

**FOUNDATION FOR ORGANIZATION RESEARCH AND EDUCATION (FORE) SCHOOL OF MANAGEMENT**

**NEW DELHI**

**Academic Year 2023 – 25**

**MACHINE LEARNING FOR MANAGERS**

**PROGECT REPORT**

**PROJECT 2**

**Submitted to: Prof. AMARNATH MITRA**

**Submitted by:**

**SUBHAJIT PAUL**

**ROLL: 321173**

**SEC: C**

**Project 2: Report**

**Project Title: Classification of Consumer Data into Segments | Clusters | Classes**

**1. Project Objectives | Problem Statements**

1.1. PO1 | PS1: Classification of Consumer Data into Segments | Clusters | Classes using Supervised Learning Classification Algorithms

1.2. PO2 | PS2: Determination of an Appropriate Classification Model

1.3. PO3 | PS3: Identification of Important | Contributing | Significant Variables or Features and their Thresholds for Classification

**2. Description of Data**

**2.1. Data Source, Size, Shape**

2.1.1. Data Source (Website Link)

-- <https://www.kaggle.com/datasets/parisrohan/credit-score-classification>

2.1.2. Data Size (in KB | MB | GB …)

-- 10 MB

2.1.3. Data Shape | Dimension: Number of Variables | Number of Records

-- Number of Variables: 23, Number of Records: 60000

**2.2. Description of Variables**

2.2.1. Index Variable(s): ID

2.2.2. Variables or Features having Categories | Categorical Variables or Features (CV)

Categorical variables are:

* + - ID
    - Customer\_ID
    - Name
    - Occupation
    - Credit\_Mix
    - Payment\_of\_Min\_Amount
    - Credit\_Score

2.2.2.1. Variables or Features having Nominal Categories | Categorical Variables or Features - Nominal Type:

* Name
* Occupation

2.2.2.2. Variables or Features having Ordinal Categories | Categorical Variables or Features - Ordinal Type:

* Credit\_Mix
* Payment\_of\_Min\_Amount
* Credit\_Score

2.2.3. Non-Categorical Variables or Features:

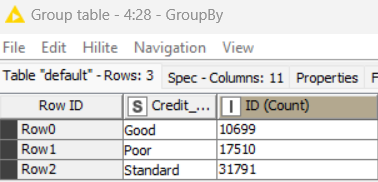
* ID
* Customer\_ID
* Age
* Annual\_Income
* Num\_Bank\_Accounts
* Num\_Credit\_cards
* Interest\_Rate
* Num\_of\_Loan
* Delay\_from\_due\_date
* Num\_of\_Delayed\_Payment
* Changed\_Credit\_Limit
* Num\_Credit\_Inquiries
* Outstanding\_Debt
* Credit\_Utilization\_Ratio
* Credit\_History\_Age\_days
* Total\_EMI\_per\_month
* Monthly\_Balance

**2.3. Descriptive Statistics**

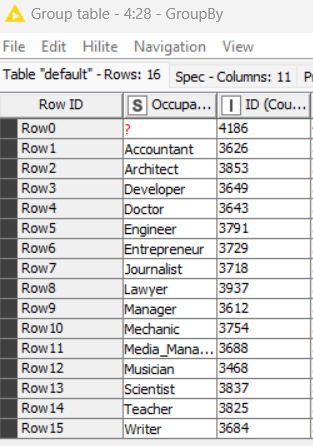
2.3.1. Descriptive Statistics: Categorical Variables or Features

2.3.1.1. Count | Frequency Statistics

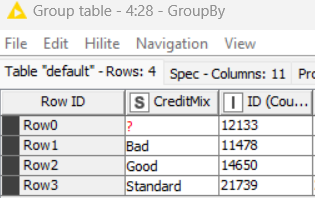
Credit Score:



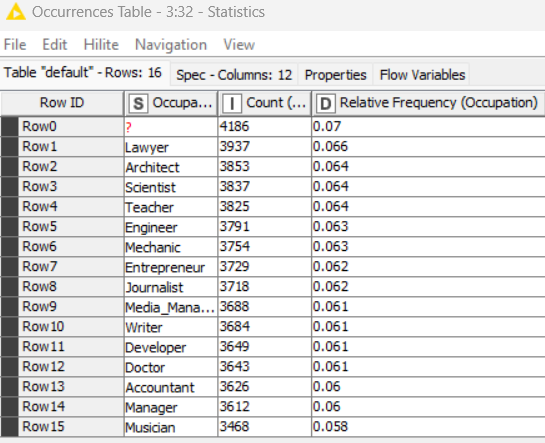
Occupation:

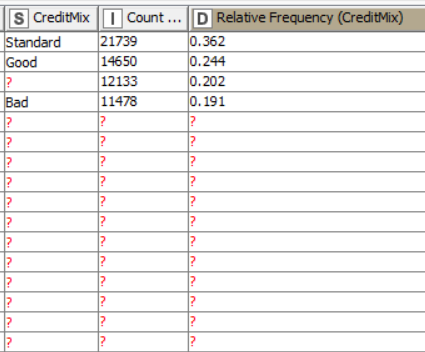


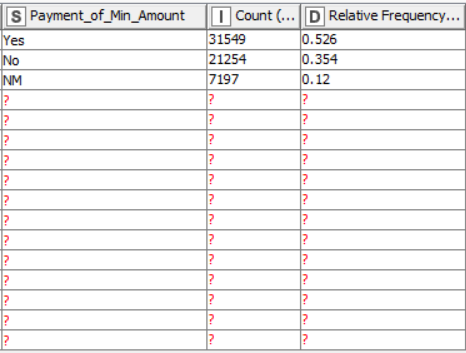
Credit Mix:

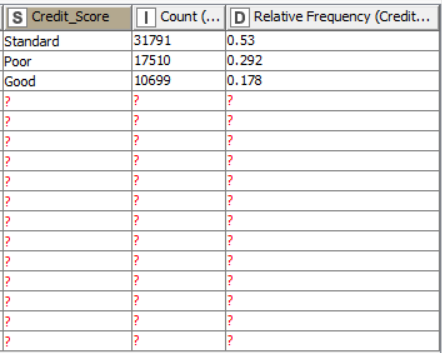


2.3.1.2. Proportion (Relative Frequency) Statistics





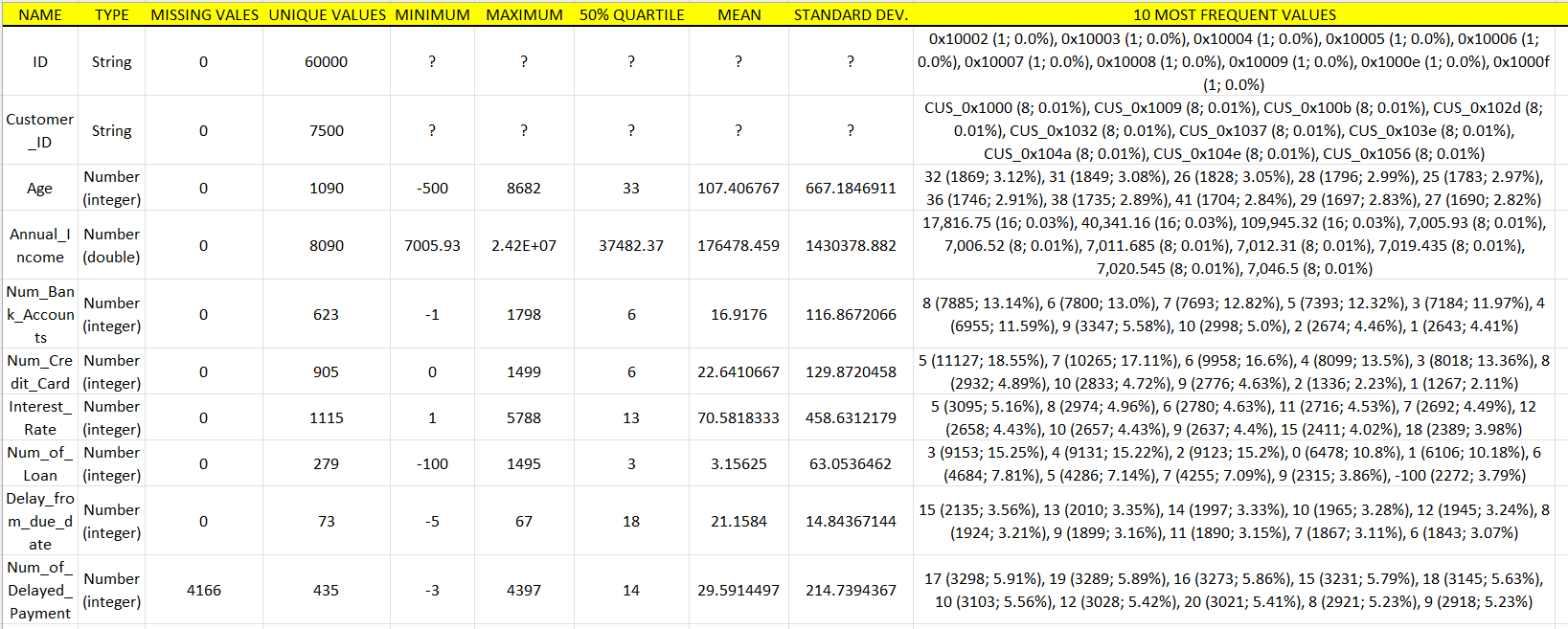


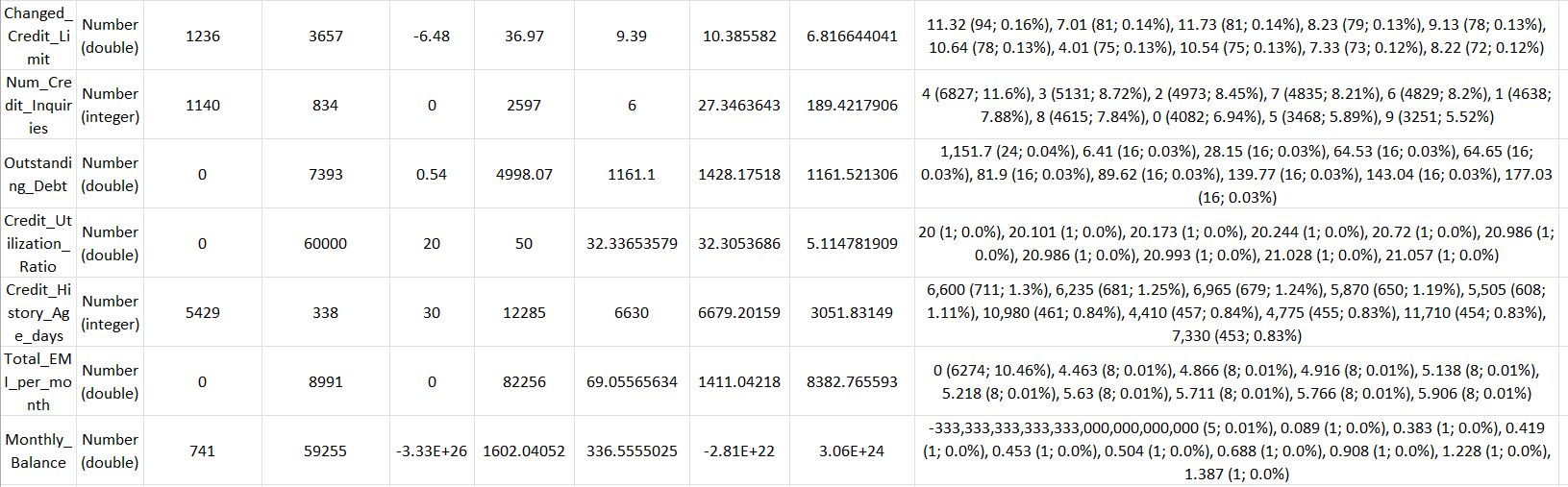


