

King Saud University

College of Computer and Information Sciences

Information Technology Department

**IT326: Data Mining**

Project "Students Exam Scores**"**

|  |  |
| --- | --- |
| Section number: 56544 | ID |
| Raghad Aldajani | 443200551 |
| Rana Alharbi | 442202346 |
| Nora Albyahi | 443200479 |
| Mariam Ahmed | 443204628 |
| Sadeem albargash | 443200679 |

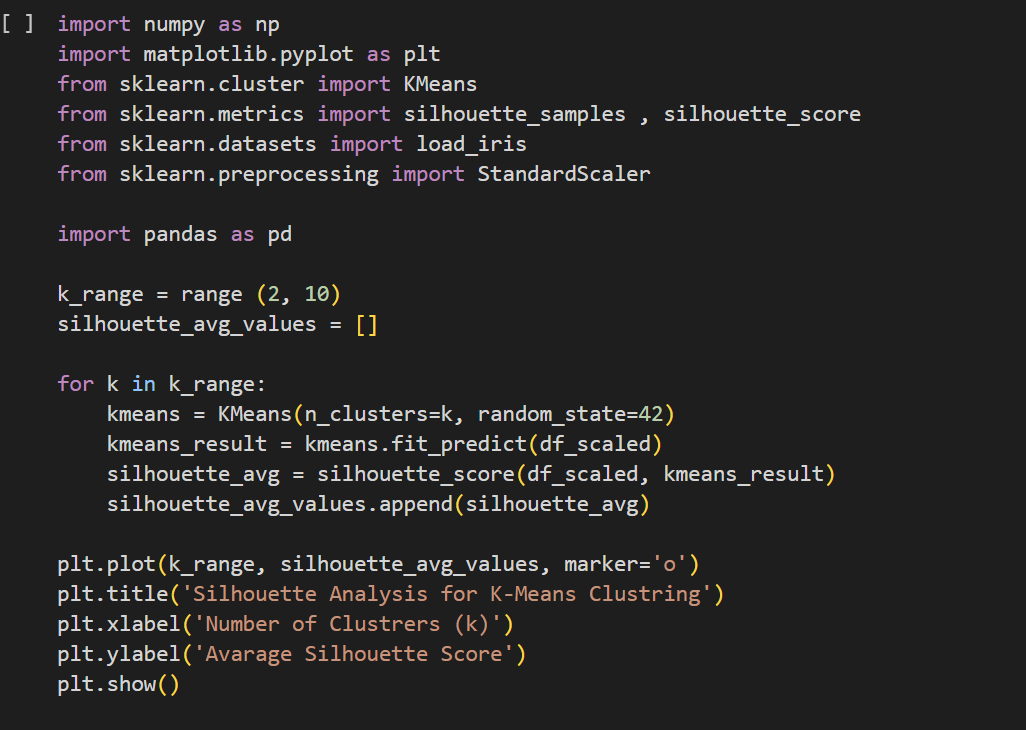
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1. **Problem**   We want to solve the problem of limited personalized educational support for students in core subjects. This is important because it directly impacts educational equity, future opportunities for students, and the overall progress of society. By working together to develop innovative solutions, we can help bridge the educational gap and empower all students to succeed.     1. **Data Mining Task**   In our project, we addressed the problem as a data mining task utilizing **classification** and **clustering** to analyze student performance.  For **classification**, we aimed to predict students' performance levels (high or low) using the target variable **Average Score**, which was converted into binary classes based on a defined threshold (e.g., 0.07). This task helped us identify the main factors influencing performance and provided insights to support underperforming students.  For **clustering**, we grouped students based on their attributes, excluding **Average Score**, to uncover natural patterns and similarities within the data. This unsupervised approach allowed us to identify distinct student groups, enabling the creation of personalized educational strategies.  By integrating both classification and clustering, our project offered a well-rounded analysis of student performance, supporting data-driven interventions and informed decision-making.   1. **Data**   **Dataset Description and Analysis**  **Dataset Overview**   * **Source**: Kaggle - Students Exam Scores Dataset [1]. * **Number of Objects (Records)**: 30,641 students * **Number of Attributes**: 14 attributes * **Main Characteristics of Attributes**:  |  |  |  |  | | --- | --- | --- | --- | | Attribute Name | Description | Data Type | Possible Values | | Gender | The gender of the student. | Categorical | "Male", "Female" | | EthnicGroup | The ethnic group of the student. | Categorical | Group A, B, C, D, E | | ParentEduc | The education level of the parent(s). | Categorical | "Some High School", "High School", "Some College", "Bachelor's", "Master's" | | LunchType | The type of lunch received by the student. | Categorical | "Standard", "Free/Reduced" | | TestPrep | Completion of test preparation courses. | Categorical | "Completed", "None" | | ParentMaritalStatus | The marital status of the parent(s). | Categorical | "Married", "Single", "Widowed", "Divorced" | | PracticeSport | The frequency of the student's sports activities. | Categorical | "Never", "Sometimes", "Regularly" | | IsFirstChild | Indicates if the student is the first child in the family. | Binary | "Yes", "No" | | NrSiblings | The number of siblings the student has. | Numeric | 0 to 7 | | TransportMeans | The student's primary means of transportation to school. | Categorical | "School Bus", "Private" | | WklyStudyHours | The number of hours the student spends studying per week. | Categorical | "< 5 hours", "5 - 10 hours", "> 10 hours" | | MathScore | The student's score in mathematics. | Numeric | 0 to 100 | | ReadingScore | The student's score in reading. | Numeric | 0 to 100 | | WritingScore | The student's score in writing. | Numeric | 0 to 100 | | Average Score | The average of MathScore, ReadingScore, and WritingScore, added during preprocessing for analysis purposes. | Numeric | Continuous (0 to 100) |   **Key Dataset Characteristics:**   1. **Missing Values**:     A screenshot of a table   1. **Distributions**:   A screen shot of a computer code  Description automatically generated  A graph of a bar chart  Description automatically generated with medium confidence  **Statistical Summary for Numeric Attributes:**   1. **Five Number& Variance Summary**:   A screenshot of a computer  Description automatically generated**A screenshot of a computer  Description automatically generated**      **Graphical Presentation:**   |  |  |  | | --- | --- | --- | | Name of Graph | Picture of Graph | Description | | Pie Chart | **A blue and pink circle with text  Description automatically generated** | **The distribution of "Gender" shows an almost equal ratio of male and female students.** | | Pie Chart | **A pie chart with numbers and a group of people  Description automatically generated** | **A pie chart depicting the distribution of students among various ethnic groups.** | | Pie Chart | **A colorful pie chart with text  Description automatically generated** | **Highlights the distribution of students by parental education levels.** | | Pie Chart | **A pie chart with a blue and orange circle  Description automatically generated** | **Displays the ratio of students with standard versus free/reduced lunch types.** | | Pie Chart |  | **Visualizes the marital status distribution of students' parents.** | | Scatter Plot | **A graph with a red line  Description automatically generated** | **Indicates no clear correlation between being a first child and academic performance.