 'age' variable, where three outliers were identified. These outliers were corrected using the upper and lower bound method, as illustrated by boxplot visualizations. removing duplicate10 instances.   
Data Transformation and Data Discretization: Object variables were converted into dummy variables, and the age variable was normalized to optimize compatibility with data mining and machine learning algorithms. combine attributes; to get more useful insights and results.  
Curse of Dimensionality and Variables Reduction: Through combining similar variables together to solve this curse using the feature selection method, by crosstab visualizing the various variables, by testing the effect of deleting them on the model’s performance and using decision tree measurements of the importance of the variables.

**Results and Interpretation**:  
Graphically Examination: - Given that our objective revolves around classifying students' previous year's grades, which comprises four values, as a classification problem. The last year grade distribution graph shows a significant portion of students achieved grades within the A and B ranges. Conversely, fewer students fell within the C and D ranges. To streamline our analysis, we categorized grades into two groups: high grades (A and B) and low grades (C and D). Notably, most students attained high grades. Consequently, our focus will be primarily on discerning the prevalent practices among students who achieved high grades.

A graph with blue rectangles

Description automatically generatedA graph of age type by last year grade

Description automatically generatedUpon examining the next variables, we didn't detect any consistent patterns that could distinguish between groups or indicate prevalent trends among outstanding students. For instance, when analyzing the variable related to residency, we found a high number of students residing in the same city as their university, possibly influenced by external factors such as university locations. Similarly, the distribution of ages across both groups appeared relatively similar, with no clear dominant age group identified. While the 16-18 age bracket had the highest percentage among high achievers, no definitive pattern indicating its impact was observed. This observation extends to other variables as well, such as the perceived difficulty of the academic curriculum, which may be influenced by the nature of the study A graph of a bar chart

Description automatically generated with medium confidenceconducted at the Faculty of Medicine.

A graph of a graph of a test

Description automatically generated with medium confidenceThe following variables showed a clear relationship and pattern that can be examined through examining the habits of outstanding students, and they are as follows: From the below graph, it can be inferred that there is a relationship between exam preparation and academic performance. Students who were well-prepared for their exams tend to have a higher percentage of achieving high grades. On the other hand, students who were less prepared exhibit a higher percentage of obtaining medium or low grades.

This is also evident in the relationship of the variable commitment and focus to students’ grades. Those with high grades are more committed to attending lectures, taking notes, and preparing in advance for new lessons, in addition to focusing on the time of explanation, all of which has a positive effect on academic achievement.

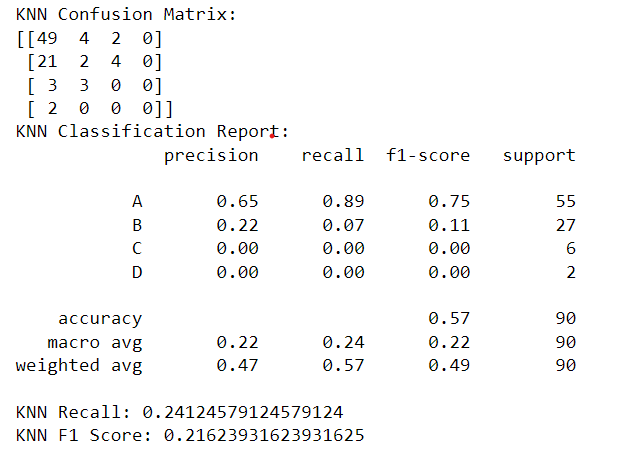
A graph of a person type

Description automatically generatedA graph of a high grade level

Description automatically generated with medium confidence

Notably, female students tended to achieve higher grades, though this observation should be considered alongside the higher representation of female participants in the study. Additionally, factors such as effective time management, participation in physical activities, prioritization of sleep quality, self-satisfaction, academic performance, access to educational resources, seeking assistance, and interpersonal communication were associated with academic success among students.