2.3.2. Descriptive Statistics: Non-Categorical Variables or Features

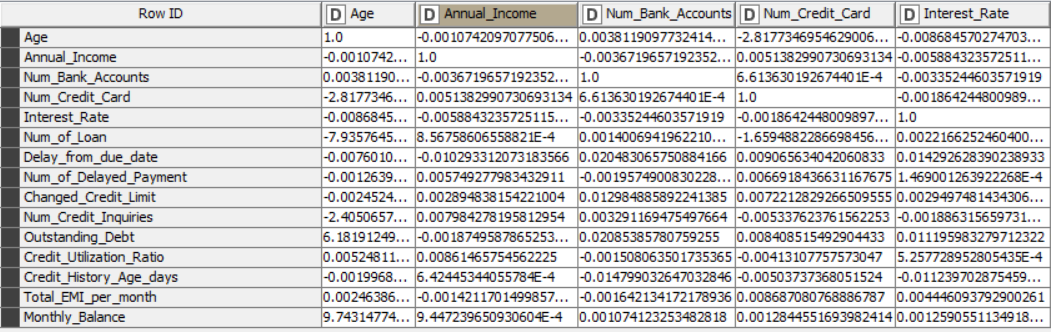
2.3.2.1. Measures of Central Tendency

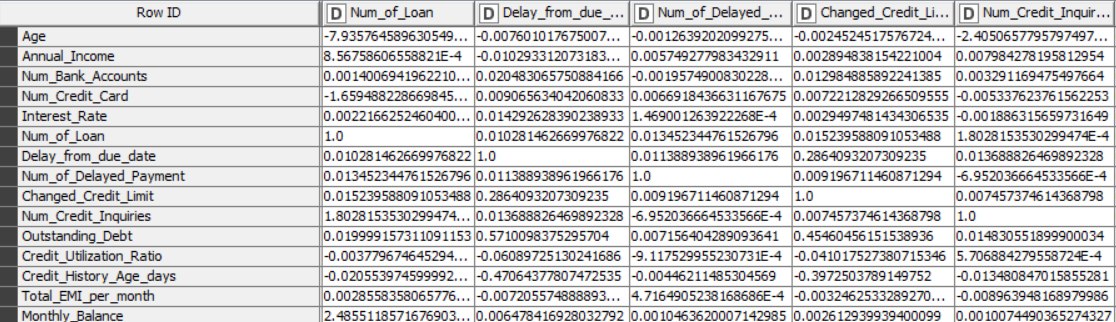
2.3.2.2. Measures of Dispersion

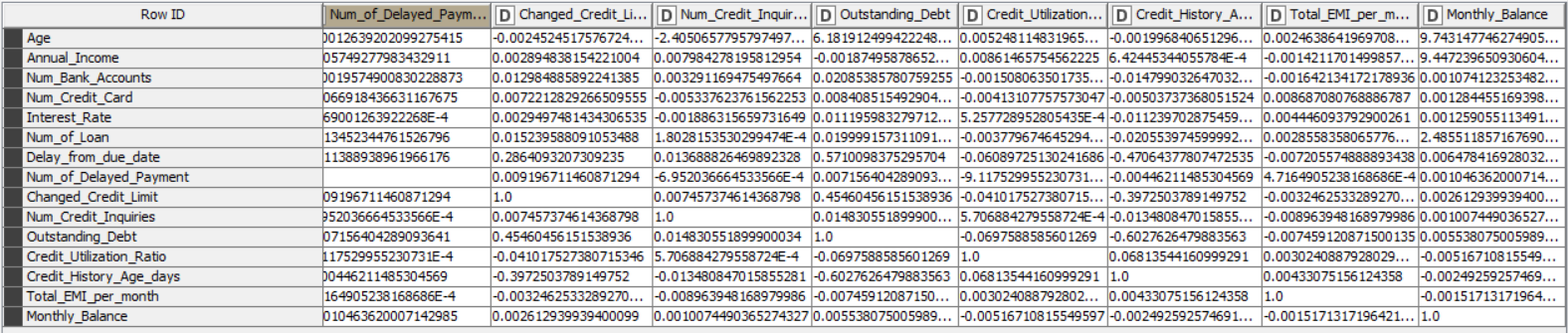