** | | Boxplot |  | **Compares average scores between male and female students, with similar performance ranges.** | | Boxplot | **A chart of different groups  Description automatically generated with medium confidence** | **Highlights score distributions across different ethnic groups, showing similar variability.** | | Boxplot | **A graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of  Description automatically generated** | **Reveals higher parental education correlates with improved student performance.** | | Boxplot | **A graph of a number of bars and a number of numbers** | **Shows students with >10 hours of weekly study tend to achieve higher scores.** | | Boxplot | **A graph of a test preparation  Description automatically generated** | **Students who completed test preparation have notably higher average scores.** | | Boxplot |  | **Indicates that regular sports participation does not significantly affect academic performance.** | | Boxplot | **A graph of a number of siblings  Description automatically generated** | **Displays no strong correlation between the number of siblings and academic performance.** | | Boxplot | **A chart of a graph** | **Compares scores for students using private transport versus school buses, showing minimal impact.** |  1. **Data preprocessing**    **Filling the missing values:**   The columns EthnicGroup, ParentEduc, ParentMaritalStatus, and IsFirstChild were processed by filling their missing values using their respective modes, which is the most frequent value in each column. This approach is suitable for handling missing categorical data as it maintains consistency without altering the original distribution. As a result, missing values were replaced with the most common value in each column, ensuring the dataset remains complete while preserving its original data pattern.     **Detect and Removing Outliers:**  **Justification for Not Handling Outliers:** Although outliers were detected in the dataset, their number is relatively small compared to the total sample size. This minimal proportion of outliers does not significantly impact the credibility or validity of the data. Therefore, handling outliers in this case is not deemed necessary, as their influence on the overall analysis and model performance is negligible. **Data Transmission:**  * **Encoding:**     Label Encoding was applied to transform categorical columns ('Gender', 'EthnicGroup', 'ParentEduc', 'TestPrep', 'PracticeSport', 'IsFirstChild') into numerical format, as machine learning models require numerical data. This method assigns a unique integer to each category, making the data suitable for analysis. Label Encoding is particularly effective for nominal data with no intrinsic order, providing a simple and efficient conversion. By applying this preprocessing step, the dataset is now ready for machine learning models, ensuring accurate model training and analysis.       * **Normalization:**     Data preprocessing was applied to ensure the dataset is suitable for machine learning models, specifically through **Decimal Scaling Normalization** of the Average Score column. This technique scales values by dividing them by a power of 10 based on the maximum absolute value, ensuring all values are within a similar range. Normalization was necessary to prevent features with larger magnitudes from dominating the model, improving its stability and performance, especially for distance-based or gradient-based algorithms. Decimal scaling was chosen for its simplicity, efficiency, and suitability when values are within a reasonable range. The processed dataset is now ready for further analysis or model training.       * **Aggregation:**   The dataset is loaded from Processed\_dataset.csv, which has been preprocessed (e.g., encoding, normalization). The data is grouped by WklyStudyHours and TestPrep, calculating the average score for each combination of these factors using the. mean() function. Grouping allows us to analyze how study hours and test preparation affect academic performance by comparing the average scores across these categories, helping to understand the impact of these factors on overall performance. This technique is useful when exploring how different factors influence outcomes like average scores.       * **Discretization:**   The dataset is loaded from Processed\_dataset.csv, and the continuous NrSiblings column is discretized into categorical bins using pd.cut().The bins are defined as [-1, 0, 2, 5, 7], with values categorized as 'None', 'Few', 'Moderate', and 'Many'. Discretization helps transform continuous data into more interpretable categories, which is useful for certain analyses or machine learning models. It simplifies the data, making it easier to analyze or model by grouping continuous values into predefined segments.       * **Balance Data:**   Before starting the Data Mining Technique, we investigated whether the data was  balanced or not:      Based on the calculated values, the "Average Score" appears to be relatively balanced. The mean score is 68.12, which is very close to the median of 68.33, indicating a central distribution of scores around this value. Additionally, the standard deviation of 14.45 suggests moderate variance among the scores, meaning that scores are not widely spread from the mean. Overall, the "Average Score" can be considered reasonably balanced, with scores distributed around a stable central value without significant dispersion.   