****

**FORE SCHOOL OF MANAGEMENT**

**NEW DELHI**

**Academic Session: 2023-2025**

**Machine Learning for Managers – II**

**“Segmentation of Employee Compensation Data using Supervised Machine Learning Algorithms”**

**PGDM 32 Section: A**

**Submitted to: Submitted by:**

**Prof. Amarnath Mitra Arushi Khanna**

**Roll No.: 321012**

**TABLE OF CONTENTS**

|  |  |
| --- | --- |
| **Sr. No.** | **Content** |
| 1. | Project Objective |
| 2. | Data Description |
| 3 | Data Analysis |
| 4. | Results and Observation |
| 5. | Managerial Insights |

1. **Project Objective**

→ Segmentation of Employee Compensation Data using Supervised Machine Learning Algorithms such as Decision Tree.

→ The second objective is to determine the number of appropriate classification model by comparing and contrast using logistic regression, KNN (k-nearest neighbor) and SVM (support vector machine).

→ The third objective is to identify significant variables or features and their thresholds for classification

1. **Data Description**

2.1 Dimension of Data

2.1.1 Data Source: https://www.kaggle.com/datasets/san-francisco/sf-employee-compensation

2.1.2 Data Size: 26 MB (Kaggle), 153 MB (Excel csv File)

2.1.3 Number of Variables: The number of variables in the csv file is 22.

2.1.4 Number of records: The number of records in the csv file is 6,83,277 (excluding naming column).

**2.2 Description of Variables**

Index Variables:

Organization Group Code: Gives the organisation groups an identification

Job Family Code: Gives the Job families an identification

Job Code: Gives the particular from a Job Family an identification

Year: The particular year for the variables and records

Department Code: An identifier for a department in an organisation

Union Code: An identifier for a particular union

Employee Identifier: An employee’s identification code

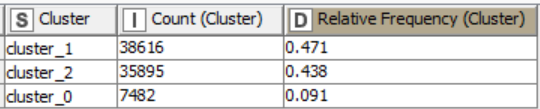
**Categorical Variables:**

|  |
| --- |
| Year Type |
| Organization Group |
| Department |
| Union |
| Job Family |
| Job  **Non – Categorical Variables:** |
| Salaries |
| Overtime |
| Other Salaries |
| Total Salary |
| Retirement |
| Health and Dental |
| Other Benefits |
| Total Benefits |
| Total Compensation |

2.3 **Descriptive Statistics**

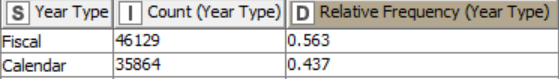
**2.3.1. Descriptive Statistics of Outcome Categorical Variables**

It provides the statistics of cluster variable (categorical variable) by giving frequency as well as relative frequency

