DEPARTMENT OF INFORMATION AND ELECTRONIC ENGINNERING

TITLE

Data Analysis using Data Mining Algorithms

Student : Moumoulidou Maria

Student ID: 185341

Table of Contents

1 Classification 4

1.1 Introduction 4

1.2 Data Preprocessing 4

1.3 Results and Confusion Matrix 4

1.4 Conclusions 8

2 Clustering 9

2.1 Introduction 9

2.2 Data Preprocessing 9

2.3 K-means Clustering Method 10

2.3.1K-means Clustering Results 10

2.4 Αgglomerative Clustering Method 11

2.4.1Αgglomerative Clustering Results 11

2.5 Clustering with DBSCAN Algorithm 12

2.5.1Results with DBSCAN Algorithm 13

3 Association Rule Mining 14

3.1 Introduction 14

3.2 Data Preparation 14

3.3 Rule Extraction 14

**Abstract**

In this project, two datasets were studied and analyzed. The first one was the Spam dataset, which involved the study and classification of email messages into "email" and "spam." The second one was the StudentsPerformance dataset, on which three clustering methods (K-means, DBSCAN, and Αgglomerative Clustering) were applied in order to classify students based on their performance and to compare the results of each method.

In addition, in the final part of this project, the Apriori algorithm was applied to generate association rules, which were then discussed.

For the implementation of the project, the Python programming language and its relevant libraries were used.

# 1 Classification

## **Introduction**

In this section, the classification of electronic messages into two main categories, “email” and “spam,” was examined using and comparing the following classification algorithms: (a) Decision Trees, (b) k-Nearest Neighbors, (c) Naive Bayes, and (d) Random Forest. The *spam.csv* dataset was used for this purpose.

## **1.2 Data Preprocessing**

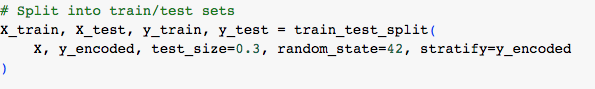
The process began with loading and inspecting the dataset. In the *class* column, there was a typographical error where “email” had been recorded as “emai,” which was corrected. Then, using *LabelEncoder*, the values “email” and “spam” were transformed into 0 and 1, respectively. The data was subsequently split into training and testing sets with a 70/30 ratio.

|  |  |
| --- | --- |
| Category | Number of Records |
| Email | 2788 |
| Spam | 1813 |

