

 The Campus of Tomorrow

Higher
Colleges of
Technology



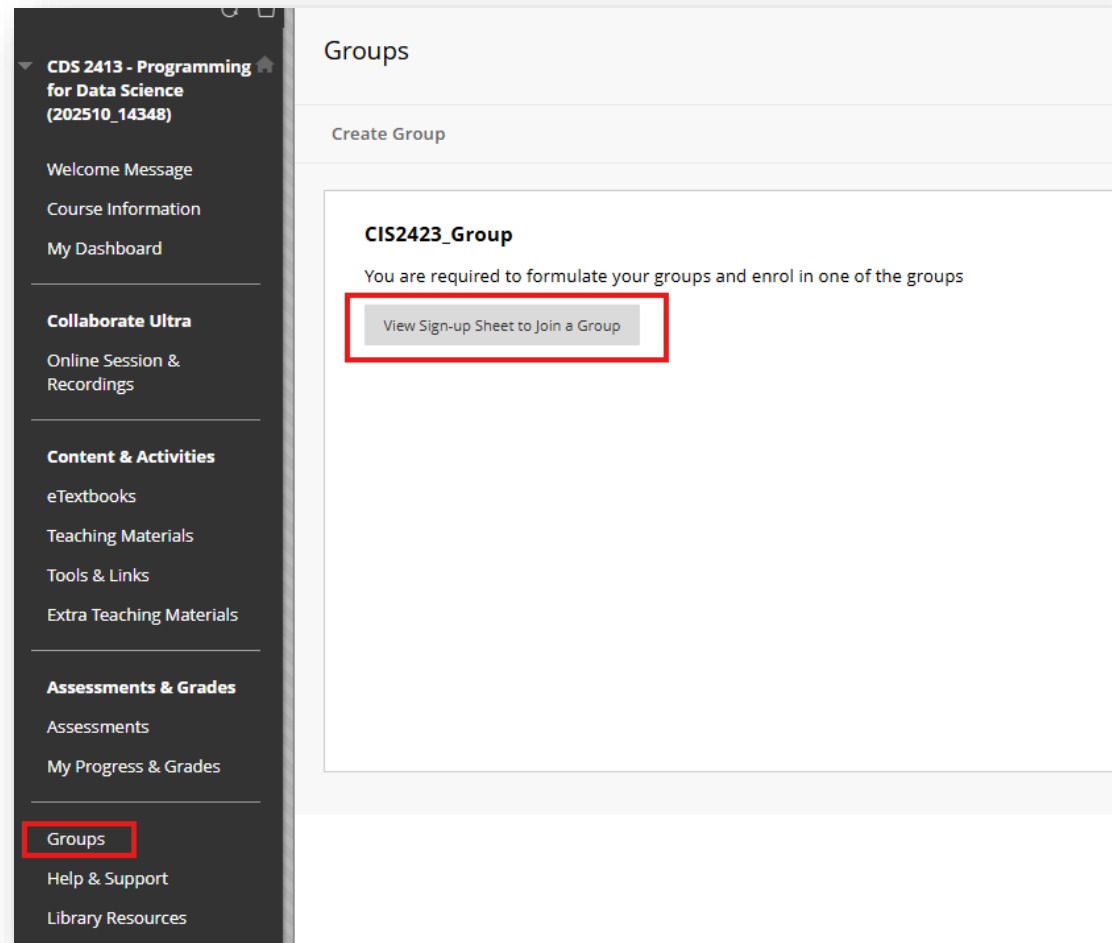
كليات
التقنية
العليا

PROJECT

CDS2413

FORM GROUPS

- 2 Students in Each group



The screenshot displays a user interface for a course management system. On the left is a dark sidebar with a list of navigation items. The 'Groups' item is highlighted with a red rectangular box. The main content area on the right is titled 'Groups' and includes a 'Create Group' button. Below this, a section titled 'CIS2423_Group' contains the text 'You are required to formulate your groups and enrol in one of the groups'. A button labeled 'View Sign-up Sheet to Join a Group' is highlighted with a red rectangular box.