1. **Data Mining Technique**   This project employs **classification** and **clustering** techniques to analyze students' academic performance. The primary objective is to classify students based on their performance and group them into clusters based on shared characteristics. These approaches help uncover patterns and relationships that can be used to improve educational outcomes.  **Classification**  For classification, the **Decision Tree** algorithm was used due to its transparency and ability to handle both categorical and numerical data. The following steps were implemented:   1. **Splitting the Data:**    * The dataset was split into training and testing sets using different ratios:      + **90% Training / 10% Testing**      + **80% Training / 20% Testing**      + **70% Training / 30% Testing** 2. **Criteria for Splitting:**    * **Information Gain (Entropy):** Measures the reduction in uncertainty when splitting the data.    * **Gini Index:** Evaluates the "impurity" of a dataset split to determine its quality. 3. **Evaluation Metrics:** The model performance was evaluated using:    * **Accuracy:** Percentage of correct predictions.    * **Sensitivity (Recall):** Proportion of actual positive cases identified correctly.    * **Specificity:** Proportion of actual negative cases identified correctly.    * **Precision:** Proportion of positive predictions that were correct.    * **Error Rate:** Percentage of incorrect predictions. 4. **Visualization:**    * **Decision Tree Diagrams** were used to display the hierarchical structure of feature splits.    * **Confusion Matrices** illustrated the true positives, true negatives, false positives, and false negatives for each split ratio.   The results showed that the **90%-10% split** achieved the best balance between accuracy and specificity when using the **Information Gain (Entropy)** criterion.  **Clustering**  For clustering, the **K-means** algorithm was applied to group students into clusters based on their shared attributes. This technique provides insights into the distribution and similarities among students.   1. **Preprocessing:**    * The **target column (Average Score)** was excluded to ensure clustering relied solely on other features.    * Features were scaled using normalization to ensure consistent scaling across attributes. 2. **Number of Clusters (KKK):**    * Different values of KKK were tested to identify the optimal number of clusters:      + **Elbow Method** was used to determine K=5K = 5K=5, where adding more clusters resulted in diminishing returns in terms of compactness.      + **Silhouette Analysis** revealed that K=9K = 9K=9 achieved the highest silhouette score, indicating the best separation between clusters. 3. **Visualization:**    * The **Elbow Method Graph** was used to display the relationship between KKK and the within-cluster sum of squares (WSS), identifying K=5K = 5K=5 as the optimal point.    * The **Silhouette Plot** visualized the cohesion and separation of clusters, demonstrating the quality of clustering at different values of KKK.   **Python Packages and Methods**   * **Data Preprocessing:**   + pandas for data manipulation.   + StandardScaler from sklearn.preprocessing for scaling features. * **Classification:**   + DecisionTreeClassifier from sklearn.tree for building and evaluating the decision tree.   + train\_test\_split from sklearn.model\_selection for splitting the data.   + plot\_tree and matplotlib for visualizing the decision tree.   + confusion\_matrix from sklearn.metrics for evaluating model performance. * **Clustering:**   + KMeans from sklearn.cluster for clustering.   + silhouette\_score from sklearn.metrics for evaluating cluster quality.   + matplotlib for creating the Elbow and Silhouette plots.  1. **Evaluation and Comparison**  * **Classification:**   Classification [90% training, 10% testing] Information Gain:  **Figure (1) (decision tree):**    **Figure (2) (confusion matrix):**    Classification [80% training, 20% testing] Information Gain:  **Figure (1) (decision tree):**    **Figure (2) (confusion matrix):**  صورة تحتوي على نص, لقطة شاشة, رقم, رسم بياني  تم إنشاء الوصف تلقائياً  Classification [70% training, 30% testing] Information Gain:  **Figure (1) (decision tree):**    **Figure (2) (confusion matrix):**  صورة تحتوي على نص, لقطة شاشة, رقم, الخط  تم إنشاء الوصف تلقائياً        Classification [90% training, 10% testing] Gini Index:  **Figure (1) (decision tree):**    **Figure (2) (confusion matrix):**    Classification [80% training, 20% testing] Gini Index:  **Figure (1) (decision tree):**    **Figure (2) (confusion matrix):**    Classification [70% training, 30% testing] Gini Index:  **Figure (1) (decision tree):**    **Figure (2) (confusion matrix):**           * **The better partitioning:**   The 90%-10% splits (both Gini and Entropy) generally perform best across metrics such as accuracy, error rate, specificity, and precision, while the 70%-30% split (especially Gini) performs better on sensitivity and true positives. Depending on the importance of each metric for your use case (e.g., prioritizing accuracy over sensitivity), the 90%-10% Gini or Entropy model could be the optimal choice for a balanced performance, while the 70%-30% Gini split might be preferred if capturing positives is more critical. |

