APPLIED DATA SCIENCE -1
ASSIGNMENT - 2: CLUSTERING AND FITTING

Name: PAVITHIRA SEENIVASAGAN
Student id: 23095934
Mail id: ps24abe@herts.ac.uk
Github link: https://github.com/Pavithiraseenivasagan/Clustering-and-Fitting
Dataset link: https://www.kaggle.com/datasets/ikynahidwin/depression-student-dataset

Introduction

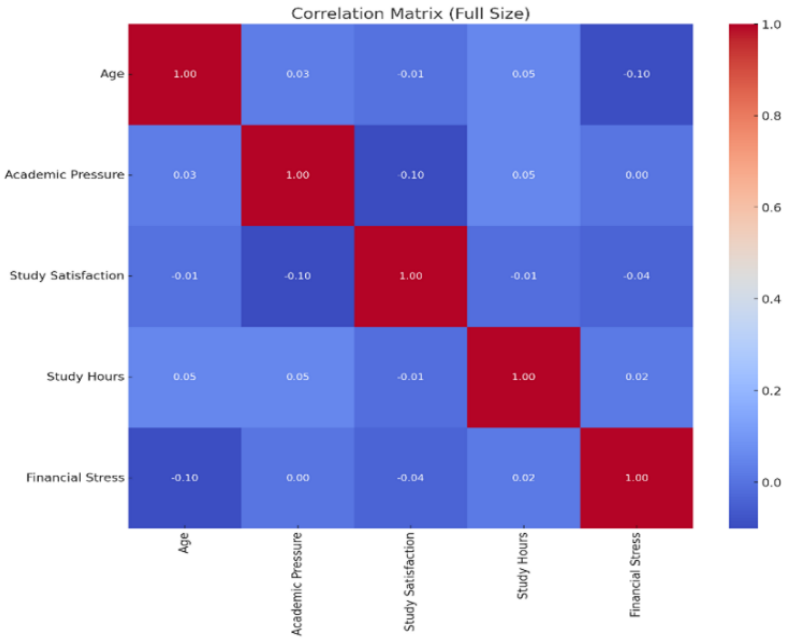
Student depression is a significant concern. This study aims to analyse student data to understand factors contributing to depression. By using machine learning, we will:

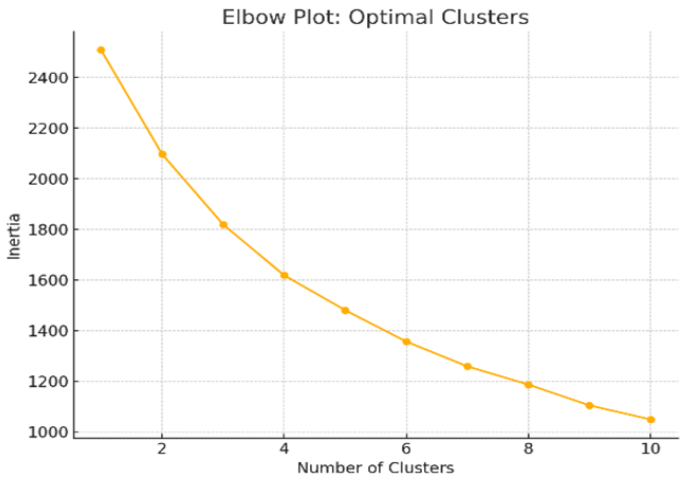
1. Identify distinct student groups using K-means clustering.
2. Analyse depression patterns within these groups.
3. Predict depression risk using linear regression. Understanding these patterns will help develop effective support strategies for student mental health.

Abstract

This study delves into student depression using machine learning. Data was pre-processed and exploratory analysis revealed weak feature correlations. K-means clustering identified three distinct student groups with varying depression levels. Linear regression modelled the relationship between student factors and depression, providing a predictive model. A histogram visualized age distribution. These findings can inform targeted interventions for student mental health.

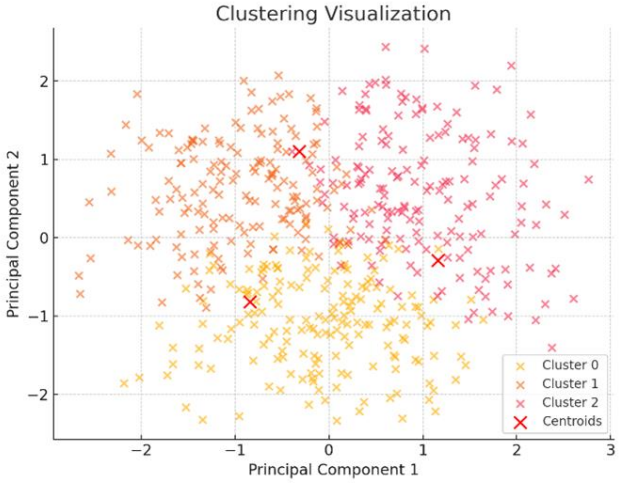
Correlation Matrix: The correlation matrix highlights the relationships between the numerical variables in the dataset. Most correlations are weak or near zero, indicating a general lack of strong relationships between the features. For instance, "Academic Pressure" and "Study Satisfaction" have a slight negative correlation (-0.10), suggesting that increased pressure might slightly lower satisfaction. Similarly, "Age" and "Financial Stress" show a weak negative relationship (-0.10), hinting that younger individuals may experience more financial stress. The diagonal values of 1.00 confirm perfect self-correlation for each variable, while the shades of red and blue in the heatmap indicate positive and negative relationships, respectively. Overall, the weak correlations suggest that deeper patterns may only be uncovered using more advanced techniques like clustering or regression.