**Macintosh HD:Users:georgemoumoulidis:Desktop:Στιγμιότυπο 2025-08-24, 12.44.37 πμ.png**

**Macintosh HD:Users:georgemoumoulidis:Desktop:Στιγμιότυπο 2025-08-24, 12.45.38 πμ.png**

**Macintosh HD:Users:georgemoumoulidis:Desktop:Στιγμιότυπο 2025-08-24, 12.45.46 πμ.png**

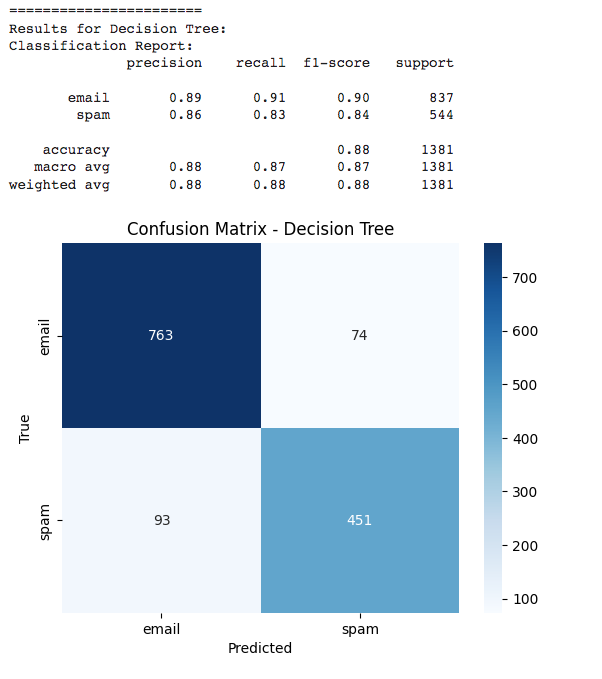
****

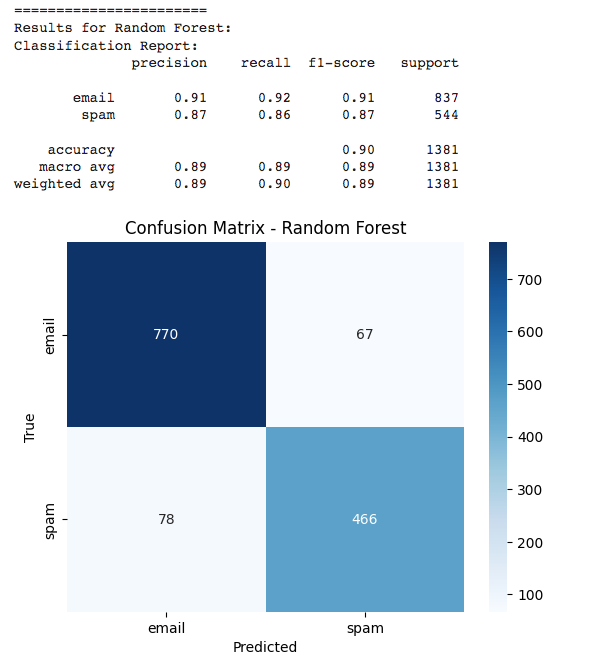
## **1.3 Results and Confusion Matrix**

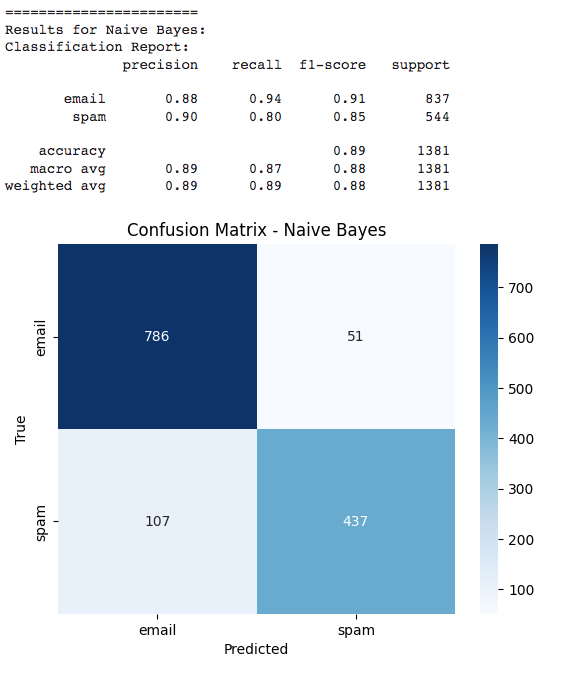
The algorithms were implemented using Python's scikit-learn library.

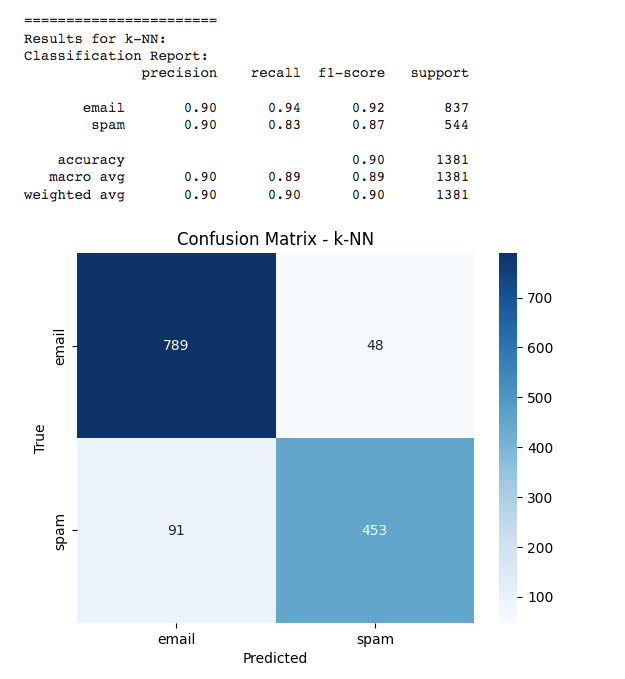
For spam detection:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Algorithm | Accuracy | Precision | Recall | F1- score |
| Decision Tree | 0.88 | 0.85 | 0.85 | 0.85 |
| Random Forest | 0.89 | 0.85 | 0.87 | 0.86 |
| Naive Bayes | 0.89 | 0.90 | 0.80 | 0.85 |
| KNN | 0.90 | 0.88 | 0.85 | 0.87 |