* **Clustering:**

Based on the outcomes of the validation techniques, we select three different sizes [5,8,9], and we then utilize these sizes to calculate the k-means clustering.

**1. Silhouette Method**

The Silhouette approach assesses clustering quality by comparing how well each point fits into its own cluster vs others. ratings range from -1 to 1, with higher ratings indicating clearly identifiable clusters.



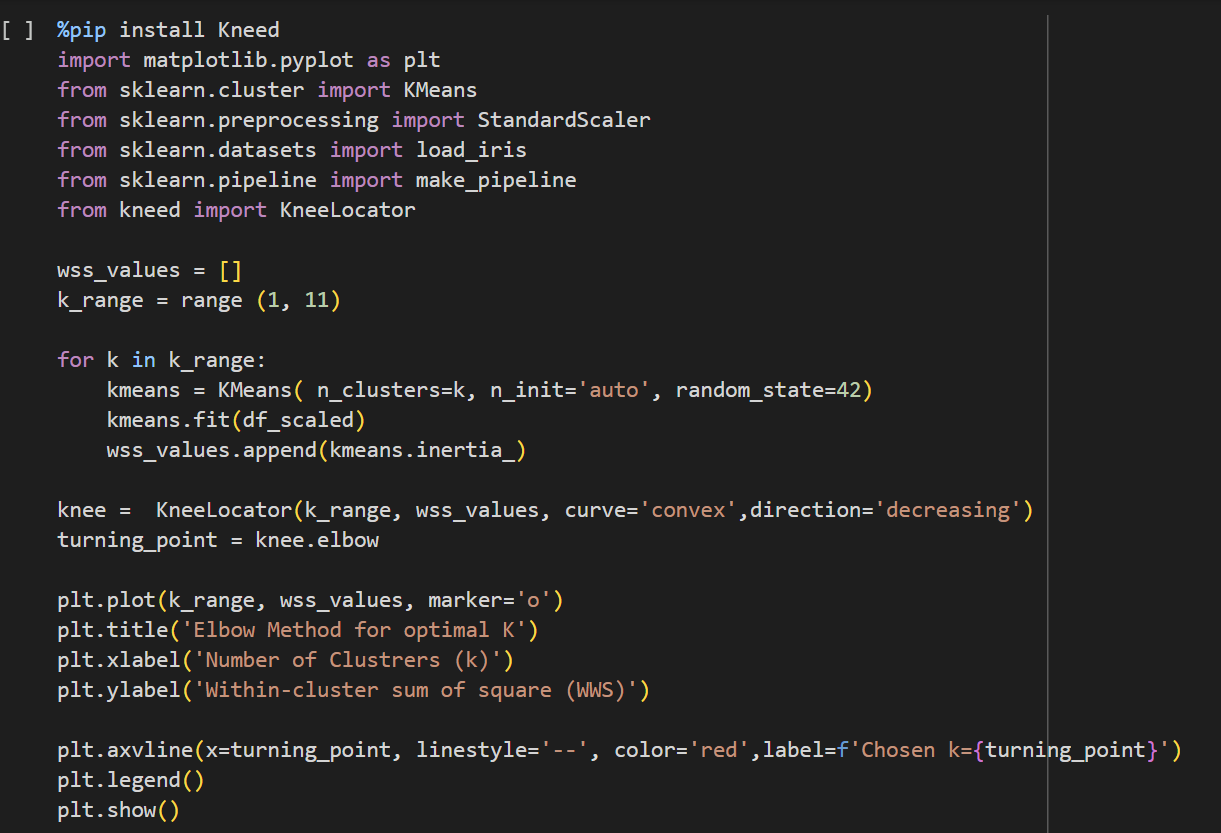
صورة تحتوي على نص, خط, رسم بياني, تخطيط

تم إنشاء الوصف تلقائياً

As seen above, we discovered that the ideal number of clusters (k) for maximizing the average Silhouette coefficient is 9, which will be our initial K-means option. and the second highest average Silhouette coefficient is 8.

**2. Elbow method**

The Elbow Method plots the within-cluster sum of squares (WSS) against (K) to find the ideal number of clusters. The ideal (K) is selected at this elbow point.