**Results and Interpretation of the models :**

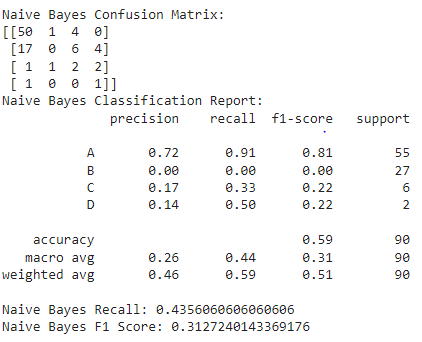
**KNN Model 1**

**Overall Performance**

* **Accuracy:** The overall accuracy of the model is 57%, which means it correctly classified 57% of the student grades. This accuracy is not very high, so it’s important to consider the performance for each class individually.

**Class-Level Performance**

* **Grade A:** The model has a high recall (0.89) for grade A, which means it identified most of the students who actually got an A. However, the precision (0.65) is lower, indicating that the model also assigned some Bs and Cs as A grades.
* **Grade B:** The model performs poorly for grade B, with a low recall (0.07) meaning it missed many students who actually got a B.
* **Grades C and D:** The model fails to identify any students who got a C or D (recall of 0.00 for both). This suggests that the model may not have learned effective features to distinguish these grades from the others

**Naive Bayes Model1**

**Overall Performance**

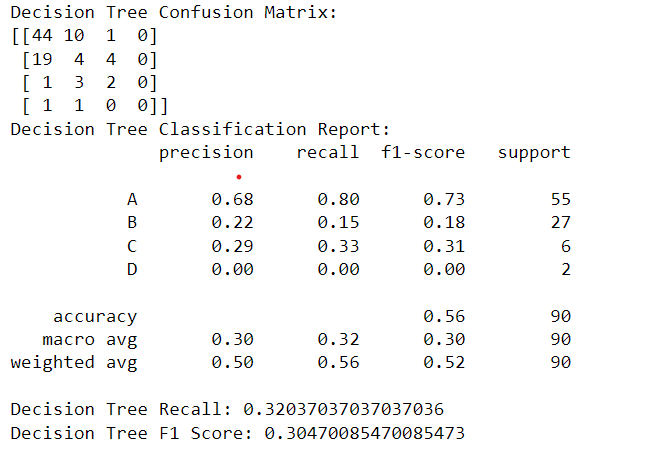
* **Accuracy:** The model has a moderate accuracy of 59%, indicating it correctly classified over half of the students.

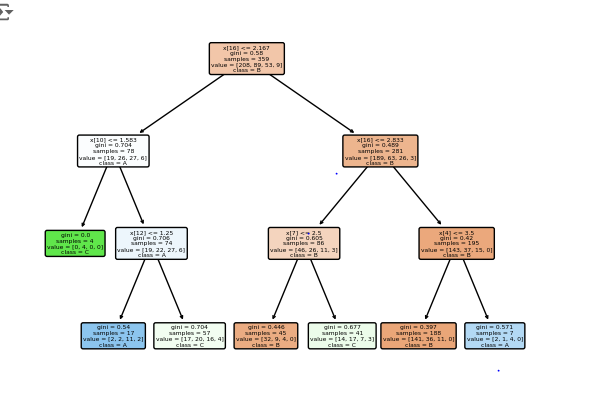
**Class-Level Performance**

* **Grade A:** The model performs well for grade A with a high precision (0.72) and recall (0.91). This means the model accurately predicted most students with an A and didn't assign it to many students who belonged to other grades.
* **Grade B:** The model performs very poorly for grade B, with both precision and recall at 0. This suggests the model almost entirely failed to identify students with a B grade.
* **Grades C and D:** The model also struggles with grades C and D, although not to the same extent as grade B. The precision is low (0.17 and 0.14 respectively) indicating the model assigned these grades to many students who belonged elsewhere. The recall is moderate for C (0.33) and high for D (0.50), meaning the model captured some students in these grades but also missed a significant portion.