2.3.2.3. Correlation Statistics (with Test of Correlation)







Where the p value is less than 0.05 in the matrix, there the two variables are correlated. For example,

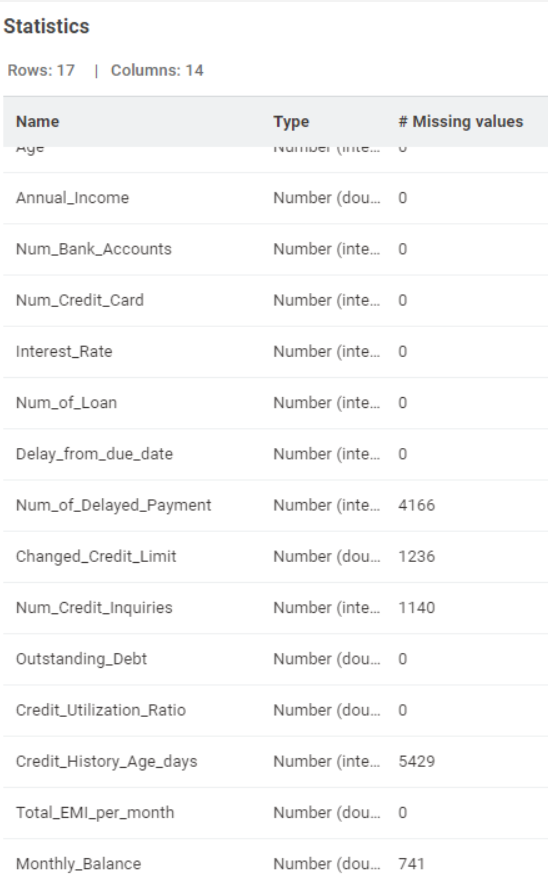
Annual income and Changed Credit Limit are correlated.