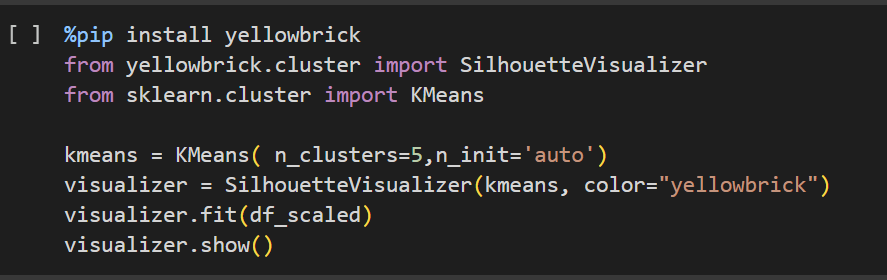
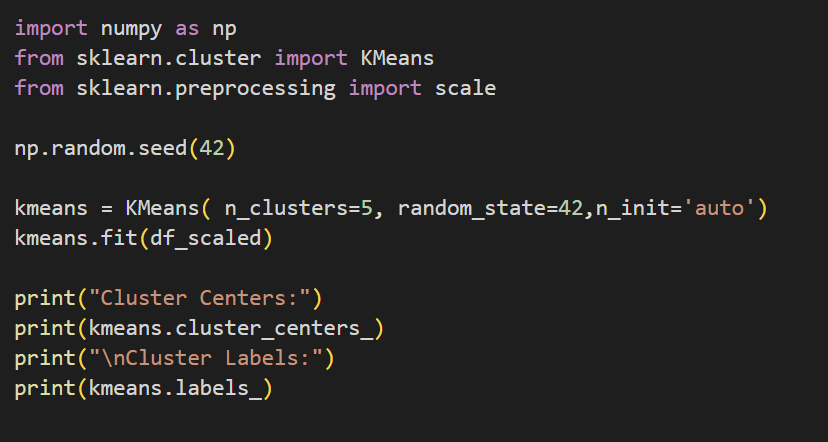


صورة تحتوي على نص, لقطة شاشة, خط, تخطيط

تم إنشاء الوصف تلقائياً

Our analysis of the elbow plot revealed a second turning point at (k = 5 ). We will choose (k = 5 ) for the third K-means clustering, as this reflects a new cluster structure. By taking into account this extra turning point, we hope to capture a varied spectrum of cluster forms while maximizing potential clustering performance.

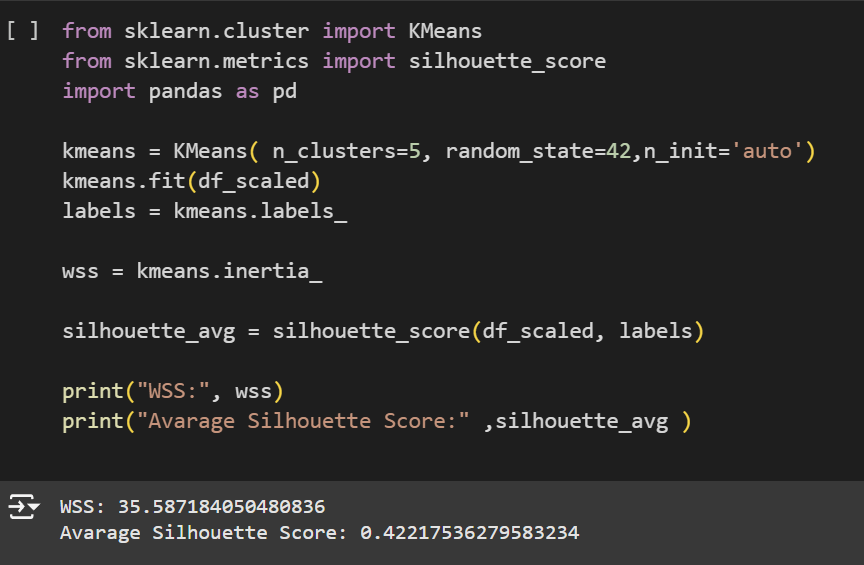
* Trail 1 : Silhouette scores [K =5]



صورة تحتوي على نص, لقطة شاشة, تخطيط, خط

تم إنشاء الوصف تلقائياً

the majority of the silhouette scores are positive indicates that the samples are well-matched to their clusters and are separated from nearby clusters. This shows that the clustering algorithm successfully divided the data into discrete, well-defined groups.

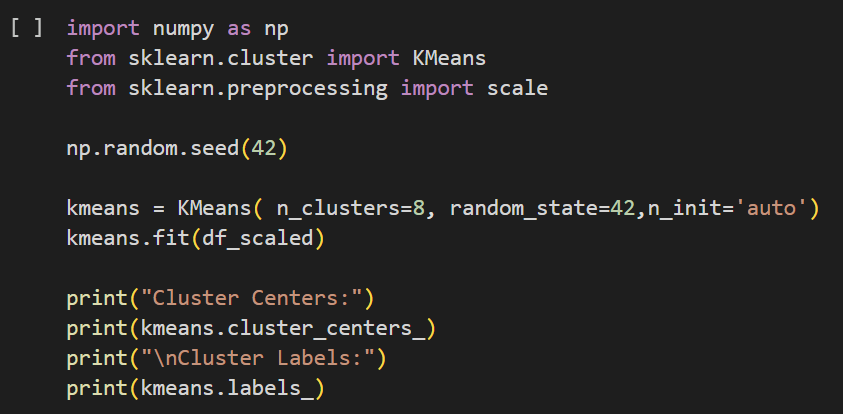


The WSS value of 35.5872 represents the total variation within the clusters, with lower values indicating tighter clusters. While this value is not very low, it suggests that the clusters are somewhat separated but still may have room for improvement in terms of compactness.

The Average Silhouette Score of 0.4222 is relatively good, indicating that the clusters are fairly well differentiated. However, there is still some overlap or ambiguity between the clusters, suggesting that the clustering quality might be further improved.