****

****

****

****

## **1.4 Conclusions**

From the experiments conducted with the four classification algorithms, it can be observed that Random Forest achieved the best performance. The kNN algorithm was also effective but with slightly lower metrics. The Decision Tree provided good interpretability, while Naive Bayes was the simplest and fastest, though with slightly lower accuracy.

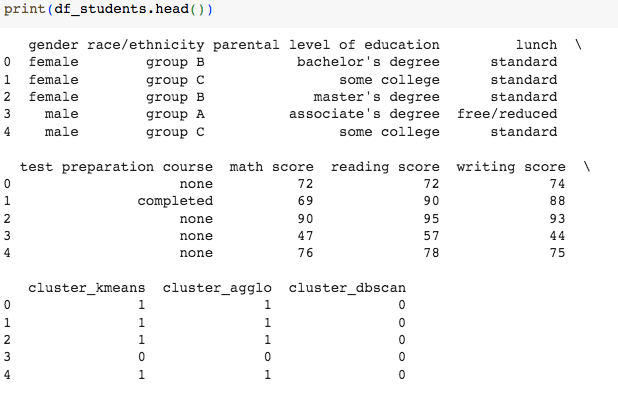
# 2 Clustering

## **2.1 Introduction**

In this section, clustering techniques were applied to the StudentPerformance dataset. The methods used were: (a) K-means clustering, (b) Hierarchical clustering, and (c) clustering with the DBSCAN algorithm.

## **2.2 Data Preprocessing**

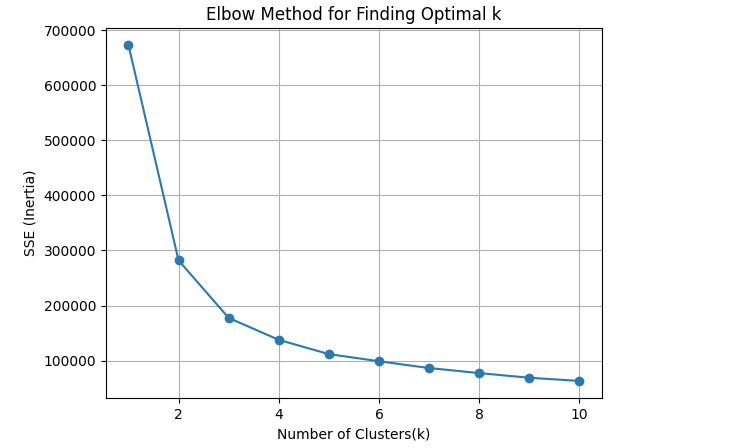
Only the numerical attributes math score, reading score, and writing score were used.



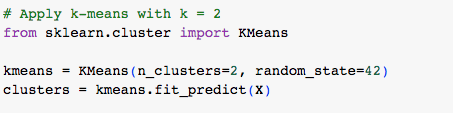
Macintosh HD:Users:georgemoumoulidis:Desktop:Στιγμιότυπο 2025-08-24, 3.23.55 μμ.png

## **2.3 K-means Clustering Method**

To determine the optimal number of clusters, the Elbow method was applied

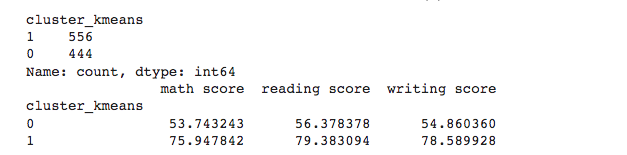


As shown in the plot, the ‘elbow’ point appears at k = 2, indicating that the dataset is divided into two main groups.