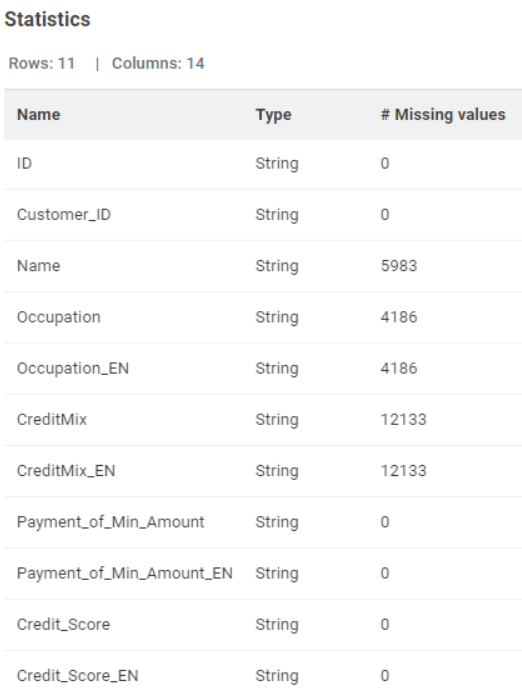
**3. Analysis of Data**

**3.1. Data Pre-Processing**

**3.1.1. Missing Data Statistics and Treatment**

3.1.1.1.1. Missing Data Statistics: Records

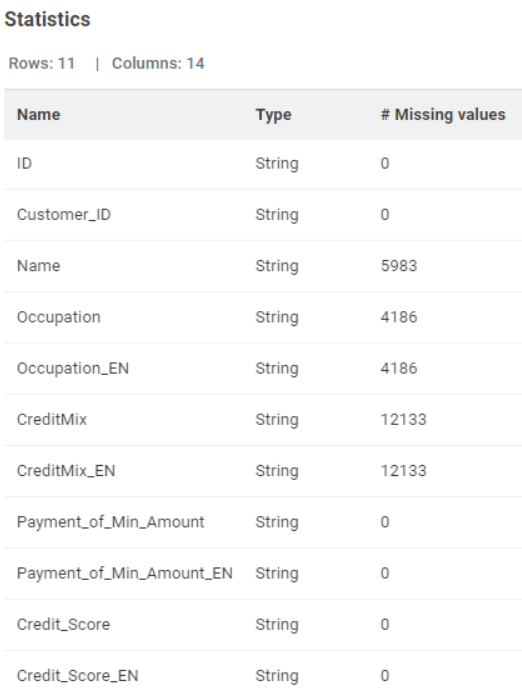




3.1.1.1.2. Missing Data Treatment: Records

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

3.1.1.2.1. Missing Data Statistics: Categorical Variables or Features

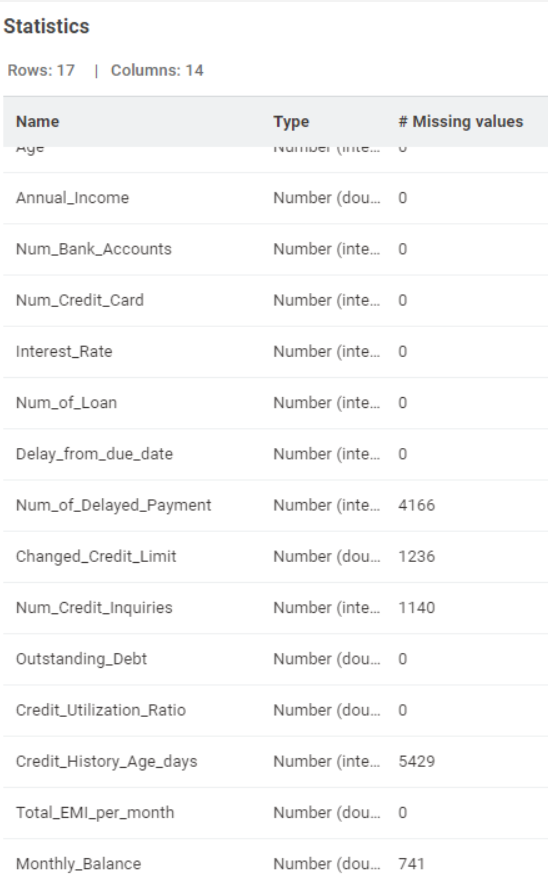


3.1.1.2.2. Missing Data Treatment: Categorical Variables or Features

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

3.1.1.2.2.2. Imputation of Missing Data using Descriptive Statistics: Mode – Most Frequent Value