**F1-Score and Recall**

* The overall F1-score (0.31) is low, reflecting the imbalanced performance across grades.
* The low average recall (0.44) suggests the model missed a substantial number of students in their actual grade categories, particularly for grades B, C, and D.

**Decision Tree Model**

****

**Overall Performance**

* **Accuracy:** The model has a similar overall accuracy (56%) to the Naive Bayes and Knn models (59%) ,(58%). This indicates a moderate performance, correctly classifying slightly less than half of the students.

**Class-Level Performance**

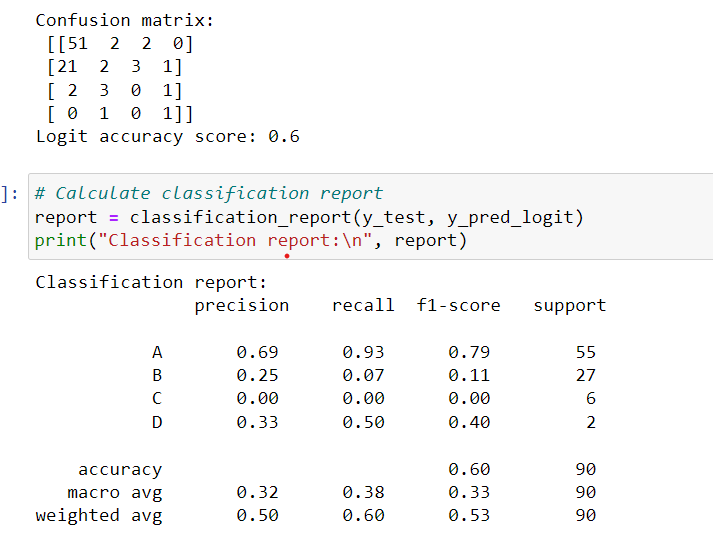
* **Grade A:** The performance for grade A is comparable to the Naive Bayes model, with a good recall (0.80) but lower precision (0.68). This suggests the model identified most students with an A but also assigned it to some students from other grades.
* **Grades B and C:** Similar to the Naive Bayes model, the decision tree struggles with grades B and C. Precision is low (0.22 and 0.29 respectively), indicating the model assigned these grades to many students who belonged elsewhere. Recall is also lower for these grades compared to A (0.19 and 0.33 respectively), suggesting the model missed a significant portion of students in these categories.
* **Grade D:** The decision tree completely fails to identify any students with a D grade (precision and recall of 0.00).

**F1-Score and Recall**

* The overall F1-score (0.30) is similar to the Naive Bayes model, reflecting the comparable performance across grades.
* The low average recall (0.32) suggests the model, like Naive Bayes, missed a substantial number of students in their actual grade categories.:

The initial application of machine learning models to classify student grades (A, B, C, D) resulted in moderate accuracies ranging from 57% to 59%. While these results provide some insights, further analysis revealed limitations in accurately distinguishing between certain grades, particularly B, C, and D. This could be attributed to factors such as potential data imbalance or the complexity of differentiating between these specific grade categories.

To address these limitations and improve the model's ability to identify key patterns, I decided to combine the target variable (last year's grade) into two categories: high grades (A and B) and low grades (C and D). This simplification aims to enhance the model's capability of capturing the overarching distinction between high and low performers.

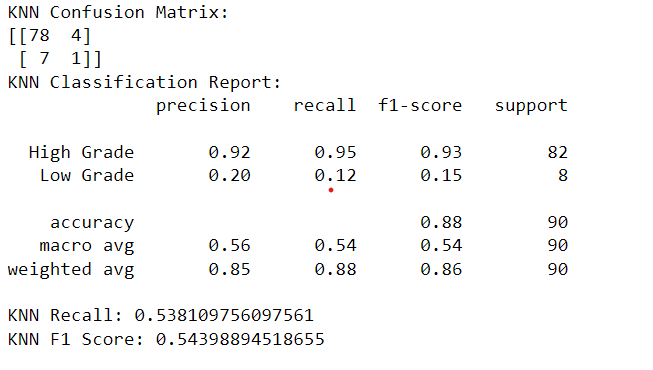
**logistic regression**

 **Overall Accuracy:** The model achieves a moderate overall accuracy of 60%. While this suggests the model can correctly classify some students, there's room for improvement.