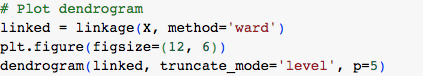
### 2.3.1 K-means Clustering Results

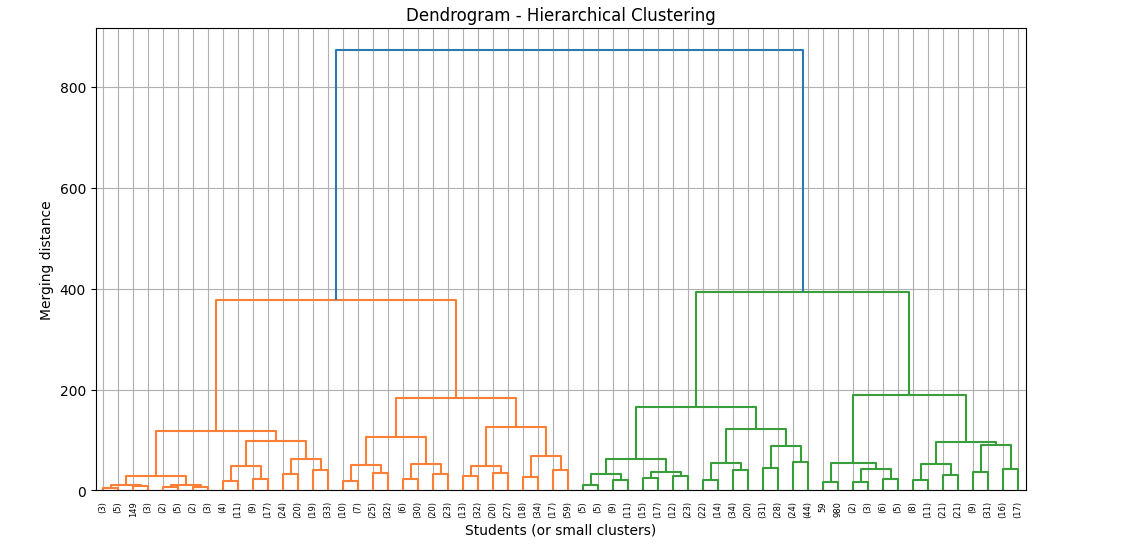
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Cluster | Count | Math Score | Reading Score | Writing Score |
| 0 | 444 | 53.74 | 56.38 | 54.86 |
| 1 | 556 | 75.95 | 79.38 | 78.59 |



Therefore, the algorithm divided the students into two clusters: Cluster 0, consisting of students with low performance, and Cluster 1, consisting of students with high performance

## **2.4 Αgglomerative Clustering Method**

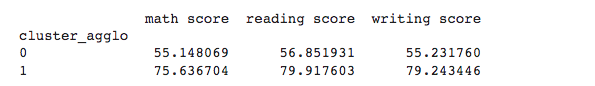
Using the *Ward* linkage method, a dendrogram was generated.  




As observed in the dendrogram, a gap appears between approximately 400 and 600, suggesting the formation of 2 clusters.

### 2.4.1 Αgglomerative Clustering Results

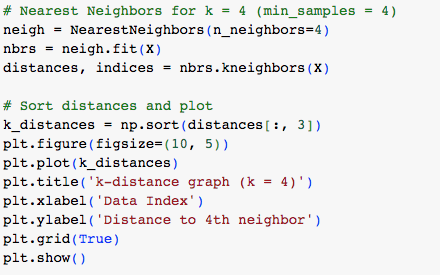
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Cluster | Count | Math Score | Reading Score | Writing Score |
| 0 | 444 | 55.15 | 56.85 | 55.23 |
| 1 | 556 | 75.63 | 79.91 | 79.24 |