CDS 2413 - Programming for Data Science (202510_14348)

- Welcome Message
- Course Information
- My Dashboard

Collaborate Ultra

- Online Session & Recordings

Content & Activities

- eTextbooks
- Teaching Materials
- Tools & Links
- Extra Teaching Materials

Assessments & Grades

- Assessments
- My Progress & Grades

Groups

- Help & Support
- Library Resources

Groups

Create Group

CIS2423_Group

You are required to formulate your groups and enrol in one of the groups

[View Sign-up Sheet to Join a Group](#)

CHOOSING A DATASET

- When you pick a dataset, think of it like choosing the right puzzle to solve. The pieces need to fit together so you can practice all the steps of data analysis. Here's what to look for:

It should have enough data

- Look for datasets with at least a few hundred rows.
- If it's too small, your models won't make sense.
- Example: A dataset with only 20 people's test scores is too small. But one with 1,000 students' scores is good.

UAE Dataset

Data Card Code (0) Discussion (0) Suggestions (0)

UAE.csv (18.93 MB)

Detail Compact Column 10 of 17 columns

Type	Title	# Area	# Bedrooms	# Bathrooms	# Price
Apartment 64%					
Villa 19%	65662 unique values				
Other (14233) 17%		150 240k	1 21	1 30	1000
Villa	Beach Pool Facing I Limited Offer @ A La Carte	2350	4	5	2049989.
Villa	Beach Pool Facing I Limited Offer @ A La Carte	2350	4	5	2049989.
Villa	Beach Pool Facing I Limited Offer @ A La Carte	2350	4	5	2049989.
Villa	Beach Pool Facing I Limited Offer @ A La Carte	2350	4	5	2049989.
Villa	A La Carte Amazing Villas High-End Finishing	2350	4	5	2100000.
Villa	A La Carte Amazing Villas High-End Finishing	2350	4	5	2100000.

CHOOSING A DATASET

It should have the right kind of columns


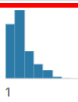


- You need:
 - One main outcome (dependent variable) → something you want to predict (e.g., exam score, house price, survival on Titanic).
 - Several input features (independent variables) → things that might affect the outcome (e.g., study hours, house size, age, gender).
- Avoid datasets with just one or two columns — you won't be able to do much with them.

UAE Dataset

Data Card Code (0) Discussion (0) Suggestions (0)

UAE.csv (18.93 MB)

Detail Compact Column 10 of 17 columns

Type	Title	Area	Bedrooms	Bathrooms	Price
Apartment 64%	65662 unique values				
Villa 19%		150 240k	1 21	1 30	1000
Other (14233) 17%					
Villa	Beach Pool Facing I Limited Offer @ A La Carte	2350	4	5	2049989.
Villa	Beach Pool Facing I Limited Offer @ A La Carte	2350	4	5	2049989.
Villa	Beach Pool Facing I Limited Offer @ A La Carte	2350	4	5	2049989.
Villa	Beach Pool Facing I Limited Offer @ A La Carte	2350	4	5	2049989.
Villa	A La Carte Amazing Villas High-End Finishing	2350	4	5	2100000.
Villa	A La Carte Amazing Villas High-End Finishing	2350	4	5	2100000.

CHOOSING A DATASET

It should be clean and understandable

- Columns should have clear names and meanings (e.g., “Age,” “Salary” instead of “col1,” “col2”).
- Avoid datasets full of missing values or strange symbols.
- If you can look at the first few rows and quickly understand what’s being measured, it’s a good sign.

UAE Dataset

Data Card Code (0) Discussion (0) Suggestions (0)

UAE.csv (18.93 MB)