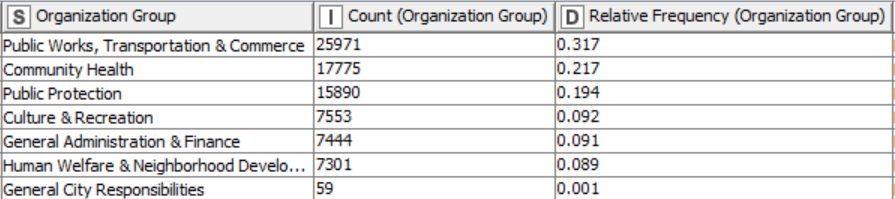
****

**2.3.2. Descriptive Statistics of Input Categorical Variables**

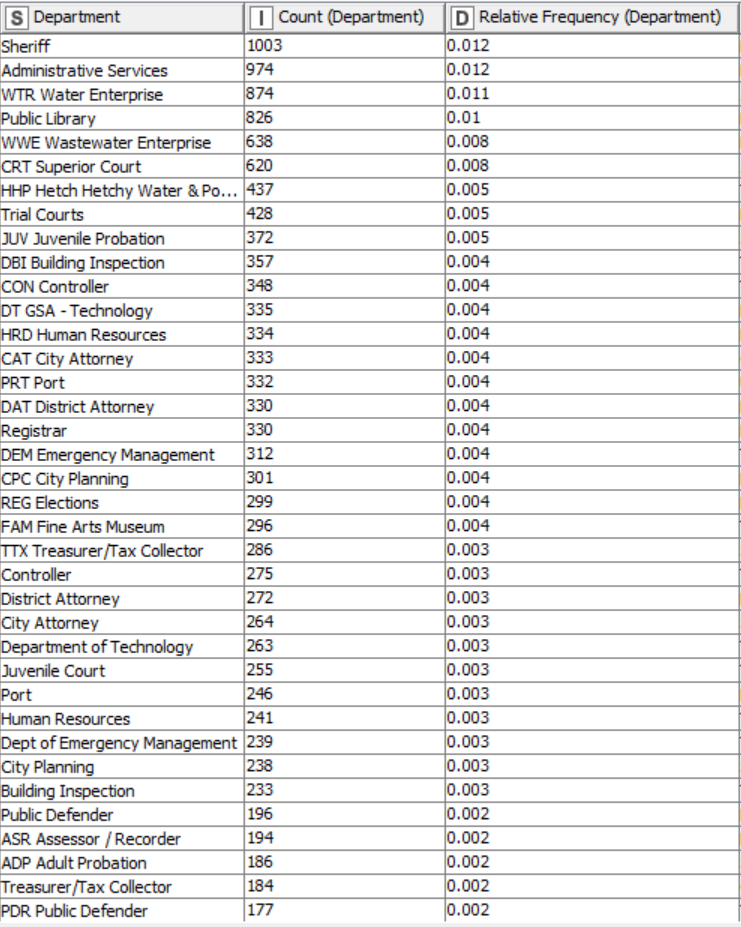
2.3.2.1 Year type

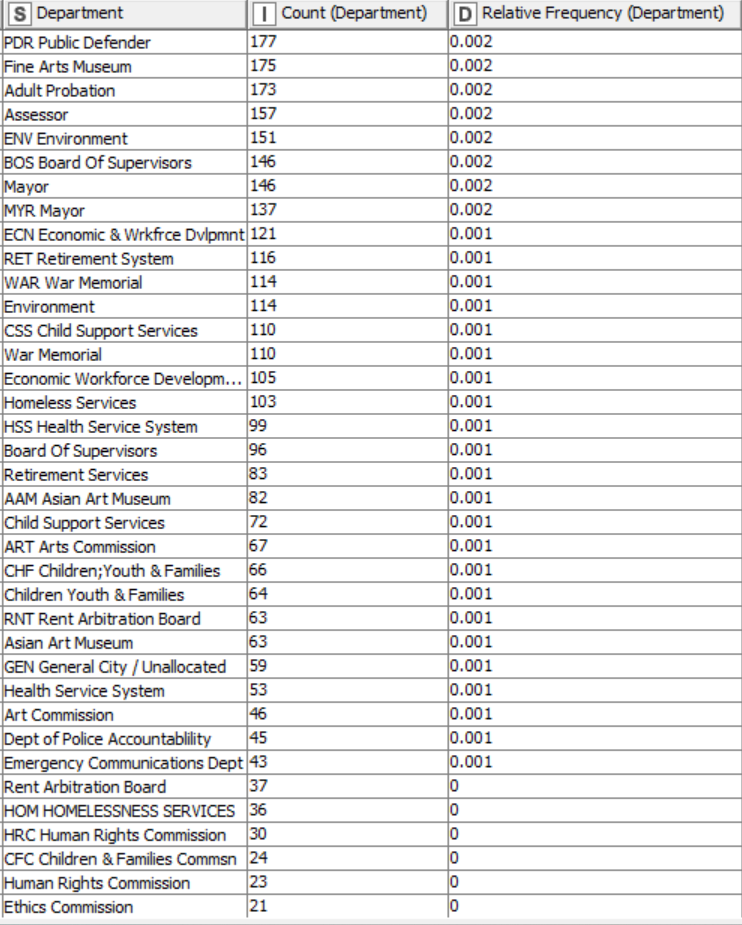


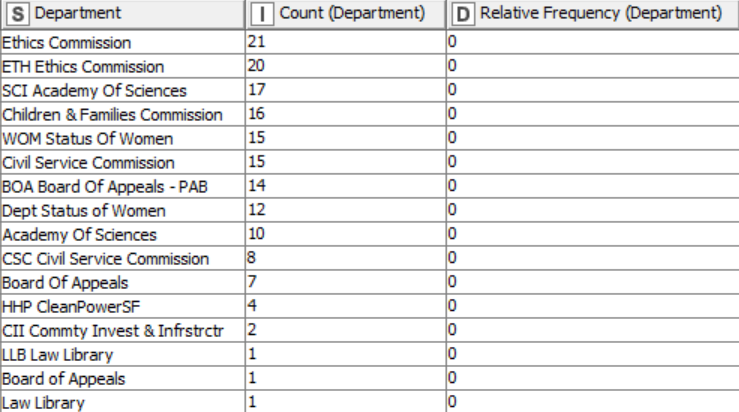
2.3.2.2 Organisation Group



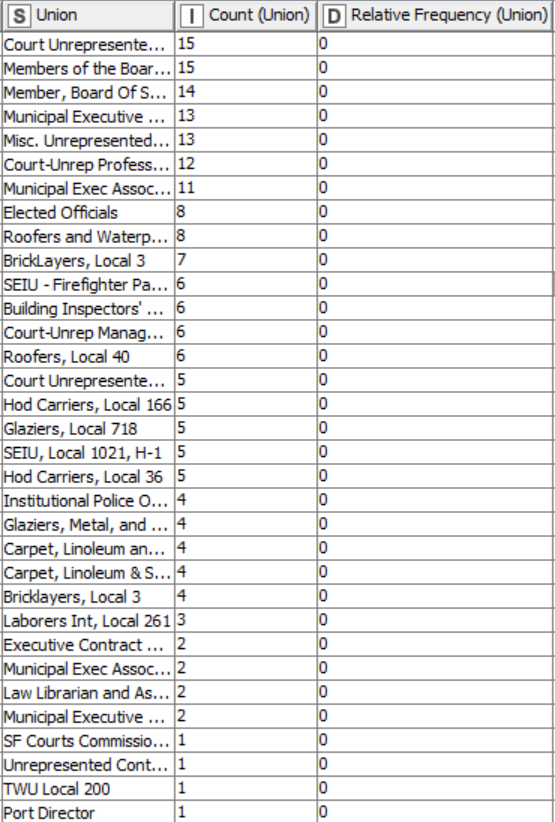
2.3.2.3 Department



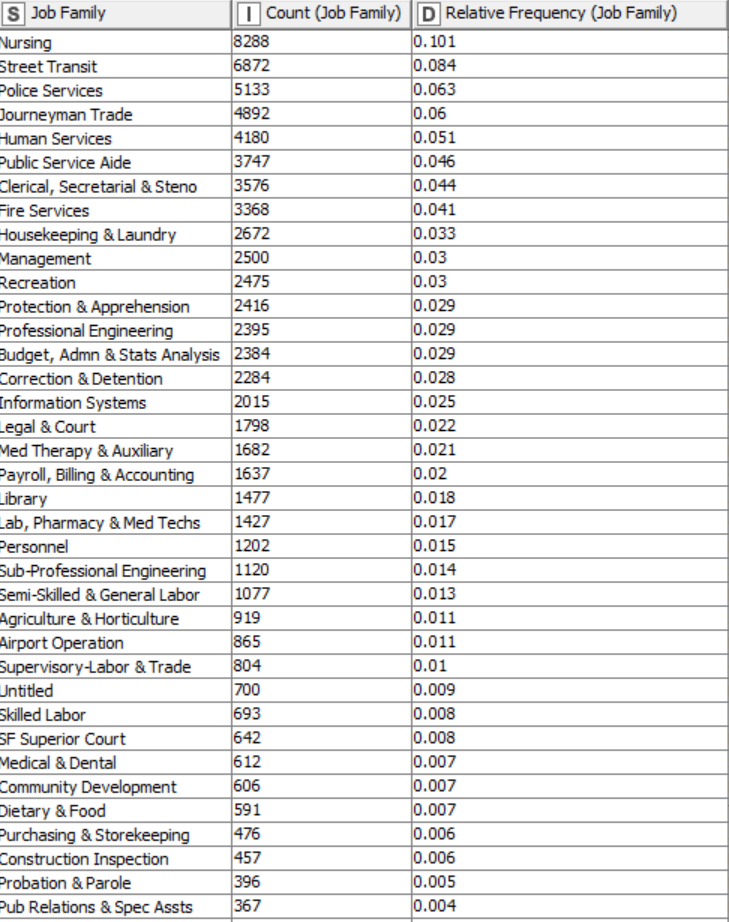


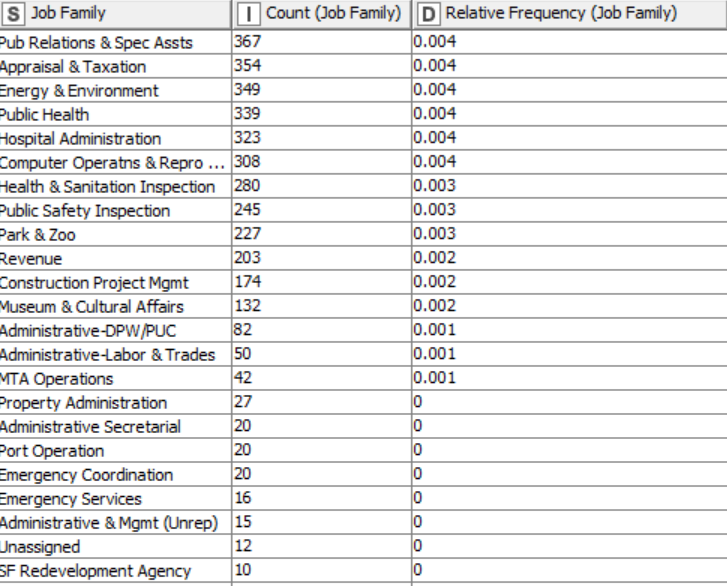


2.3.2.4 Union

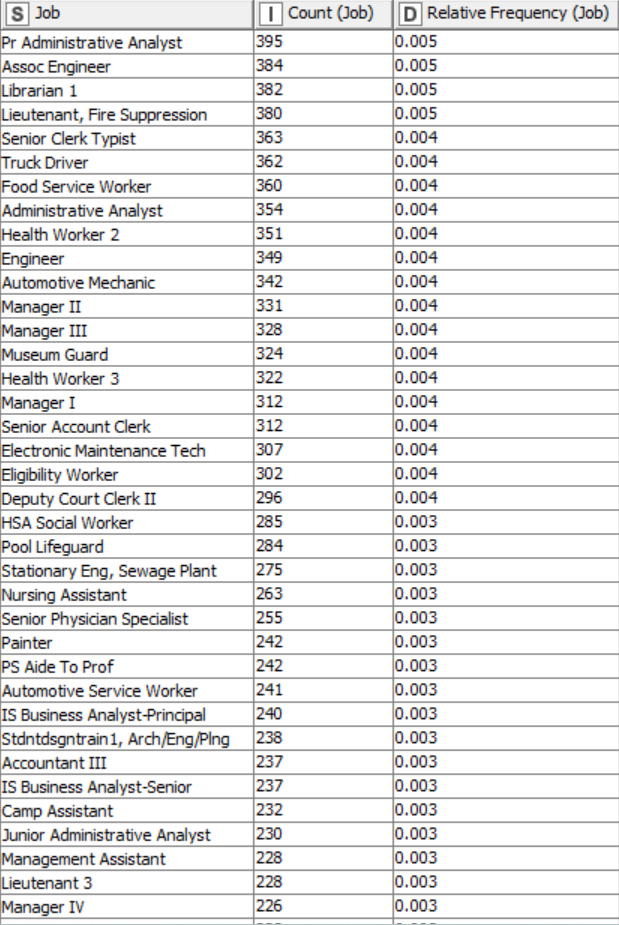


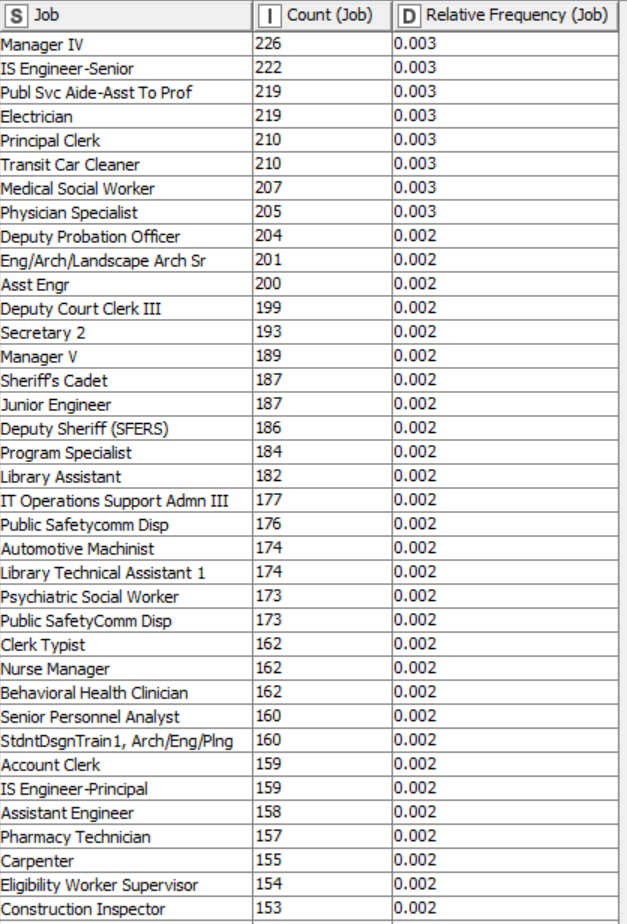
2.3.2.5 Job Family





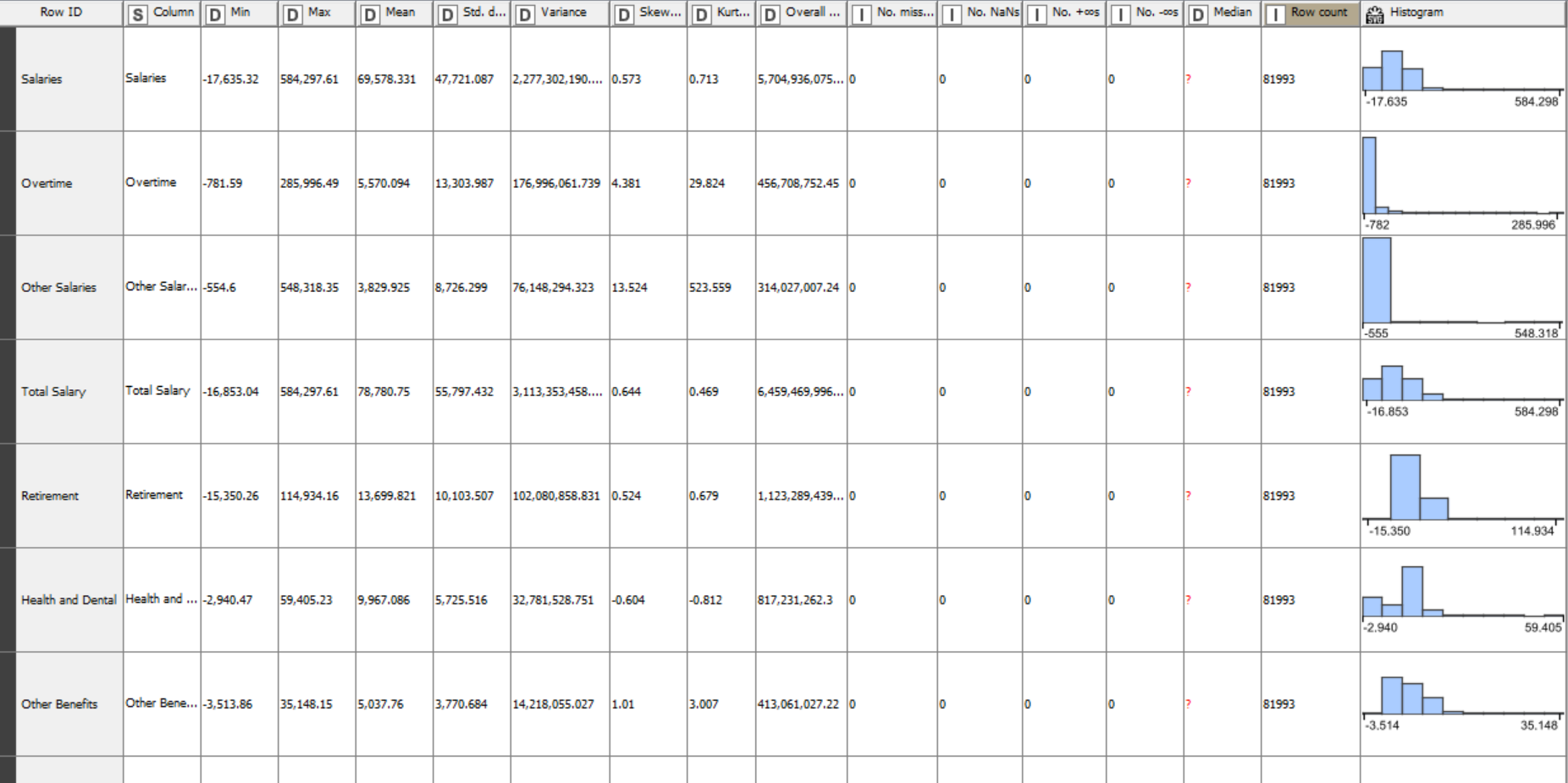
2.3.2.6 Job

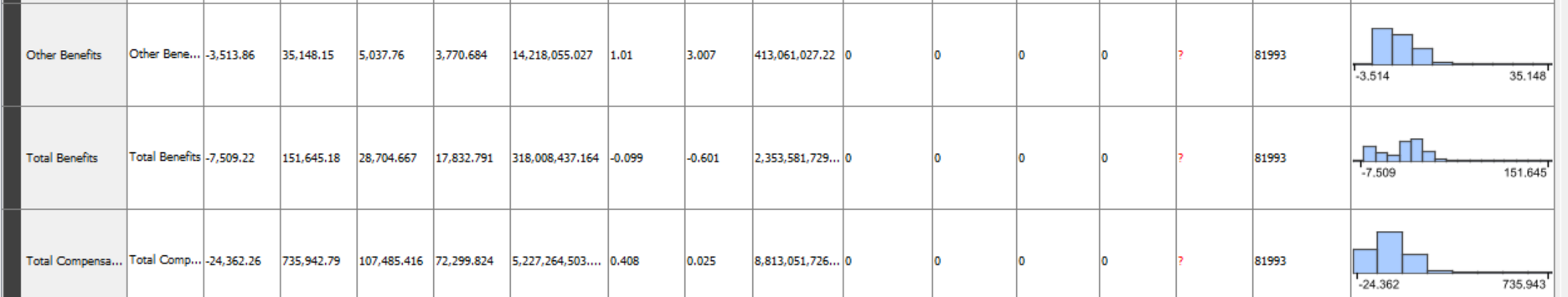




And so on and so forth

2.3.3 Descriptive Statistics: Non-Categorical Variables





1. **Analysis Of Data**

3.1 Data Pre-Processing

3.1.1 Missing Data Statistics and Treatment

3.1.1.1 Missing Data Statistics: 0

3.1.1.2 Missing Data Treatment: 0

3.1.1.2.1 Removal of Records with More Than 50% Missing Data: None

3.1.1.3 Missing Data Statistics of categorical Variables: 0

3.1.1.3.1 Missing Data Treatment: Categorical Variables or Features: 0

3.1.1.3.1.1 Removal of Variables or Features with More Than 50% Missing Data: None

3.1.1.4 Missing Data Statistics of non-categorical Variables: 0

3.1.1.4.1 Missing Data Treatment of non-categorical Variables: 0

3.1.1.4.1.1 Removal of Variables or Features with More Than 50% Missing Data: None

**3.1.2 Numerical Encoding of Variables**

In this case, category to variable node will be used to encode categorical data into numbers such as:

Year Type:

Calendar = 0, Fiscal = 1

Organization Group:

Public Protection = 0

Public Works, Transportation & Commerce = 1

Human Welfare & Neighbourhood Development = 2

Community Health = 3

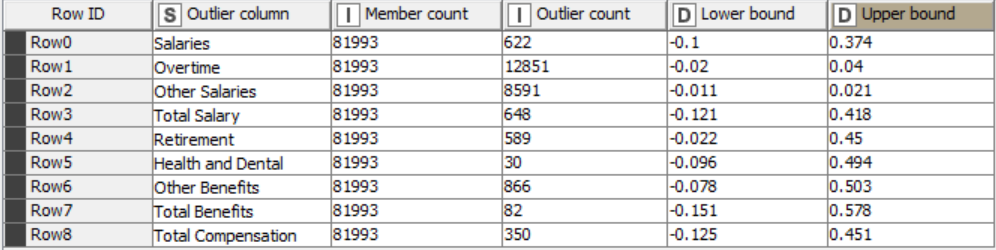
Culture & Recreation = 4

General Administration & Finance = 5

And so on and so forth

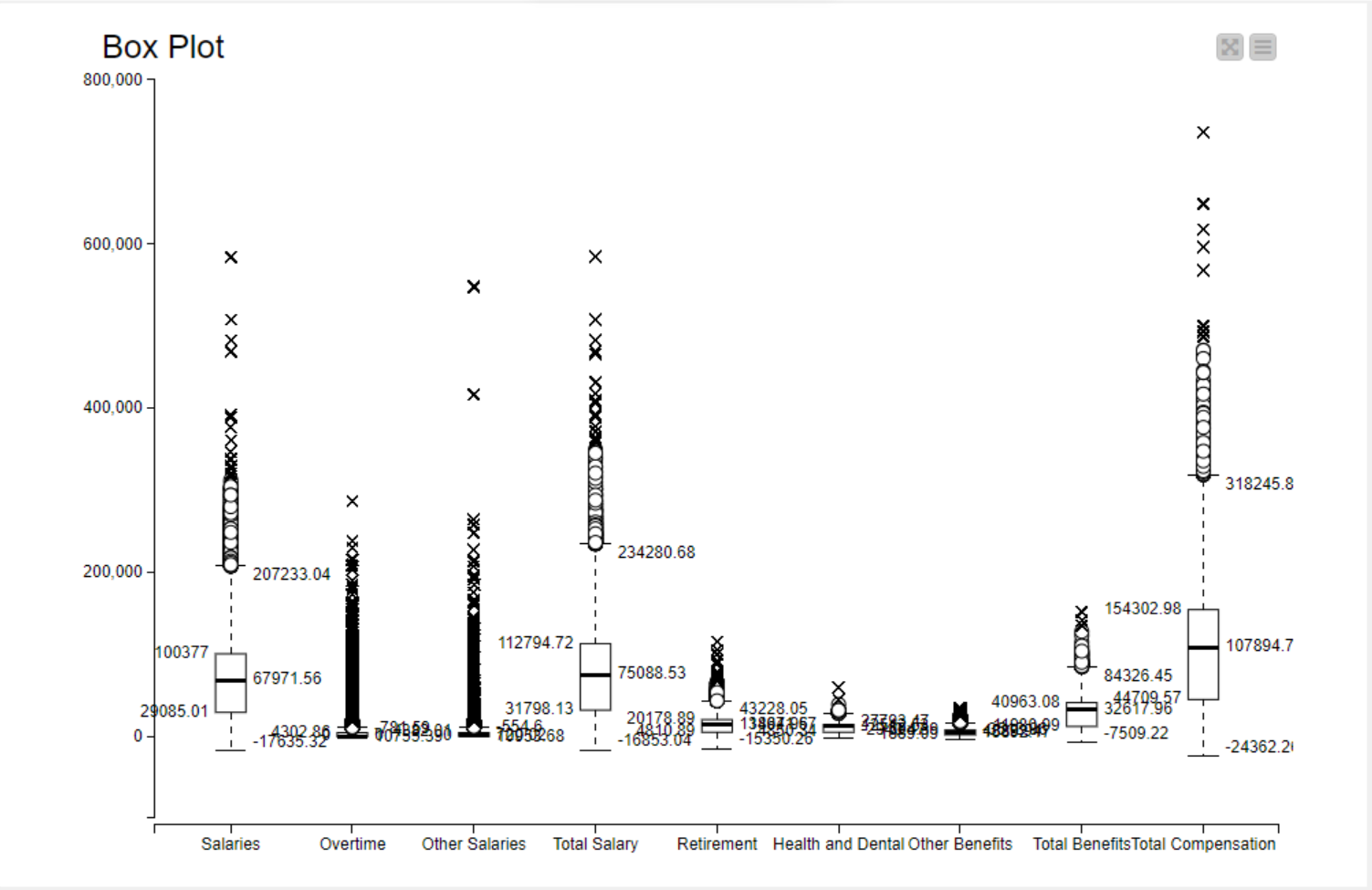
3.1.3 Outlier Statistics Treatment

3.1.3.1 Outlier Statistics: Non-Categorical Variables

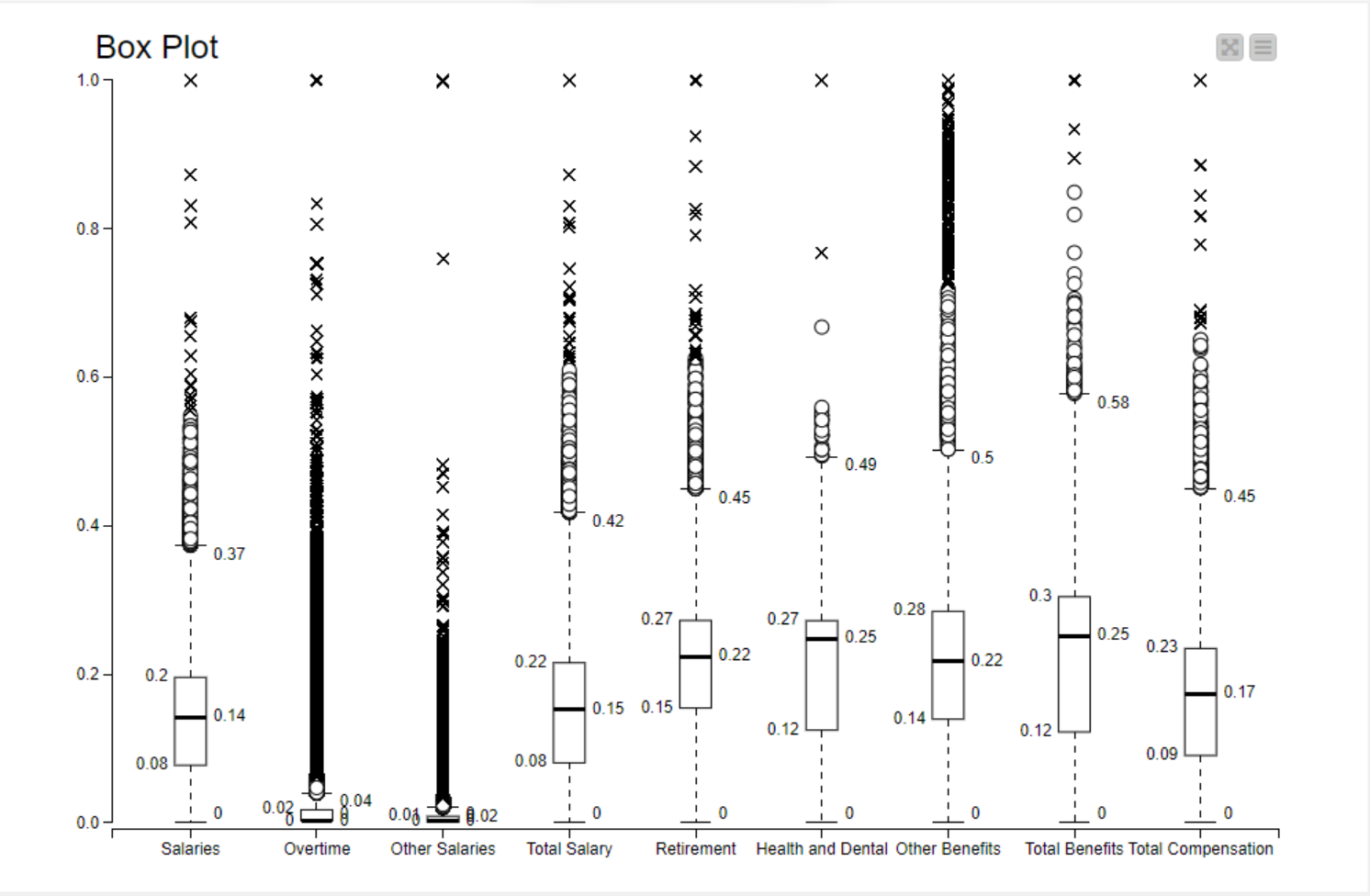


3.1.3.2 Normalization using Min-Max Scaler

Before Normalization

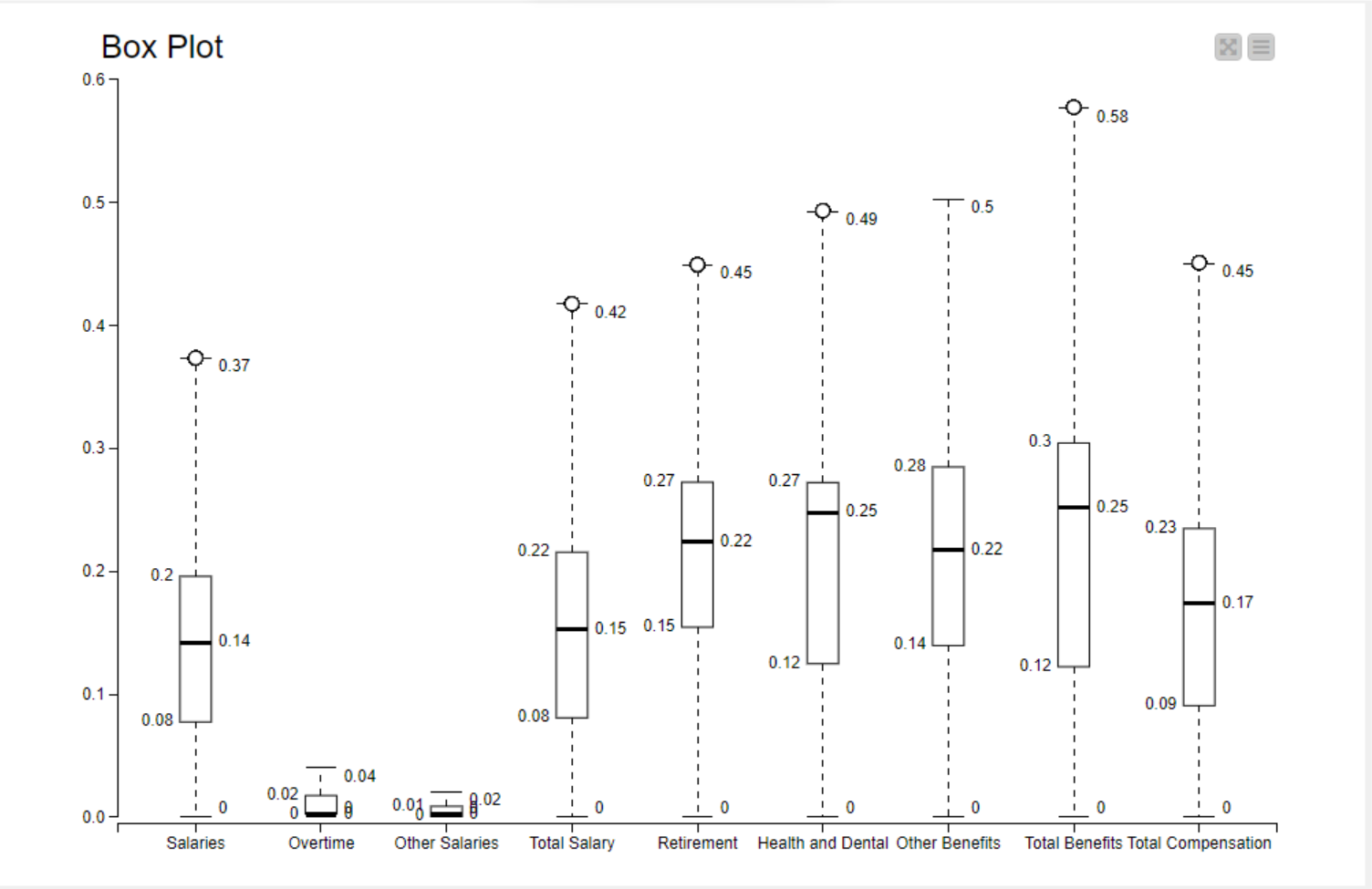


After Normalization



After treating Outliers

Outlier Treatment using replacement strategy where the values are replaced to the closest permitted value



**3.2. Data Analysis**

3.2.1 Supervised Machine Learning Algorithm

3.2.1.1. Supervised Machine Learning Classification Algorithm: Decision Tree

➔ A decision tree is a supervised machine learning algorithm used for both classification and regression tasks. It works by recursively partitioning the input space into smaller regions based on feature values, creating a tree-like structure of decisions. At each node of the tree a decision is made based on the value of a specific feature, and the data is split into subsets. This process continues until a stopping criterion is met, such as reaching a maximum depth or no further improvement in impurity reduction.

➔ In this project, decision tree will be the classification algorithm used for unsupervised learning. The metrics used in decision tree is Gini coefficient.

➔ When using decision tree, we will be also seeing comparison when no pruning method is used and when pruning method is used.