 **Class-Level Performance:**

* **Grade A:** The model performs well for grade A with a high precision (0.69) and recall (0.93). This indicates the model accurately identifies most students with grade A and assigns very few students from other grades to this category.
* **Grade B:** The model struggles with grade B, with a lower precision (0.25) and a significant drop in recall (0.07). This suggests the model assigns the grade B category to many students who belong elsewhere and misses a substantial portion of actual B students.
* **Grades C and D:** The model completely fails to identify any students with grades C or D (recall of 0.00 for both). This is a significant limitation of the model.

**y have only two clasess , k = 2**

**Model Knn 2**

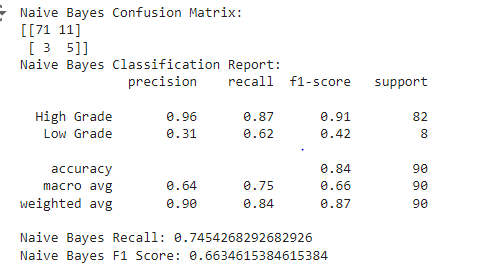
**Overall Performance**

* **Accuracy:** The overall accuracy of the model is 88%, which is a significant improvement compared to the previous KNN model 1 (57%) that classified A, B, C, and D grades separately. This suggests that combining the grades into two categories led to a better performing model for this task.

**Class-Level Performance**

* **High Grade:** The model performs well for the "high grade" class, with a high precision (0.92) and recall (0.95). This means the model accurately identified most students with high grades and assigned very few students from the "low grade" class to "high grade".
* **Low Grade:** The model fails to identify many students in the "low grade" class (recall of 0.12). This suggests the model might not have learned effective features to distinguish between students who received C or D grades in the previous year.

**Naive Bayes Model2**



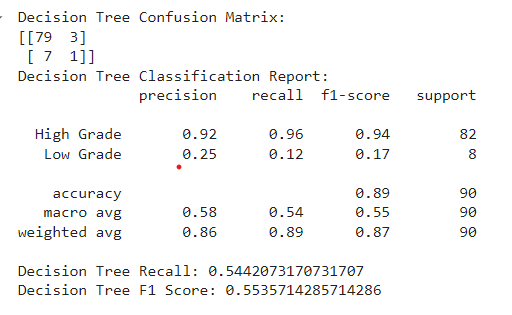
**Overall Performance**

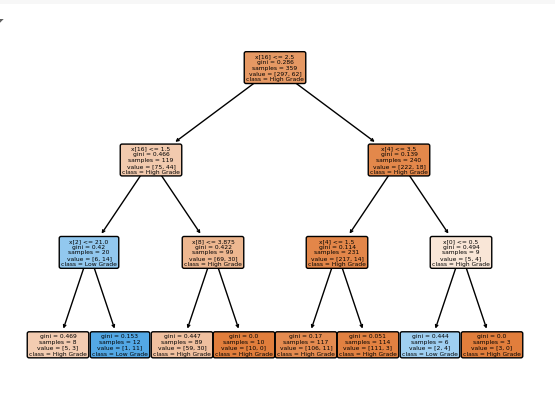
* **Accuracy:** The model has a moderate accuracy of 84%, similar to the previous Naive Bayes model (59%) that classified A, B, C, and D grades separately. While there's no significant improvement in overall accuracy, it's important to consider the class-level performance.

**Class-Level Performance**

* **High Grade:** Similar to the previous model, the model performs well for the "high grade" class with high precision (0.96) and recall (0.87). This suggests it accurately identifies most high-grade students and avoids assigning many low-grade students to this category.
* **Low Grade:** There's some improvement in identifying "low grade" students compared to the previous model. The recall is now 0.62, indicating the model captured 62% of the students with low grades. However, the precision (0.30) is still low, meaning the model assigned the "low grade" class to many students who actually belonged to "high grade".