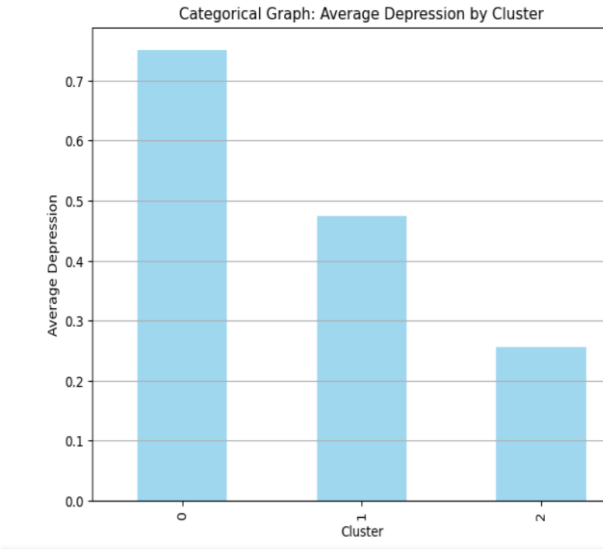




Elbow Plot: The elbow plot demonstrates how the inertia, or within-cluster sum of squares, decreases as the number of clusters increases. A steep drop in inertia is visible up to three clusters, after which the decrease slows significantly, creating an "elbow" at three clusters. This suggests that three clusters are the optimal choice for the dataset, balancing simplicity with effectiveness. Adding more clusters beyond this point results in diminishing returns, with little improvement in the model's performance. By selecting three clusters, the data is segmented meaningfully without overcomplicating the clustering process.

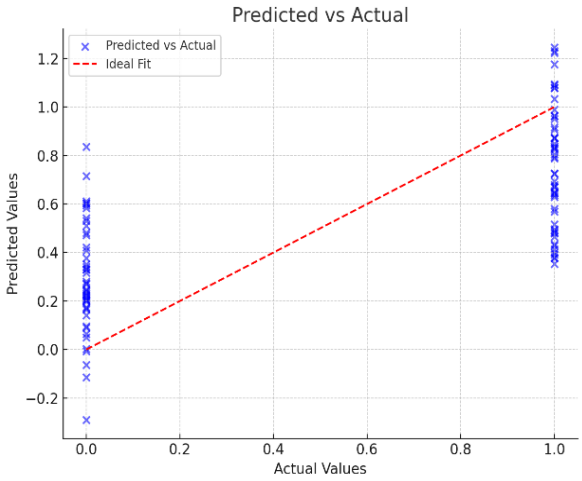
Clustering Visualization: This plot visually represents the three clusters identified by the K-means algorithm, reduced to two dimensions using PCA for simplicity. Each point corresponds to a data entry, with colors distinguishing the clusters and red "X" markers indicating the cluster centroids. The clear separation between clusters suggests distinct groupings, while overlapping areas hint at shared characteristics among some data points. The use of PCA preserves the data's structure, allowing for an interpretable visualization of the clusters. This result aligns with the elbow plot, confirming that three clusters effectively capture the natural segmentation of the dataset.

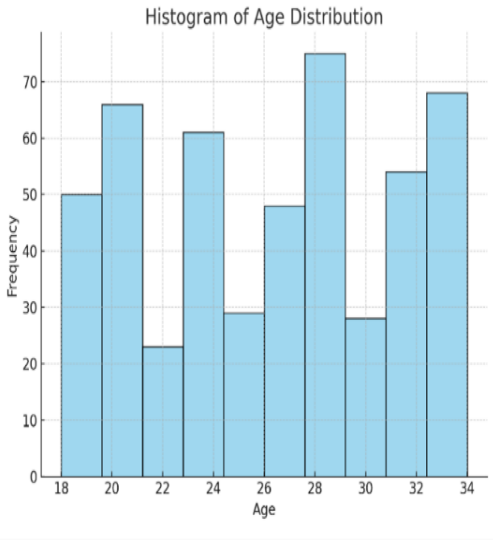




Average Depression by Cluster: The bar chart illustrates the average depression levels across the three clusters identified by K-means. Cluster 0 has the lowest average depression, indicating that individuals in this group report fewer depressive symptoms. Cluster 1 shows a moderate average, suggesting slightly elevated depression levels compared to Cluster 0. Cluster 2 stands out with the highest average depression, marking it as the group most affected by depressive symptoms. This visualization highlights distinct behavioural patterns, helping to pinpoint groups that might benefit most from targeted mental health interventions.

Predicted vs. Actual Depression (Regression): The scatter plot compares the predicted depression values against the actual ones from the regression model. The red dashed line represents an ideal scenario where predictions perfectly match the actual values. While many points align closely to the line, some deviations, especially at the extremes, suggest areas where the model's predictions could be improved. These discrepancies provide valuable insights into where the model struggles, offering opportunities for further refinement. Overall, the plot demonstrates that the regression model is reasonably effective but has room for optimization.





Age Distribution (Histogram): The histogram showcases the distribution of students' ages in the dataset. The x-axis represents age ranges, while the y-axis indicates the frequency of individuals in each range. The peaks around ages 20, 28, and 32 suggest that these age groups are the most represented, while dips around ages 22 and 30 indicate fewer individuals in those age brackets. This visualization provides a clear picture of the age demographics in the dataset, helping to contextualize other findings and tailor interventions to the most common age groups.

Conclusion

This study successfully used machine learning to analyse student data. K-means identified distinct student groups with varying depression levels, emphasizing the need for tailored interventions. Linear regression provided a predictive model for identifying at-risk students. These findings highlight the importance of comprehensive mental health support. By addressing the unique needs of different student groups, we can promote better mental well-being and create supportive learning environments.