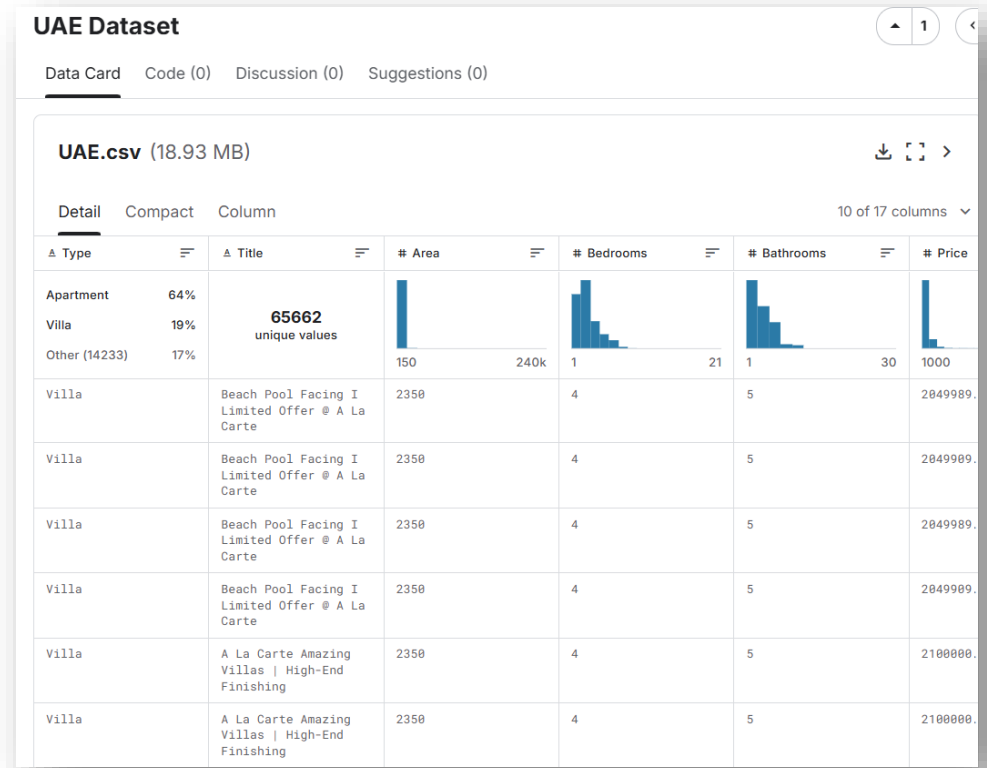
Detail Compact Column 10 of 17 columns

Type	Title	# Area	# Bedrooms	# Bathrooms	Price
Apartment 64%	65662 unique values				
Villa 19%		150 240k	1 21	1 30	1000
Other (14233) 17%					
Villa	Beach Pool Facing I Limited Offer @ A La Carte	2350	4	5	2049989
Villa	Beach Pool Facing I Limited Offer @ A La Carte	2350	4	5	2049989
Villa	Beach Pool Facing I Limited Offer @ A La Carte	2350	4	5	2049989
Villa	Beach Pool Facing I Limited Offer @ A La Carte	2350	4	5	2049989
Villa	A La Carte Amazing Villas High-End Finishing	2350	4	5	2100000
Villa	A La Carte Amazing Villas High-End Finishing	2350	4	5	2100000

CHOOSING A DATASET

It should allow different analyses

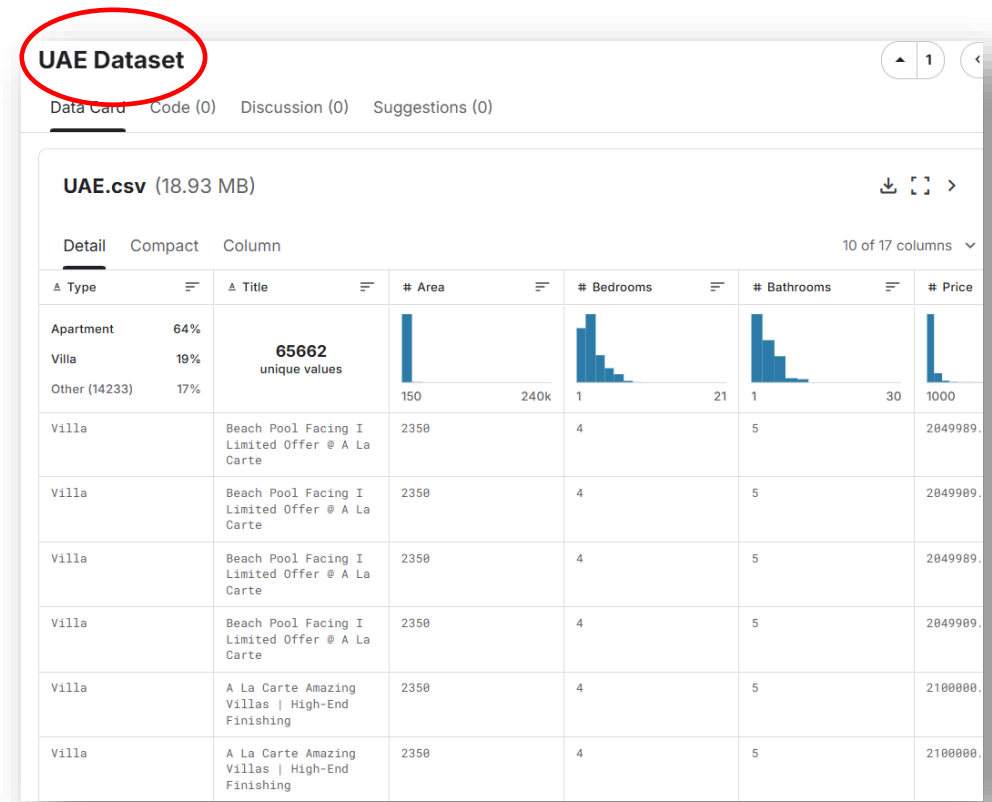
- The dataset must be rich enough to let you practice:
 - Descriptive statistics (averages, counts, min/max).
 - Visualizations (scatter plots, histograms, box plots).
 - Regression (predicting numbers, like house price).
 - Classification (predicting categories, like pass/fail).
 - Clustering (grouping things, like customer types).
- Example: The Titanic dataset is great because you can predict survival (classification), check ages and fares (descriptive), and make graphs.