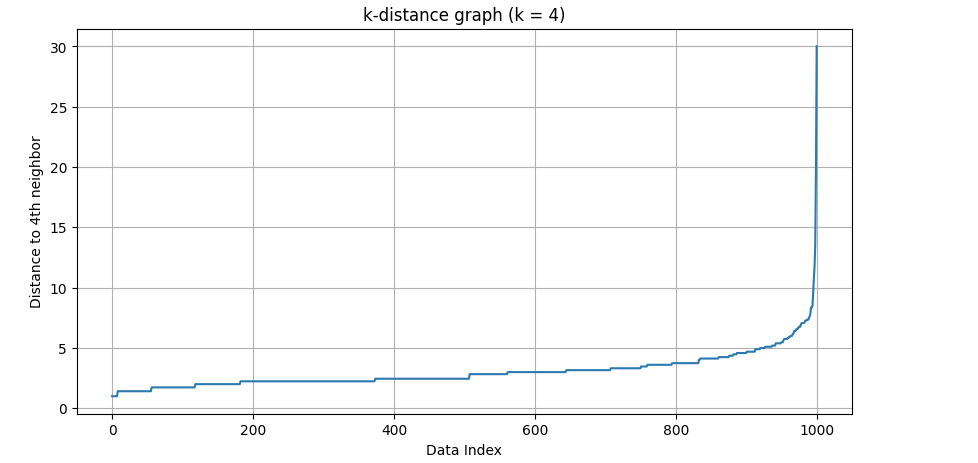
The agglomerative Clustering method also divided the students into two clusters based on their performance levels.

## **2.5 Clustering with DBSCAN Algorithm**

The DBSCAN algorithm does not require the number of clusters to be specified in advance. However, it does require parameter values for *eps* and *min\_samples*.

To determine an appropriate *eps* value, the k-distance graph was generated with *k = 4*, and the "knee" of the curve was used as the chosen value for *eps*. The parameter *min\_samples* was set to 4, meaning that at least four neighboring points are required for a region to be considered dense.





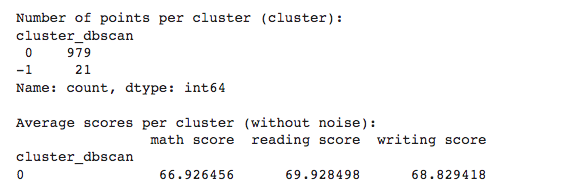
In the plot, the ‘elbow’ appears around 7, therefore we select this as the value of eps

### 2.5.1 Results with DBSCAN Algorithm

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Cluster | Πλήθος | Math Score | Reading Score | Writing Score |
| 0 | 979 | 66.93 | 69.93 | 68.83 |
| -1 (noise) | 21 | - | - | - |



As a result, the algorithm created only one cluster with 979 students and identified 21 students as noise. We observe that DBSCAN did not separate students into high- and low-performance groups, but instead detected a compact group of students with relatively similar scores, while labeling the rest as noise.



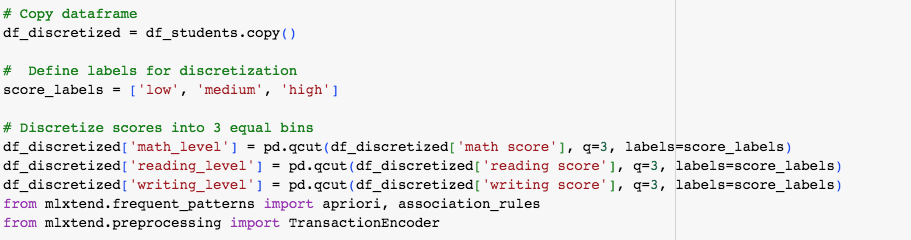
# Association Rule Mining

## **3.1 Introduction**

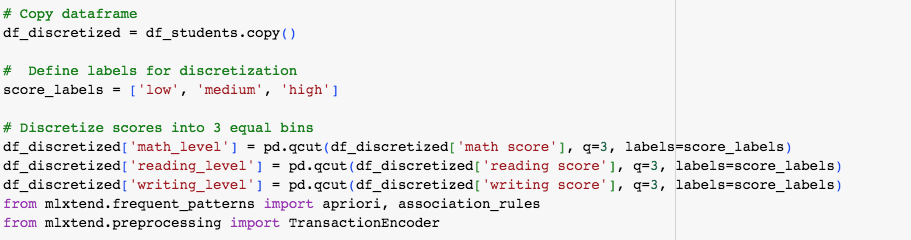
Our goal was to discover association rules in the *StudentsPerformance* dataset using the *Lift* measure.

## **3.2 Data Preparation**

To perform association rule analysis, the three numerical variables were categorized into three classes—‘low’, ‘medium’, and ‘high’—based on quantiles.



Only the categorical data was then selected for analysis.