**3.2.1.2. Supervised Machine Learning Classification: Other Methods**

**Logistic Regression**

It is a supervised learning algorithm used for binary classification tasks. It models the probability of the input belonging to a particular class using the logistic function. The algorithm learns the relationship between input features and the probability of the binary outcome, making it suitable for predicting categorical outcomes. In this project, logistic regression will be used and the metric used in logistic regression is iteratively reweighted least squares (solver method).

**K-Nearest Neighbours**

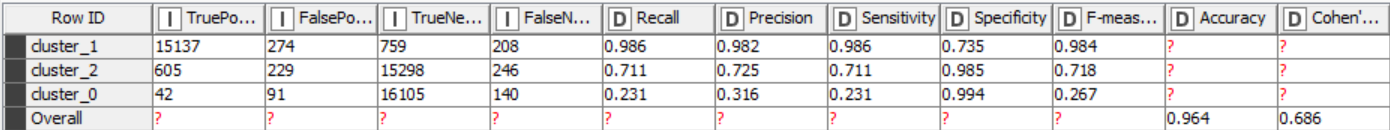
K-Nearest Neighbours (KNN) is a supervised learning algorithm that is also used for both classification and regression tasks. It predicts the classification of a data point by finding the majority class among its k nearest neighbours in the feature space. KNN's performance heavily depends on the choice of distance metric and the value of k, making it sensitive to the dataset's characteristics. In this project, KNN will be used and the metric used is Euclidean distance. For comparison, we will be using k =7 till k=19 in steps of 2 i.e. k=7,9,11,13,15,17 and 19.

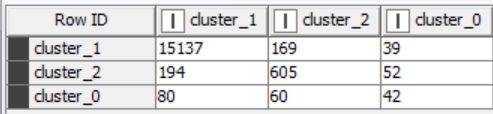
**Support Vector Machines**

Support Vector Machine (SVM) is a powerful supervised learning algorithm used for classification and regression tasks. It works by finding the hyperplane that best separates the classes in the feature space, maximizing the margin between them. SVM can handle high-dimensional data and is effective even in cases where the number of features exceeds the number of samples. In this project, the kernel used will be polynomial and the parameters are power = 1, bias = 1 and gamma = 1.

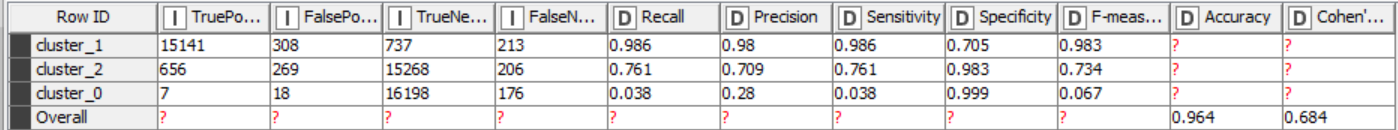
**3.2.2.1.1. Classification Model Performance Evaluation of Decision Tree by using Confusion Matrix**

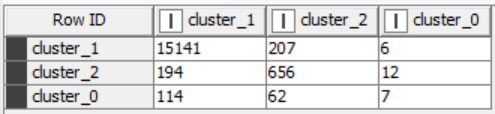
Without Pruning





With Pruning





Cluster 1

• This cluster has a high number of true positives and true negatives indicating that the model correctly classified most instances within this cluster.

• The precision and recall scores are both very high suggesting that the model effectively identifies true positives while also minimizing false positives.

Cluster 2

• This cluster has a lower recall and precision compared to cluster 0, indicating that the model's performance is not as strong for this segment.

• The number of false positives is relatively high, suggesting that the model may misclassify some instances within this cluster.

• Despite the lower performance metrics, the specificity is very high indicating that the model correctly identifies true negatives within this cluster.

Cluster 0

• This cluster has the lowest recall and precision.

• The number of false positives is relatively low suggesting that the model effectively minimizes misclassifications within this cluster.

• Specificity scores are high indicating that the model correctly identifies the true negatives within this cluster.

**Comparative analysis of decision tree with and without pruning**

• Pruning generally improves precision and specificity while slightly reducing recall and sensitivity.

• In cluster\_1, the precision and specificity decreased with pruning but the recall is the same, indicating that the model becomes less conservative in predicting positive cases resulting in more positives but also more true positives.

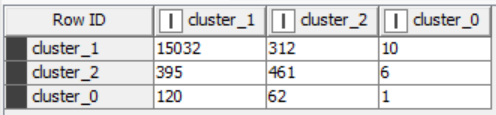
• In cluster\_0 and cluster\_2, pruning led to a slight decrease in precision and very insignificant change in specificity, change in recall and sensitivity indicating that the model becomes more conservative resulting in less true positives but also less false positives.

• The choice of whether to prune the decision tree depends on the specific requirements of the problem and the trade-off between precision and recall. If minimizing false positives is crucial (can be used for risk assessment) pruning may be preferred. If capturing as many true positives as possible is more important (can be used for customer retention) pruning may be avoided.

The logistic regression model achieves high accuracy and Cohen's Kappa showing its effectiveness in classifying instances into the correct clusters

**3.2.2.2. Classification Model Performance Evaluation of Other Supervised Learning methods by using confusion matrix**

**Logistic Regression**

****

The confusion matrix shows the performance of a logistic regression model classifying data points into three classes: cluster\_0, cluster\_1 and cluster\_2. Here's a breakdown of the table:

**Values in each cell represent the count of data points that fall into that classification.**

**Looking at the values in the table:**

* **15032:** Out of the data points that actually belong to cluster\_1 (positive class), the model correctly classified 15032 of them (True Positives).
* **312:** The model incorrectly classified 312 data points that don't belong to cluster\_1 (negative class) as cluster\_1 (False Positives).
* **10:** Out of the data points that don't belong to cluster\_1, the model correctly classified 10 of them (True Negatives). This number is quite low compared to the other values, indicating the model might be struggling to identify points that are not cluster\_1.

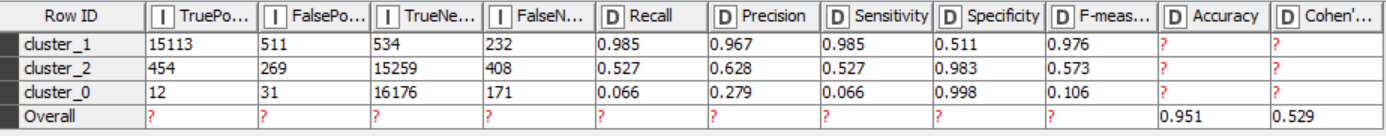
The same could be followed up for cluster\_2 and cluster\_0.

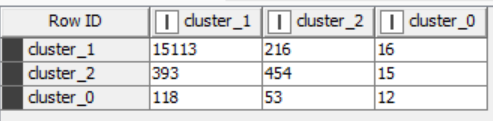
**Overall Performance:**

* Based on these values, the model seems to be good at identifying actual cluster\_1 data points (high True Positives). However, it's also making a significant number of mistakes (False Positives) by classifying data points that don't belong to cluster\_1 as cluster\_1. It's also missing some actual cluster\_1 data points (False Negatives).