CHOOSING A DATASET

It should interest you

- If you're curious about the topic, you'll enjoy working with it more.
- Ideas:
 - Sports stats (player performance, match outcomes).
 - Health data (exercise, diet, disease rates).
 - Finance (house prices, salaries, sales).
 - Education (student grades, study habits).
- Bonus if your data is related to UAE!



EXAMPLES

- **UAE Cancer Patient Dataset** – Synthetic health records of over 10,000 cancer patients in the UAE, including demographics and medical history.
- **UAE Real Estate 2024 Dataset** – Property listings in Dubai with details like location, property type, size, and prices.
- **UAE Used Car Prices & Features (10k Listings)** – Car listings in the UAE with information on make, model, year, mileage, and prices.
- **Dubai Real Estate Goldmine – UAE Rental Market** – Rental property data across Dubai, Abu Dhabi, and Sharjah, including features and rental prices.
- **UAE Population by Emirate, Nationality and Gender** – Demographic data from 1975 to 2005 showing population by emirate, nationality, and gender.

TEAMS

- **UAE Cancer Patient Dataset** – Synthetic health records of over 10,000 cancer patients in the UAE, including demographics and medical history.
- **UAE Real Estate 2024 Dataset** – Property listings in Dubai with details like location, property type, size, and prices.
- **UAE Used Car Prices & Features (10k Listings)** – Car listings in the UAE with information on make, model, year, mileage, and prices.
- **Dubai Real Estate Goldmine – UAE Rental Market** – Rental property data across Dubai, Abu Dhabi, and Sharjah, including features and rental prices.
- **UAE Population by Emirate, Nationality and Gender** – Demographic data from 1975 to 2005 showing population by emirate, nationality, and gender.

MULTIPLE DATASETS

■ CLO1, CLO2, CLO3 (Regression part)

- For all tasks related to **defining the dataset (CLO1)**, **descriptive analysis and visualization (CLO2)**, and **regression modeling (CLO3 – regression part)** → 👉 Students must use **the same dataset**.
- Example: If you pick the *UAE Housing 2024 dataset*, you will:
 - Explain the purpose of analyzing it (CLO1).
 - Do descriptive statistics, graphs, hypothesis testing (CLO2).
 - Build regression models to predict housing prices (CLO3).

■ CLO3 (Classification & Clustering part)

- For the **classification** and **clustering** tasks in CLO3, students can either:
 - Continue using the **same dataset** (if it has categorical variables/classes to predict).
 - Or, if their chosen dataset is not suitable for classification/clustering, they are allowed to pick a **different dataset**.
- Example:
 - UAE Historical Weather works for both regression and classification → can use the same one.
 - A **housing dataset** (prices are continuous) is great for regression, but not ideal for classification → here, students may choose a second dataset (Preferably related to housing).

REPORT

I	Define the purpose of data analysis for the chosen dataset	2
	Identify and Justify the type of programming used for data analysis	2
	Identify the type and purpose of the machine learning algorithm to be implemented for the chosen dataset	3
	Identify and Justify the independent and dependent variables for the chosen dataset.	3
	Will you do the sampling? Identify and justify the type of sampling to be used for the chosen dataset	