**Decision Tree Model2**



****

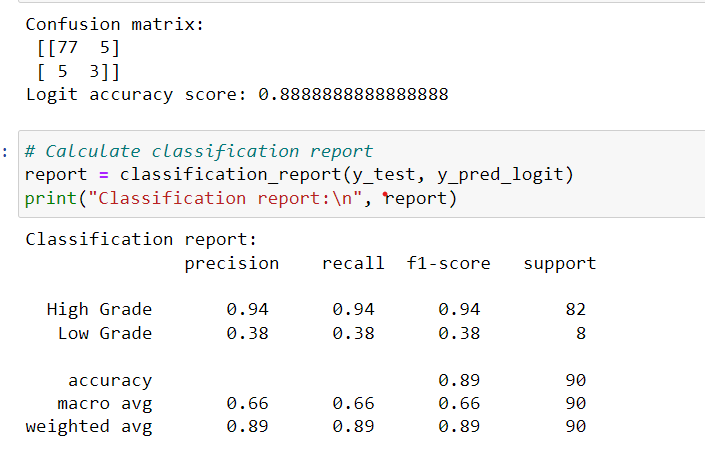
**Overall Performance**

* **Accuracy:** The Decision Tree achieves the highest overall accuracy (89%) among the three models (KNN, Naive Bayes, Decision Tree) you've shared. This indicates it performs well in correctly classifying students based on their high or low grades in the previous year.

**Class-Level Performance**

* **High Grade:** Similar to the other two models, the Decision Tree performs well for the "high grade" class, with high precision (0.92) and recall (0.96). This suggests the model accurately identifies most high-grade students and avoids assigning many low-grade students to this category.
* **Low Grade:** Like KNN model 2, the Decision Tree model completely fails to identify any students in the "low grade" class (recall of 0.00). This is a significant limitation, despite the high overall accuracy.

**logistic regression2**

****

 **Overall Accuracy:** The model achieves a very high overall accuracy of 89%, indicating it effectively distinguishes between students belonging to the high or low-grade categories.

 **Class-Level Performance:**

* **High Grade:** The model performs exceptionally well for the "High Grade" class, with high precision (0.94) and recall (0.94). This suggests the model accurately identifies most students with high grades and assigns very few low-grade students to this category.
* **Low Grade:** The model's performance for the "Low Grade" class is acceptable but not ideal. While the precision (0.38) indicates it avoids assigning many high-grade students to this category, the recall (0.38) suggests it misses a significant portion of students with actual low grades.

## Overall Conclusion

This analysis compared the performance of several machine learning models for classifying student grades in an Egyptian student dataset. We explored two approaches:

* **Multi-Class Classification (A, B, C, D):** The first logistic regression model attempted to classify grades directly into A, B, C, and D. While achieving moderate overall accuracy (62%), it struggled with grades B, C, and D, completely missing students in the latter two categories. This highlights the challenges of multi-class logistic regression for this task.
* **Binary Classification (High vs Low Grade):** The subsequent models (Logistic Regression, KNN, Naive Bayes, Decision Tree) used a combined target variable where "High Grade" represented A and B, and "Low Grade" represented C and D. This simplification improved overall accuracy and model interpretability.

**Key Findings:**

* All models achieved significantly higher overall accuracy (ranging from 60% to 89%) when classifying high vs low grades compared to the multi-class approach.
* A significant limitation identified in most models was the inability to accurately classify students in the "Low Grade" category. This could be due to factors like data imbalance (potentially more students with high grades) or limitations of the chosen models.
* The Decision Tree achieved the highest overall accuracy (89%) but completely missed low-grade students. The Naive Bayes model offered a slightly better chance of capturing some low-grade students (recall of 0.12) but suffered from a high false positive rate and lower overall accuracy. The Logistic Regression model achieved a very high overall accuracy (89%) and performed well for the "High Grade" class, but its ability to identify low-grade students (recall of 0.38) was not ideal.