* Trail 2: Silhouette scores [K =8]

صورة تحتوي على نص, لقطة شاشة, برمجيات, برامج الوسائط المتعددة

تم إنشاء الوصف تلقائياً

صورة تحتوي على نص, لقطة شاشة, تخطيط, خط

تم إنشاء الوصف تلقائياً

Most of the silhouette scores with a positive value reinforce the notion that the samples are well-matched to their clusters and are distant from neighboring clusters. This indicates that the clustering solution has successfully separated the data points into distinct and well-defined clusters.

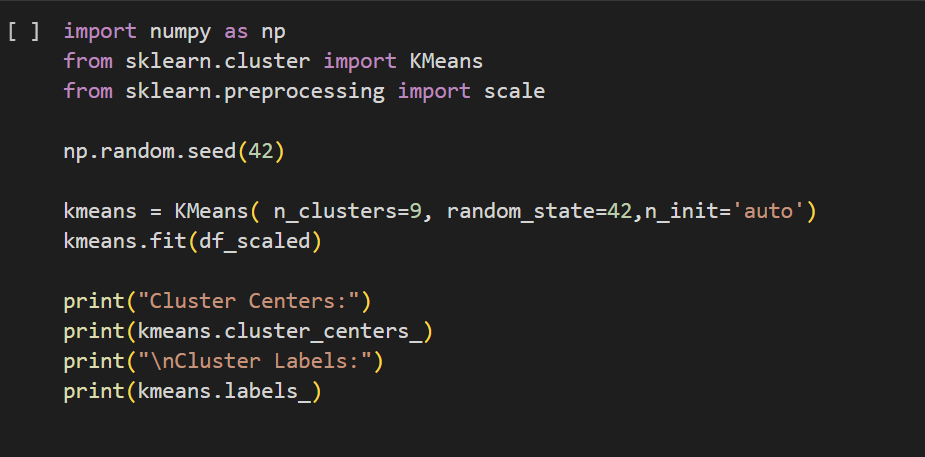
صورة تحتوي على نص, لقطة شاشة, الخط, برمجيات

تم إنشاء الوصف تلقائياً

WSS: 20.09324711660112 - The relatively low WSS value suggests that the clusters are reasonably well-separated and compact, though not as optimal as some lower WSS values might indicate. The clusters are still relatively tight but could potentially benefit from some refinement.

Average Silhouette Score: 0.41622939498931744 - The moderately high score indicates that there is still some degree of overlap or ambiguity in the cluster assignments, but the clusters are generally well-defined and distinct. While not perfect, the clusters appear to exhibit acceptable separation overall.

* Trail 3: Silhouette scores [K =9]

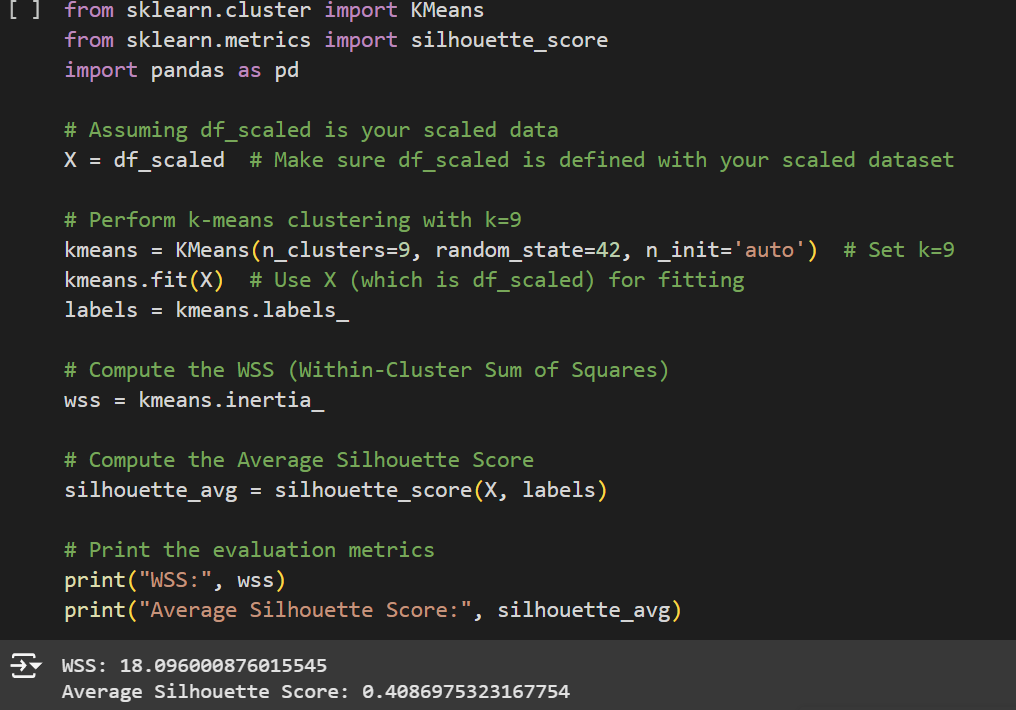


صورة تحتوي على نص, لقطة شاشة, الخط

تم إنشاء الوصف تلقائياًصورة تحتوي على نص, لقطة شاشة, خط, تخطيط

تم إنشاء الوصف تلقائياً

the fact that most of the silhouette scores have positive values is indeed a positive indicator. Positive silhouette scores suggest that the samples are well-matched to their clusters and are relatively distant from neighboring clusters. This reinforces the notion that the clustering solution has successfully separated the data points into distinct and well-defined clusters.