REPORT

2	I. Justify why you want to perform the descriptive analysis for the chosen dataset.	1
	I. Create a script to develop a Python function for descriptive statistics. The input for the function should be the sample and the field to perform the descriptive statistics.	1
	I. Create a program to random sampling of size 150 and find the descriptive statistics for the dependent variable from the sample [Apply the descriptive function which you created].	1
	I. Create a script for systematic sampling by giving certain conditions and finding the desc stat for the dependent variable from the sample [Apply the descriptive function which you created].	1
	I. Create a detailed descriptive statistics report about the dependent variable of the chosen dataset.	1
	I. Visualize the dependent variable by the Graph/Chart of the following using Python Program: a. Scatter plot b. Box Plot c. Histogram d. Heat Map Hint: Use Matplot or Ski-learn library	3
	I. Perform the hypothesis test to find the correlation (Pearson and Spearman for numerical variable and chi-square test for categorical variable) between the independent variable and the dependent variable.	1
	Note: If you have more than one independent variable, then choose any one of the independent variables.	
	I. Assess the performance of the dependent variable to know whether the sample is representative of the normal population by a one-sample t-test.	1
Total		10

REPORT

3	I. Build,Train, Develop and Evaluate using Simple Regression for chosen dataset.	5
	I. Develop a script to forecast the value of the dependent variable from all the relevant independent variables using Multiple Linear Regression	5
	I. Predict the value of the dependent variable from the different classifier such as Logistic Regression, KNN, Naïve-Bayes and Decision Tree.	17
	I. Evaluate the performance of each model using confusion matrix and accuracy and identify the best fit classifier for the chosen dataset.	9
	I. Predict the dependent variable by using best-fit classifier.	1
	I. Perform the cluster analysis such as K-means and Horizontal for any field from the chosen dataset.	8
	I. Explain the strategy for improving the system after viewing the cluster diagram.	2
	Total	42

REPORT

4	I. Create a new repo for project in Git Hub	3
	I. Upload all the project files created for CLO1,CLO2 and CLO3 to the Git Hub repo	4
	I. Configure Git with GitHub	5
	I. Clone Git hub repo to Git	4
	I. Pull any file from Git Hub repo to Git	5
	I. Modify the pulled file and push the modified file to Git Hub	5
	Total	26

COLLABORATION

The screenshot shows a Google Meet interface. On the left, a presentation slide from a Word document titled 'CIS 2423 Final Project' is displayed. The slide contains a list of classification results and a table comparing different models. On the right, a code editor window is open, showing Python code for a machine learning project. A chat bubble at the bottom of the Meet window contains text about versioning with git and github.

Classification Results:

- **True Positives (TP):** The classifier correctly predicted 300 houses as expensive.
- **True Negatives (TN):** The classifier correctly predicted 323 houses as cheap.
- **False Positives (FP):** The classifier incorrectly predicted 149 houses as expensive when they were cheap.
- **False Negatives (FN):** The classifier incorrectly predicted 129 houses as cheap when they were expensive.

The classifier's accuracy is approximately 69.15%, which means that it will predict the price categories correctly for the majority of cases.

17) Predict the dependent variable by using best-fit classifier:

Comparing the classification models in terms of accuracy, precision, recall, and error rates:

Model	Accuracy	Precision	Recall	Error
Logistic Regression Classifier	74.58%	73.92%	72.05%	25.42%
KNN Classifier	70.48%	68.48%	70.40%	29.52%
Naive-Bayes Classifier	52.83%	62.50%	2.33%	47.17%
Decision Tree Classifier	69.15%	66.81%	69.93%	30.85%

After examining the confusion matrix we can say that the Logistic Regression model is the best-fit classifier for our chosen dataset (House_Data.csv). This is because it has the highest accuracy, lowest errors, and a good balance for both the precision and recall rates compared to the other classifiers.

Moreover, we used it to predict the value of the dependent variable for a given house record and it predicted it correctly as shown below:

Code Editor Content:

```
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report
from sklearn.metrics import accuracy_score
from sklearn.metrics import precision_score
from sklearn.metrics import recall_score
from sklearn.metrics import f1_score
from sklearn.metrics import roc_auc_score
from sklearn.metrics import log_loss
from sklearn.metrics import brier_score_function
from sklearn.metrics import jaccard_index_score
from sklearn.metrics import hamming_loss
from sklearn.metrics import zero_one_loss
from sklearn.metrics import cohen_kappa_score
from sklearn.metrics import matthews_correlation_coefficient
from sklearn.metrics import log_likelihood_ratio
from sklearn.metrics import log_loss
from sklearn.metrics import brier_score_function
from sklearn.metrics import jaccard_index_score
from sklearn.metrics import hamming_loss
from sklearn.metrics import zero_one_loss
from sklearn.metrics import cohen_kappa_score
from sklearn.metrics import matthews_correlation_coefficient
from sklearn.metrics import log_likelihood_ratio
```

Chat Bubble:

About Versioning part: we download git and sign up for github, create new repo

Page Footer:

Page 3 of 43 5946 words

POSTER

- At the end of the term, there will be a Posters competition.
- You will present your results as a poster.
- Judges + Voting

Introduction/Background

To explore how lifestyle and health factors influence sleep duration and quality using data analytics and machine learning.