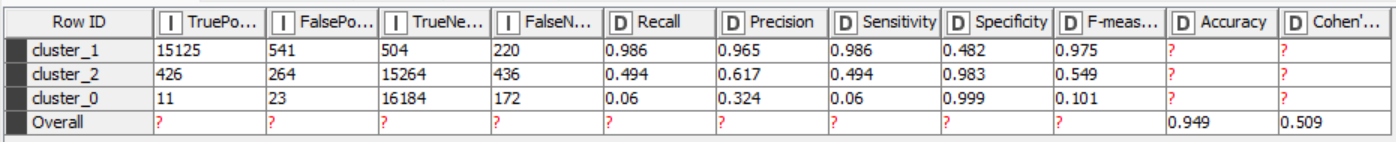
**K nearest Neighbor**

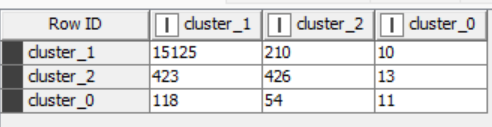
K = 7



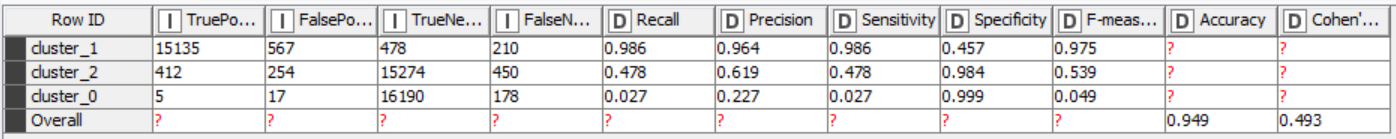


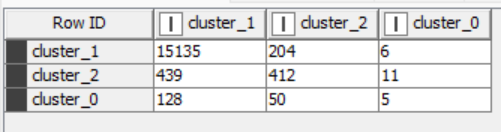
K = 9



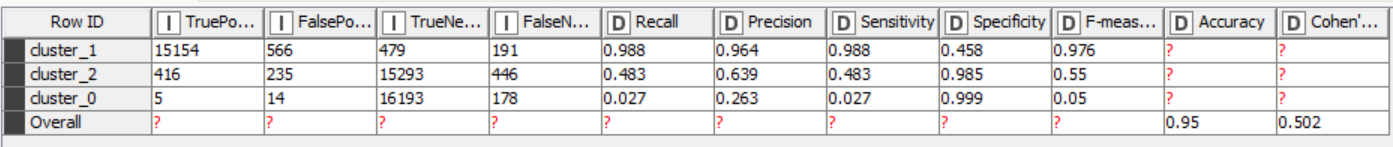


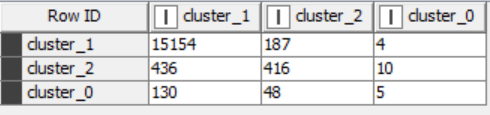
K = 11



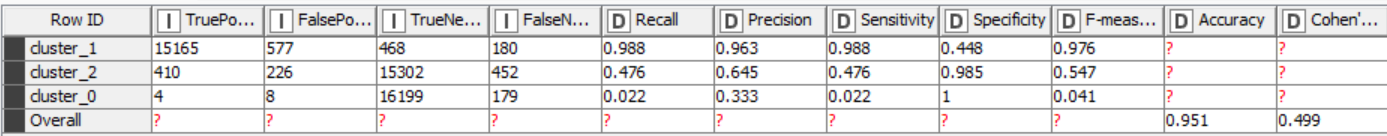


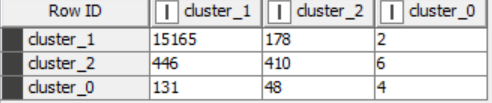
K = 13



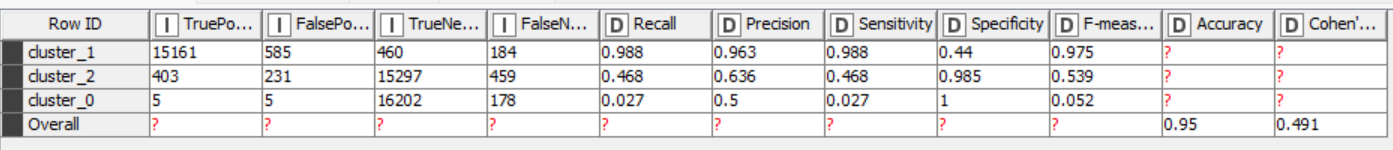


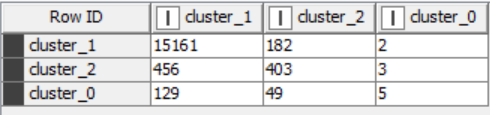
K = 15



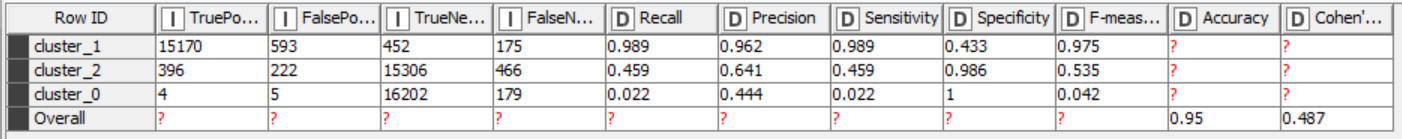


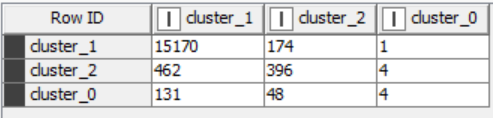
K = 17





K = 19





In KNN, the number of neighbors to be considered are from k=7 to 19. From the images, it is seen that as the number of k increases the accuracy also increases. For k=19, as the accuracy is the highest from all the other k’s, this cluster will be considered.

Cluster\_1

• True Positives: 15170, False Positives: 593, True Negatives: 452, False Negatives: 175

• Recall: 0.989, Precision: 0.962, Sensitivity: 0.989, Specificity: 0.433

• F-measure: 0.975, Accuracy: 0.95

In cluster\_0, the KNN model achieved high recall indicating that it effectively identifies true positives within this cluster. However, the precision is relatively high emphasizing a lower rate of false positives. The model's specificity is extremely low indicating that it poorly identifies true negatives.

The overall accuracy is moderate which shows that the model's performance may vary across different metrics.

Cluster \_2

• True Positives: 396, False Positives: 222, True Negatives: 15306, False Negatives: 466

• Recall: 0.459 , Precision: 0.641, Sensitivity: 0.459, Specificity: 0.986

• F-measure: 0.535, Accuracy: 0.95.

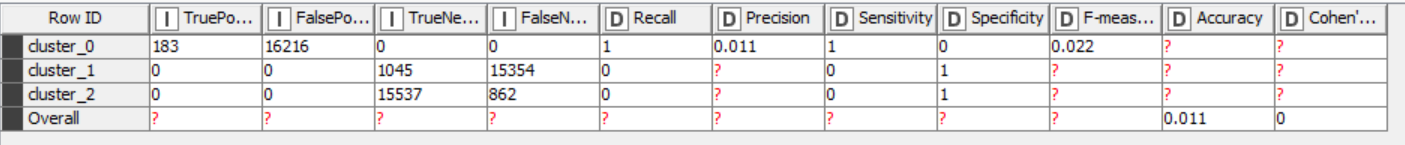
Cluster \_0

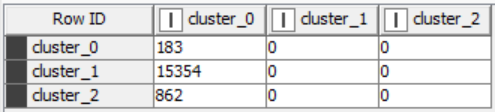
• True Positives: 4, False Positives: 5, True Negatives: 16202, False Negatives: 179

• Recall: 0.022, Precision: 0.444, Sensitivity: 0.022, Specificity: 1

• F-measure: 0.042, Accuracy: 0.95 In cluster\_0 the KNN model has low recall and precision indicating that it struggles to correctly classify instances within this cluster. However, the model exhibits high specificity showing a strong ability to identify true negatives. The overall accuracy is moderate reflecting the model's mixed performance across different metrics. The overall accuracy of the KNN model is moderate showing mixed performance across different clusters. However, Cohen's Kappa coefficient suggests very low agreement beyond chance among the predicted and actual cluster labels

**Support Vector Machines**

****

****

Cluster\_0

• True Positives: 183, False Positives: 16216, True Negatives: 0, False Negatives: 0 • Recall: 1, Precision: 0.011, Sensitivity: 1, Specificity: 0

• F-measure: 0.022, Accuracy: 0.011 In cluster\_0, the SVM model achieved perfect recall and sensitivity. It correctly identifies all positive instances within this cluster. However, the precision is extremely low as it has a high rate of false positives. The model's specificity is also very low as it poorly identifies true negatives. Overall accuracy is very low suggesting the poor performance of the model in correctly classifying instances within this cluster.

Cluster \_1

• True Positives: 0, False Positives: 0, True Negatives: 1045, False Negatives: 15354 • Recall: 0, Precision: -, Sensitivity: 0, Specificity: 1

• F-measure: -, Accuracy: - Similar to cluster\_2, the SVM model failed to identify any positive instances (true positives) for cluster\_1. Therefore, precision, F-measure and accuracy metrics are not provided. However, the specificity is 1 indicating that the model effectively identifies instances not belonging to cluster\_1.

Cluster\_2

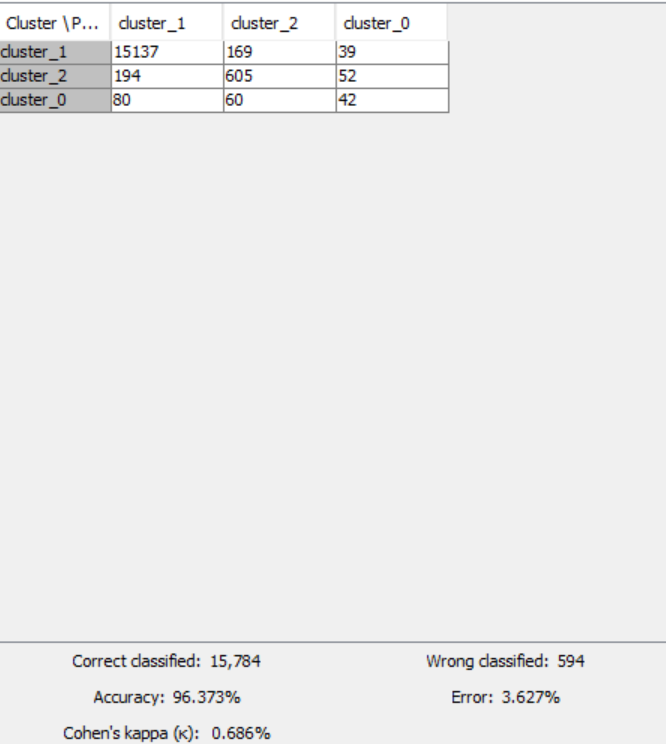
• True Positives: 0, False Positives: 0, True Negatives: 15537, False Negatives: 862• Recall: 0, Precision: -, Sensitivity: 0, Specificity: 1

• F-measure: -, Accuracy: - In cluster\_2, the SVM model failed to identify any positive instances (true positives). The precision, F-measure and accuracy for this cluster are not provided due to the absence of true positives. However, the specificity is 1 indicating that the model effectively identifies instances not belonging to cluster\_2.