WSS: 18.096000876015545 - The WSS value indicates that the cluster separation and compactness are reasonable, though not as strong as expected for ideal clustering.

Average Silhouette Score: 0.4086975323167754 - The relatively high score suggests that there is some overlap or ambiguity in the cluster assignments, but overall the clusters are reasonably well-defined and distinc.

|  |  |
| --- | --- |
| Mining task | Comparison Criteria |
| Clustering | We used 3 sizes of K,  K=5, K=8, K=9   |  |  |  |  | | --- | --- | --- | --- | |  | K=5  (best) | K=8 | K=9 | | Average Silhouette width | 0.42218 | 0.41623 | 0.40870 | | Total within-cluster sum of square | 35.587184050480836 | 20.09324711660112 | 18.096000876015545 | |

Conclusion: The K=5 model demonstrates superior clustering performance compared to the K=8 and K=9 models. It effectively separates the data into distinct and well-defined clusters, making it the most suitable choice for this particular dataset.

1. **Findings**

The goal of this analysis was to classify students based on their performance and to identify clusters of students with similar characteristics. **Classification** was carried out using the **Decision Tree algorithm**, tested with both **Information Gain (Entropy)** and **Gini Index** criteria. For **clustering**, the **K-means algorithm** was used to group students based on shared features, with different values of KKK evaluated to determine the optimal number of clusters. The results are discussed in detail below.

**Classification Results**

The **Decision Tree algorithm** was applied with varying criteria and training/testing splits. The performance of the models was evaluated using confusion matrices and metrics such as accuracy, sensitivity, specificity, and precision. The following observations were made:

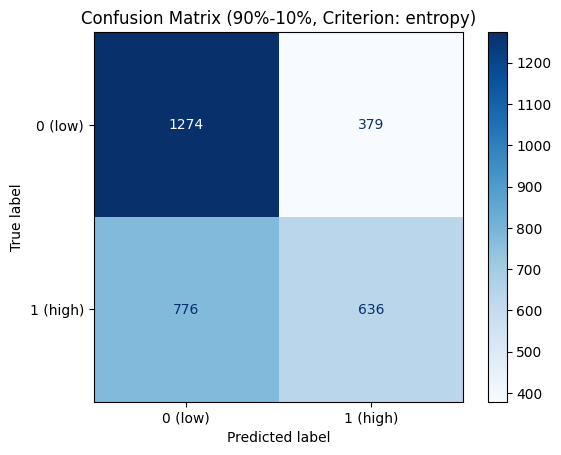
**Using Information Gain (Entropy):**

1. **Model Performance**:
   * The **90%-10% training/testing split** achieved a good balance between specificity and precision.
   * The confusion matrix revealed:
     + **True Positives (TP):** 636 (high-performing students correctly classified).
     + **True Negatives (TN):** 1274 (low-performing students correctly identified).
     + **False Positives (FP):** 379 (low performers misclassified as high performers).
     + **False Negatives (FN):** 776 (high performers misclassified as low performers).
2. **Feature Importance**:
   * The **Decision Tree diagram** highlighted **Test Preparation** as the most decisive feature, followed by **Parental Education** and **Weekly Study Hours**.

* **Decision Tree Diagram (Entropy)**:  
  A diagram of a network

  Description automatically generated

"This Decision Tree illustrates the hierarchical structure of decision-making based on the Entropy criterion. Key features such as **Test Preparation** and **Parental Education** are crucial for predicting student performance."

* **Confusion Matrix (90%-10%, Entropy)**:  
  

"The confusion matrix displays the classification performance for the Decision Tree using Information Gain, highlighting true/false positives and negatives."

**Using Gini Index:**

1. **Model Performance**:
   * The **80%-20% training/testing split** showed slightly improved sensitivity, making it more effective for detecting high-performing students.
   * The confusion matrix for this split revealed:
     + **True Positives (TP):** A higher count compared to the Entropy criterion.
     + **False Negatives (FN):** Reduced misclassification of high performers.
2. **Feature Importance**:
   * Similar to Entropy, **Test Preparation** and **Parental Education** were significant, but the tree structure was more compact, leading to faster decision-making.

* **Confusion Matrix (80%-20%, Gini)**:  
  A blue squares with white text

  Description automatically generated

"This confusion matrix showcases the results for the Decision Tree model using the Gini Index, with improved sensitivity for high-performing students."