Objectives:

- Clean and preprocess a real-world sleep health dataset.
- Apply descriptive and inferential statistics to uncover key trends.
- Build regression and classification models to predict sleep outcomes.
- Use clustering to identify patterns in lifestyle behaviors.
- Visualize findings for better understanding and communication

Analytical Methods & Key Results

Statistical & Hypothesis Tests

- Pearson & Spearman Correlation on academic factors
- Chi-Square on support and dropout categories
- One-Sample t-Test confirmed sample representativeness

Classification Models

- Logistic Regression: Accuracy = 88%
- K-Nearest Neighbors (KNN): Accuracy = 88%
- Decision Tree: Accuracy = 88%
- Naïve Bayes: Accuracy = 85%

Regression Analysis

- Multiple Linear Regression: $R^2 = 0.978$
- Admission Grade, Curriculum, Tuition Fees predicted dropout status

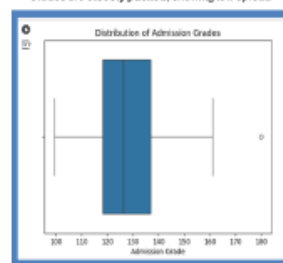
Classification Models

- K-Means: 3 clusters revealed patterns of risk
- Hierarchical Clustering confirmed similar groups

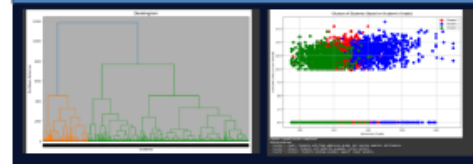
Clustering Analysis

- K-Means: 3 clusters revealed patterns of risk

- Box Plot – Admission Grade Distribution**
- Most scores fall between 120-140, median ~130
- A few outliers scored as high as 180
- Data is **right-skewed**, indicating more high scorers
- Grades are **closely packed**, showing low spread



Student Clustering using hierarchical clustering (agglomerative clustering)



Hierarchical clustering grouped students based on academic grades, confirming distinct performance clusters using a dendrogram and scatter plot.

Classification & Clustering Insights

Classification Summary:

- K-Nearest Neighbors (KNN) gave the highest accuracy (88.96%).
- Decision Tree also performed well (88.35%) with simple interpretability.
- Logistic Regression was less accurate (84%), but interpretable.
- Naïve Bayes had the lowest performance (81.6%).

KNN was selected as the best performing model for student dropout prediction.

Clustering Insights:

- 3 main clusters identified:
- Cluster 1 – High Performers
- Cluster 2 – At-Risk / Struggling
- Cluster 3 – Average Students
- Most At-Risk students dropped out
- 1 cluster of High Achievers (top 10%)
- Others mixed with good academic results
- Clustering helped in segmenting students by risk levels, aiding targeted interventions.

Data Source & Description

The dataset used in this project was sourced from Kaggle, titled *Sleep Health and Lifestyle Dataset*.

It contains 374 records with multiple health and lifestyle attributes including:

- Demographics:** Age, Gender, BMI Category
- Health Metrics:** Sleep Disorder (target), Sleep Duration, Quality of Sleep, Stress Level
- Lifestyle Factors:** Daily Steps, Physical Activity Level, Heart Rate, Occupation, Blood Pressure

The target variable is **Sleep Disorder**, categorized as *None*, *Insomnia*, or *Apnea*.

The dataset supports analysis of potential relationships between lifestyle factors and sleep quality/disorders.

Predictive Modeling and Key Findings

Classification Models (Dropout Prediction)

- Logistic Regression** – Accuracy: ~64%
Simple, interpretable; limited in handling complex patterns.
- K-Nearest Neighbors (KNN)** – Accuracy: ~70%
Best performer post-feature scaling.
- Decision Tree** – Accuracy: ~69.35%
Visual and easy to understand; can overfit.
- Naïve Bayes** – Accuracy: ~61.8%
Fast but limited precision due to assumptions.