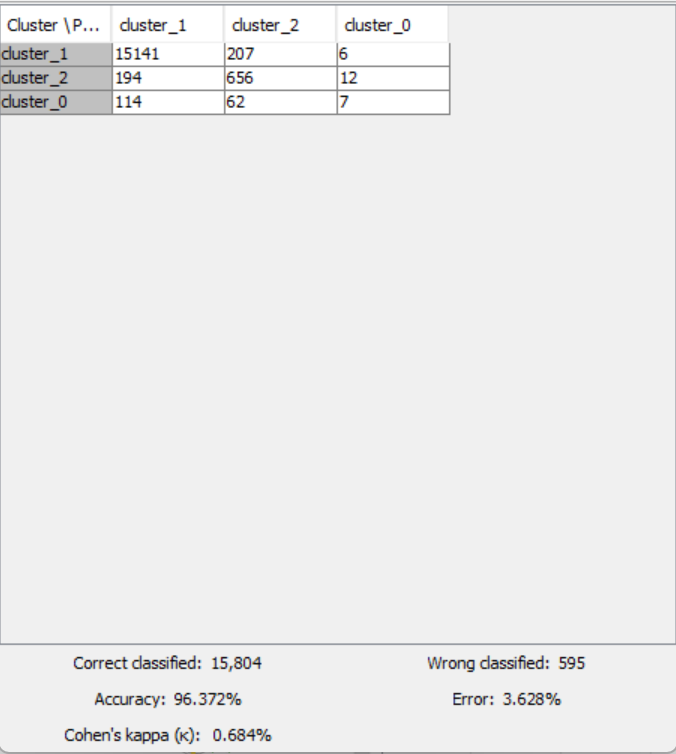
The overall performance of the SVM model is very poor with extremely low recall, precision and accuracy metrics. The absence of true positives in cluster\_2 and cluster\_1 severely impacts the model's ability to provide meaningful insights or make accurate predictions for these clusters