**The Best Model Between Information Gain and Gini Index**

After analyzing the performance of both Information Gain (Entropy) and Gini Index, the Decision Tree using Information Gain with a 90%-10% split was selected as the best model for this dataset. The reasons for this choice are:

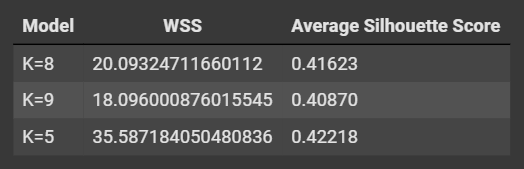
1. Balanced Performance:
   * While both criteria achieved similar accuracy, Information Gain provided a better balance between specificity (correctly identifying low performers) and precision (correctly predicting high performers).
   * This makes it more suitable for datasets where accurately identifying key groups (e.g., high performers) is critical.
2. Feature Insights:
   * The model using Information Gain highlighted Test Preparation, Parental Education, and Weekly Study Hours as the most influential features, which aligns with practical expectations and provides actionable insights.
3. Model Interpretability:
   * The splits generated by Information Gain were clearer and more intuitive compared to Gini Index, making it easier to explain the decision-making process to stakeholders.

Conclusion: The Decision Tree with Information Gain was selected as the best model due to its superior balance of performance metrics and its ability to provide interpretable results. This model is particularly useful for identifying at-risk students and designing targeted interventions to improve academic outcomes.

**Best Model for Classification:**

After evaluating both criteria, the **Decision Tree using Information Gain (Entropy)** with a **90%-10% split** was chosen as the best model. Its ability to balance specificity and precision makes it suitable for addressing the problem.

**Clustering Results  
  
For Clustering, we used K-means algorithm with 3 different K to find the optimal number of clusters, we calculated the average silhouette width for each K, and we concluded the following results:**

****

The **K-means algorithm** was applied to group students into clusters based on their shared characteristics. Several KKK values were tested to determine the optimal number of clusters.

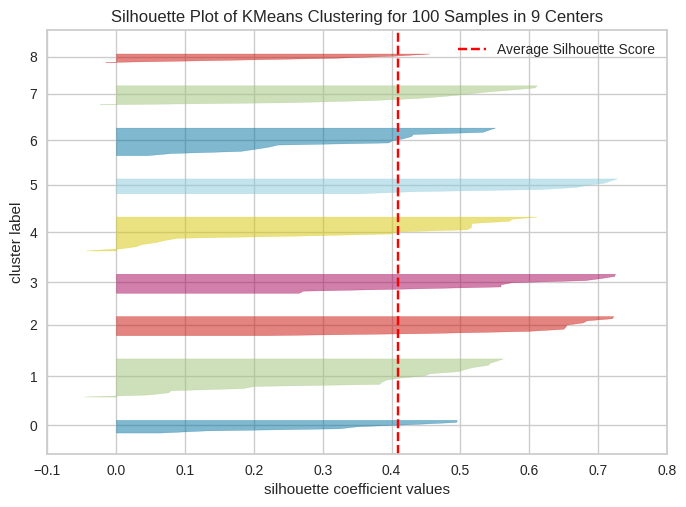
**Optimal Clustering:**

1. **Elbow Method**:
   * The **Elbow Method** identified K=5K = 5K=5 as the optimal number of clusters, balancing compactness and simplicity.
   * This ensures well-defined clusters with minimal overlap.
2. **Silhouette Score**:
   * The highest **Silhouette Score** was observed at K=9K = 9K=9, indicating the clusters are well-separated and cohesive.

* **Elbow Method Graph**:  
  A graph of a number of clusters

  Description automatically generated

"The Elbow Method graph identifies K=5K = 5K=5 as the optimal number of clusters, where adding more clusters no longer significantly reduces the Within-Cluster Sum of Squares (WSS)."

* **Silhouette Plot**:  
  

"The Silhouette Plot evaluates clustering quality, with K=9K = 9K=9 achieving the highest average score, indicating well-separated clusters."

**Cluster Insights:**

1. **Cluster Characteristics**:
   * **High-Performing Clusters**:
     + Students with consistent test preparation and high weekly study hours.
   * **Low-Performing Clusters**:
     + Students with limited parental education and lower test preparation efforts.
2. **Patterns Revealed**:
   * The clustering analysis revealed distinct groups of students based on socio-economic and study-related factors, enabling targeted interventions.

* **Cluster Distribution Scatter Plot (K = 5)**:  
  A graph of a graph with different colored lines

  Description automatically generated with medium confidence

"This scatter plot visualizes the distribution of students across clusters for K=5K = 5K=5, highlighting shared characteristics such as **Test Preparation** and **Parental Education**."

**Problem Solutions**

Based on the findings, the following solutions are proposed:

1. **Targeted Interventions**:
   * Focus on students in low-performing clusters by providing additional support for test preparation and improving study habits.
2. **Policy Recommendations**:
   * Encourage schools to provide resources for parental engagement, as **Parental Education** plays a significant role in student outcomes.
3. **Use of Decision Trees**:
   * The Decision Tree model offers a transparent framework for identifying students needing assistance and allocating resources effectively.

**Conclusion**

The analysis demonstrates the effectiveness of combining **classification** and **clustering** techniques to address the problem. The **Decision Tree using Information Gain (Entropy)** and **K-means clustering with K=5K = 5K=5** is identified as the best models for solving the problem under study. These methods not only predict student performance but also uncover actionable patterns for improving academic outcomes.

1. **References**

[1] D. Geb, "Students Exam Scores," \*Kaggle\*. [Online]. Available: https://www.kaggle.com/datasets/desalegngeb/students-exam-scores. [Accessed: Nov. 30, 2024].