Regression Insights

- Simple Regression**
 - Physical Activity → Sleep Duration: $R^2 = 0.36$
 - Stress Level → Sleep Quality: $R^2 = 0.61$
 - BMI → Sleep Disorder: $R^2 = 0.48$

Multiple Regression

- Works well for predicting Sleep Duration & Quality ($R^2 > 0.6$)
- Not suitable for predicting categorical outcomes like Sleep Disorder

Statistical Tests

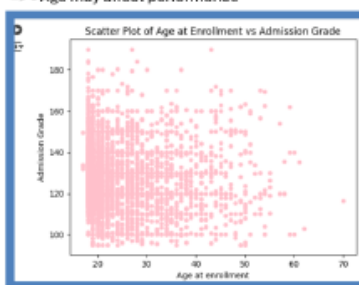
- Pearson & Spearman Correlation**
Found significant relationships (e.g., GPA vs Dropout)
- Chi-Square Test**
Significant link between academic factors and dropout
- One-Sample T-Test**
No statistically significant differences found in dropout indicators

Clustering Analysis

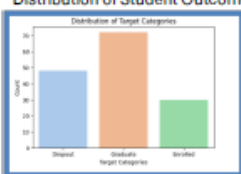
- K-Means (with PCA)**
Formed 3 distinct student clusters by dropout risk
- Hierarchical Clustering**
Confirmed the groupings visually using dendrograms
- Mean Shift**
Applied for adaptive, density-based clustering
- Naïve Bayes** – Accuracy: ~61.8%
Fast and efficient, but its strong assumptions can limit precision.

Scatter Plot - Age vs Admission Grade

- Older students show slightly lower grades;
- Most are aged 20-40; younger students show more variation.
- Some 50s-60s students are outliers.
- Age may affect performance



Distribution of Student Outcomes

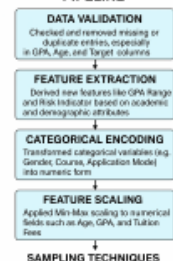


Spearman Correlation: Combined Academic Index vs Target
Weak correlation ($\rho = 0.25$), not statistically significant ($p = 0.85$).
No strong link between academic index and dropout status.



Data Preprocessing Pipeline

DATA PREPROCESSING PIPELINE



To ensure the dataset was clean, reliable, and suitable for machine learning, the following steps were applied:

Data Cleaning: Removed rows with missing or inconsistent values in critical fields like GPA and Target.

Encoding: Converted categorical variables such as Gender, Scholarship Holder, and Course into numeric format using label encoding.

Feature Scaling: Applied Min-Max scaling to standardize features like Age, GPA, and Tuition Fees to a 0-1 range.

Class Balancing: Used random sampling to split the dataset into training and testing sets, ensuring representation of both dropout and retained students.

GitHub Integration

Version control was demonstrated using GitHub and Colab.

The team pushed changes (e.g., plot edits) via standard Git commands (add, commit, push).

Notebook and code files were collaboratively managed.

Conclusion & Future Work

Conclusion:

- KNN and Decision Tree models yielded the highest accuracy (~70%) for predicting student dropout.
- Logistic Regression provided a simpler baseline, while Naïve Bayes underperformed.
- Clustering techniques revealed risk groups, and regression models ($R^2 > 0.97$) showed strong prediction capabilities for academic index.

Future Work:

- Integrate real-time data from institutional systems for ongoing dropout prediction.
- Use ensemble models or neural networks to enhance classification accuracy.
- Extend analysis to include psychological or engagement-related factors for deeper insights.
- Deploy the model as a web-based decision support tool for educators.

References

- Kaggle (2023). Sleep Health and Lifestyle Dataset. Retrieved from <https://www.kaggle.com/datasets>
- Scikit-learn Developers (2023). Scikit-learn: Machine Learning in Python. <https://scikit-learn.org>
- McKinney, W. (2010). Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference.
- Python Software Foundation. Python Language Reference, Version 3.9. <https://www.python.org>
- Seaborn & Matplotlib Documentation (2023). Python Data Visualization Libraries.

Introduction/Background

Retailers often lack real-time analytical insights on customer behavior and branch performance.

This project explores how **Python-based analytics** can uncover sales trends and optimize operations.

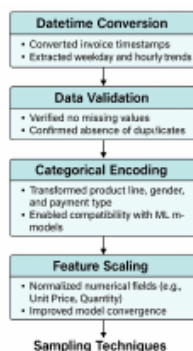
Objective: Apply data science techniques to analyze sales patterns, predict trends, and segment customers.

Data Source & Description

Dataset: Supermarket Sales from Kaggle.