1. **Results and observation**

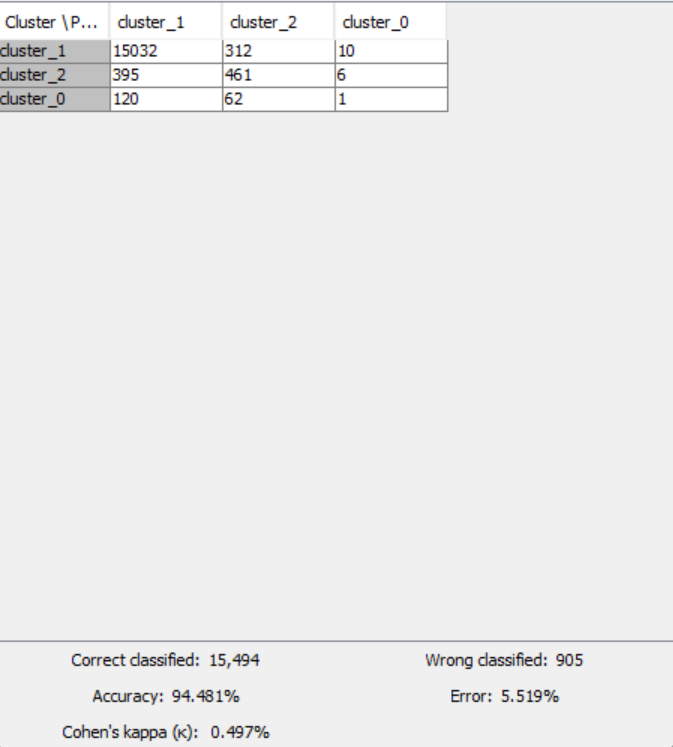
**Decision Tree (without Pruning)**

****

**Decision Tree (with Pruning)**

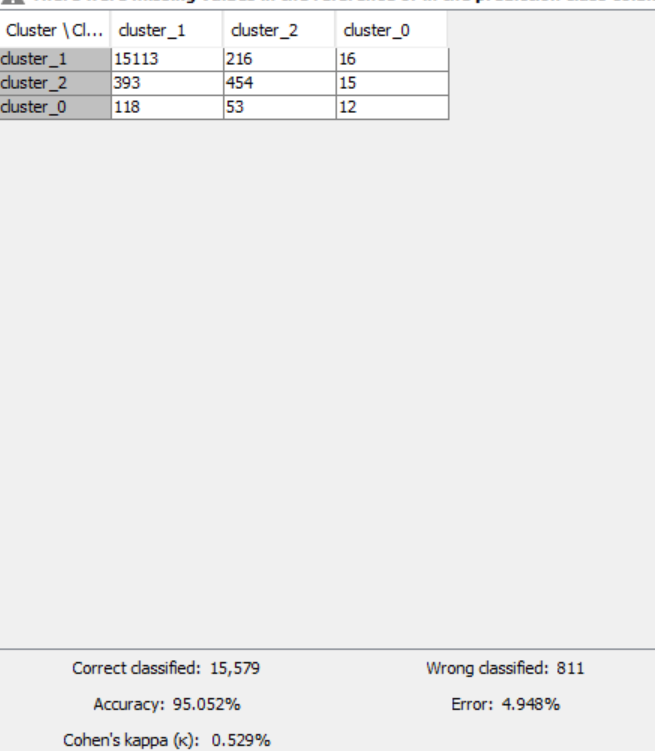
****

**Logistic Regression**

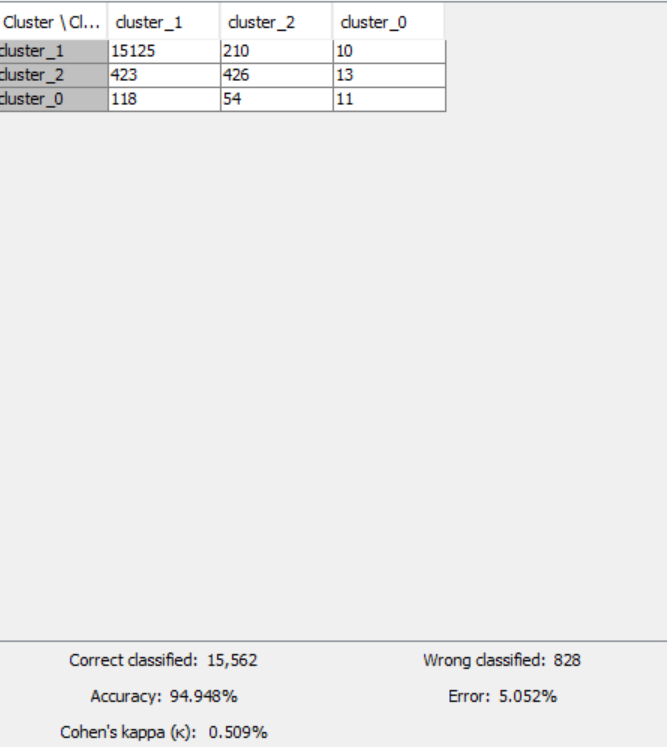
****

**KNN**

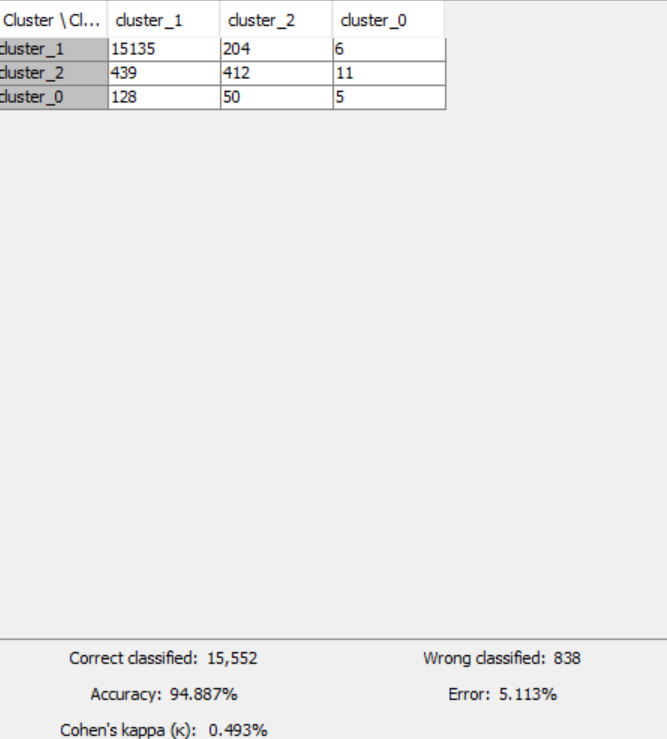
K = 7



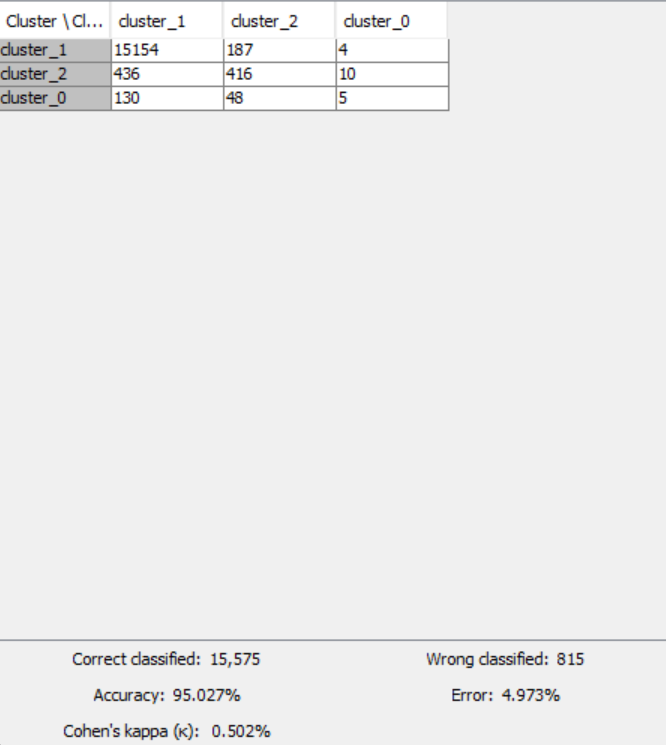
K = 9



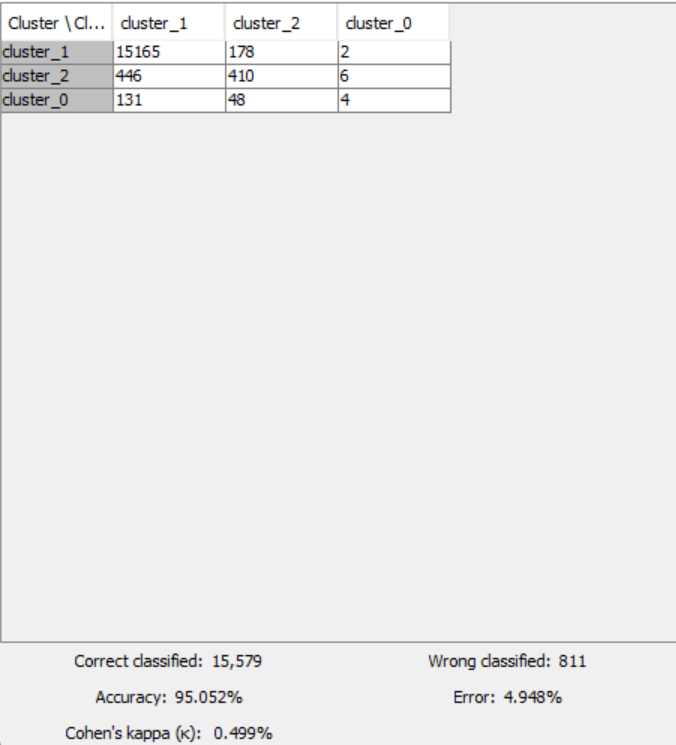
K = 11



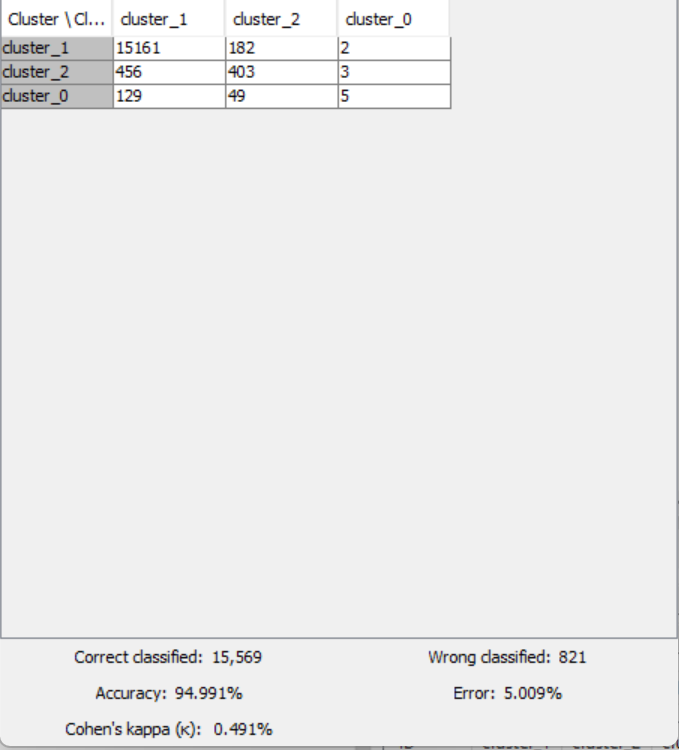
K =13



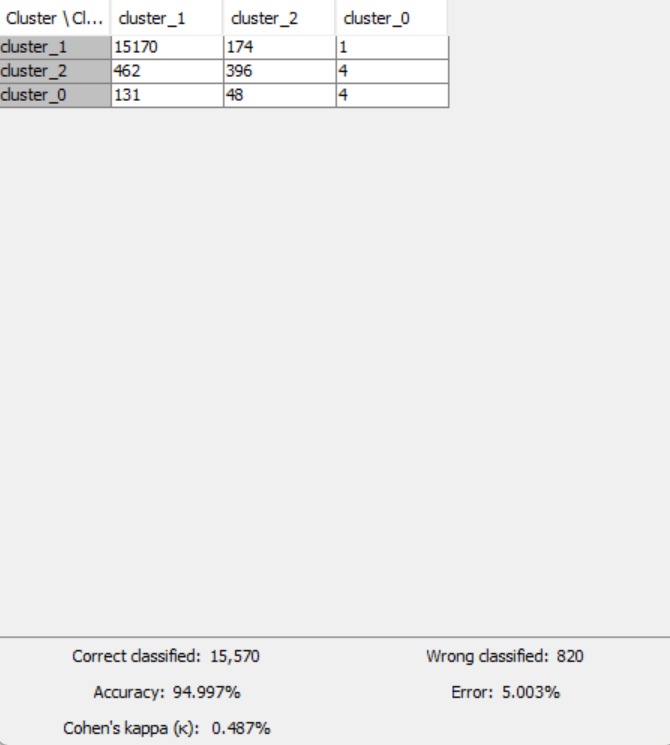
K =15



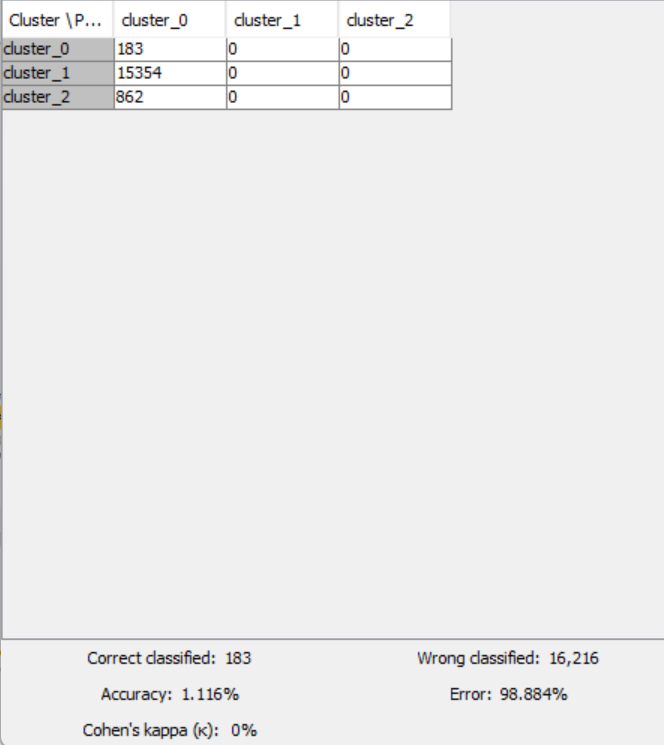
K =17



K =19



SVM



• Decision Tree models (with and without pruning), KNN and logistic regression demonstrate high accuracy and Cohen's Kappa values which shows a robust performance in classification.

• Decision Tree with pruning slightly outperforms the decision tree without pruning showcasing the importance of pruning to avoid overfitting.

• Logistic Regression also performs well, comparable to decision trees indicating its suitability for classification tasks.

• KNN with k=19 shows relatively lower accuracy and Cohen's Kappa values compared to other models suggesting its limitations in handling this particular dataset effectively.

• SVM demonstrates extremely poor performance with an accuracy of just over 1% and no Cohen's Kappa indicating a failure to effectively classify instances in this dataset.

• Overall, decision tree (with or without pruning) and logistic regression and KNN are recommended for this dataset due to their high accuracy and reliable performance. SVM are not suitable for this dataset based on the provided results.

1. **Managerial Insights**

Employee compensation data is a valuable asset for managers to understand their workforce, make informed decisions, and promote a fair and competitive work environment. Here are some key managerial insights you can glean from this data:

**1. Identify Pay Equity:**

* Analyze salary ranges for similar positions, experience levels, and demographics (gender, race, etc.) to identify any potential pay gaps.
* Use metrics like compensation ratios to compare individual salaries to the midpoint of their pay range.
* Investigate and address any pay discrepancies that can't be justified by performance or experience.

**2. Benchmarking and Competitiveness:**

* Compare your company's compensation packages (salary, benefits) to industry standards and competitors.
* Identify areas where you might be falling behind or exceeding expectations.
* Use this information to make data-driven decisions regarding salary adjustments and benefit offerings to attract and retain top talent.

**3. Performance Management and Talent Retention:**

* Analyze compensation data alongside performance reviews to identify correlations between pay and performance.
* This can help identify high performers who might be underpaid and at risk of leaving.
* Use compensation adjustments as a tool to incentivize and reward strong performance.

**4. Workforce Planning and Budgeting:**

* Analyze trends in compensation costs over time to forecast future budgetary needs.
* Identify areas where cost-saving measures might be necessary or strategic investments in talent acquisition are justified.
* Use compensation data to inform workforce planning decisions like promotions, hiring, and restructuring.

**5. Employee Satisfaction and Motivation:**

* Conduct surveys alongside compensation analysis to understand employee sentiment regarding their pay and benefits.
* Identify areas where compensation might be impacting morale or motivation.
* Use compensation data to create a transparent and competitive compensation strategy that fosters employee satisfaction.

**Here are some additional tips for using employee compensation data effectively:**

* **Maintain Data Security and Privacy:** Ensure employee compensation data is secure and used responsibly, complying with all data privacy regulations.
* **Transparency and Communication:** Communicate your compensation philosophy and how data is used to employees to build trust and understanding.
* **Regular Analysis:** Conduct regular reviews of compensation data to stay informed about trends and address any emerging issues.

By leveraging employee compensation data effectively, managers can gain valuable insights to create a fair, competitive, and motivating work environment that attracts and retains top talent.