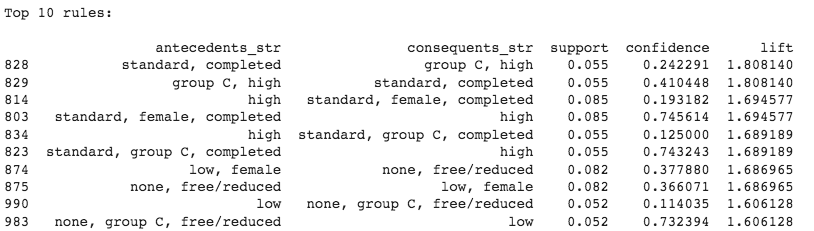
## **3.3 Rule Extraction**

For rule extraction, the Apriori algorithm was applied with a minimum support of 0.05, and *lift* was used to evaluate the significance of the rules.

Macintosh HD:Users:georgemoumoulidis:Desktop:Στιγμιότυπο 2025-08-24, 5.36.14 μμ.png

Since the resulting table was very large and difficult to study, the 10 most significant rules were selected based on the highest lift values.





* According to rule 828, 5.5% of students who have completed the standard belong to Group C with high performance. The confidence value indicates that approximately 24.2% of students with these characteristics are indeed in the high-performance Group C. The lift of 1.808 suggests that the probability of this combination occurring is 1.8 times higher than if the features were independent. Therefore, students who have completed the standard are more likely to belong to Group C and achieve high performance compared to the general population.
* According to rule 829, 5.5% of students who belong to the high-performance group and Group C have prepared for the exam and eat a regular meal. The confidence indicates that approximately 41% of students with these characteristics are in the high-performance Group C. The lift shows that the probability of this combination occurring is 1.8 times higher than if the features were independent. Therefore, students who are in the high-performance Group C are more likely to have prepared for the exam and eat a regular meal compared to the general population.
* According to rule 814, 8.5% of students have high performance. Of these, approximately 19.3% are girls who have prepared for the exam and eat regular meals. The lift indicates that the probability of this combination occurring is 1.7 times higher than if the features were independent. Therefore, high-performing students are more likely to be girls who have prepared for the exam and eat regular meals compared to the general population.
* According to rule 803, 8.5% of students are girls who have prepared for the exam and eat regular meals. Of these, approximately 74.6% achieve high performance. The lift indicates that the probability of this combination occurring is 1.7 times higher than if the features were independent. Therefore, students who have prepared, are girls, and eat regular meals are more likely to achieve high performance compared to the general population.
* According to rule 834, 5.5% of students have prepared for the exam, belong to Group C, and eat regular meals. Of these, approximately 74.3% achieve high performance. The lift indicates that the probability of this combination occurring is 1.69 times higher than if the features were independent. Therefore, students who have prepared, belong to Group C, and eat regular meals are more likely to achieve high performance compared to the general population.
* According to rule 823, 5.5% of high-performing students belong to Group C and eat regular meals. Of these, approximately 12.5% achieve high performance. The lift indicates that the probability of this combination occurring is 1.69 times higher than if the features were independent. However, the correlation is not very strong due to the low confidence value.
* According to rule 874, 8.2% of students who skip meals or have reduced meals have not prepared for the exam. Of these, approximately 36.6% have low performance and are girls. The lift indicates that the probability of this combination occurring is 1.69 times higher than if the features were independent. However, the confidence is moderate.
* According to rule 875, 8.2% of low-performing students are girls. Of these, approximately 37.7% eat very little or nothing. The lift indicates that the probability of this combination occurring is 1.69 times higher than if the features were independent. However, the correlation is moderate.
* According to rule 990 , 5.2% of students have low performance. Of these, approximately 11.4% eat very little or nothing, have not prepared for the exam, and belong to Group C. The lift indicates that the probability of this combination occurring is 1.61 times higher than if the features were independent. However, the confidence is low, so the rule has weak strength.
* According to rule 983, 5.2% of students eat very little or nothing, have not prepared for the exam, and belong to Group C. Of these, approximately 73.2% have low performance. The lift indicates that the probability of this combination occurring is 1.61 times higher than if the features were independent. However, the confidence is high, so the rule is strong.