- 1,000 records and 17 attributes, covering transactional data across multiple branches.
- Key fields: branch, product line, unit price, quantity, rating, gross income, and invoice date, offering a comprehensive view of customer behavior and store performance

Data Preprocessing Pipeline



- The flowchart highlights key steps used to prepare the dataset for analysis.
- Preprocessing ensured high data quality and model readiness.
- Enhanced model performance by addressing data consistency and scale
- Enabled effective application of statistical and machine learning techniques.

Analytical Methods & Key Results



Descriptive Statistics

- Analyzed sales spread, central tendency (mean, kurtosis) box plots)
- Identified outliers using box plot



Hypothesis Testing

- Pearson & Spearman correlation to assess relationships between continuous variables



Clustering

- K-Means: Segmented by Unit Price & Quantity (3 distinct shopper type types)



Predictive Modeling

- Simple Linear Regression ($r = 7.2$ between Unit Price & Gross Income ($R^2 = 0.72$)
- Multiple Linear Regression avoiding derived variables



Classification Models

- Naive Bayes Accuracy at 60%
- Logistic Regression at 58%
- KNN Decision Tree

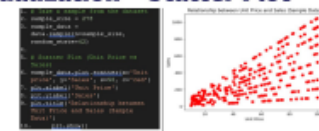


Clustering

- K-Means: Segmented by Unit Price & Quantity – 3 distinct shopper types VIP-dissatisfied

Modeling & Pattern Insights

Visualization – Scatter Plot



This plot helps visualize the relationship between the unit price of products and the total sales. The x-axis showing the unit price and the y-axis showing the sales amount. The points represent individual transactions.

Visualization – Box Plot & Histogram



Visualization – Heatmap (City vs Product Line)



Correlation & Hypothesis Testing

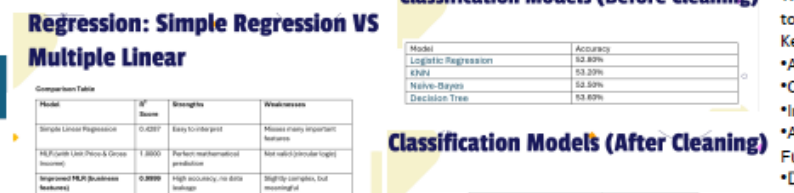


Modeling & Pattern Insights

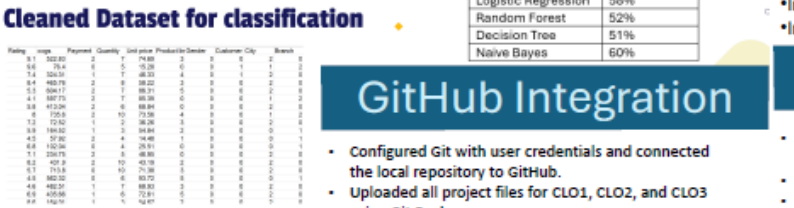
Correlation & Hypothesis Testing



Regression: Simple Regression VS Multiple Linear



Cleaned Dataset for classification



Classification Models (Before Cleaning)

Model	Accuracy
Logistic Regression	52.89%
KNN	53.29%
Naive Bayes	52.59%
Decision Tree	53.60%

Classification Models (After Cleaning)

Model	Accuracy
Logistic Regression	58%
Random Forest	52%
Decision Tree	51%
Naive Bayes	60%

GitHub Integration

- Configured Git with user credentials and connected the local repository to GitHub.
- Uploaded all project files for CLO1, CLO2, and CLO3 using Git Bash.
- Repository link: https://github.com/H00498539/project2423_1

Conclusion & Future Work

The project successfully demonstrates how Python-based data science techniques can be applied to retail analytics.

Key achievements include:

- Accurate regression models for sales prediction
 - Clear customer segmentation using clustering
 - Improved classification accuracy after data cleaning
 - Actionable business insights supporting decisions in promotions and inventory management
- Future Work**
- Develop a real-time dashboard using Streamlit
 - Extend the analysis with time-series forecasting
 - Integrate models with supermarket POS systems
 - Implement a recommender system for personalized marketing

References

- Kaggle. Supermarket Sales Dataset. <https://www.kaggle.com/datasets/aungpyaeap/supermarket-sales>
- Scikit-learn Documentation. <https://scikit-learn.org>
- Seaborn and Matplotlib Visualization Libraries
- Python for Data Analysis – Wes McKinney (O'Reilly)
- CIS2423 Course Materials – Higher Colleges of Technology