3.1.1.3.1. Missing Data Statistics: Non-Categorical Variables or Features

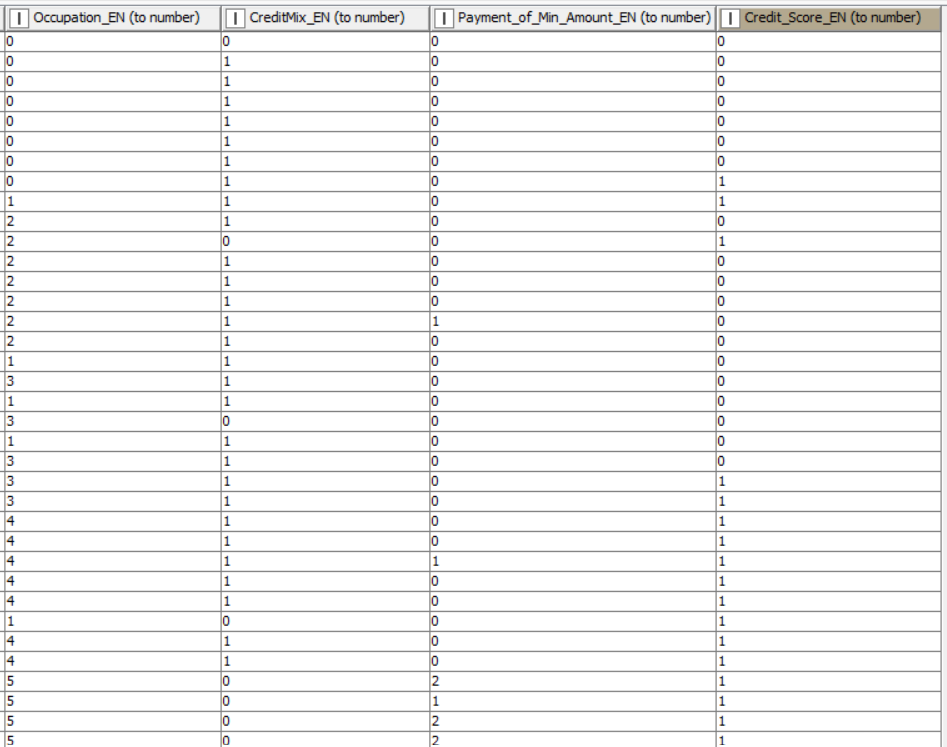


3.1.1.3.2. Missing Data Treatment: Non-Categorical Variables or Features

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

3.1.1.3.2.2. Imputation of Missing Data using Descriptive Statistics: Mean

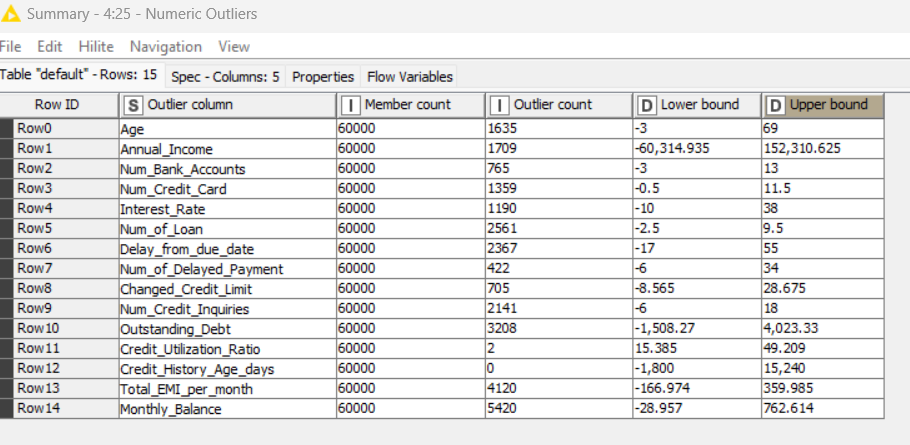
**3.1.2. Numerical Encoding of Categorical Variables or Features** (Encoding Schema - Alphanumeric Order)



Occupation, Credit Mix, Payment of minimum amount, Credit Score has been encoded to numeric characters.

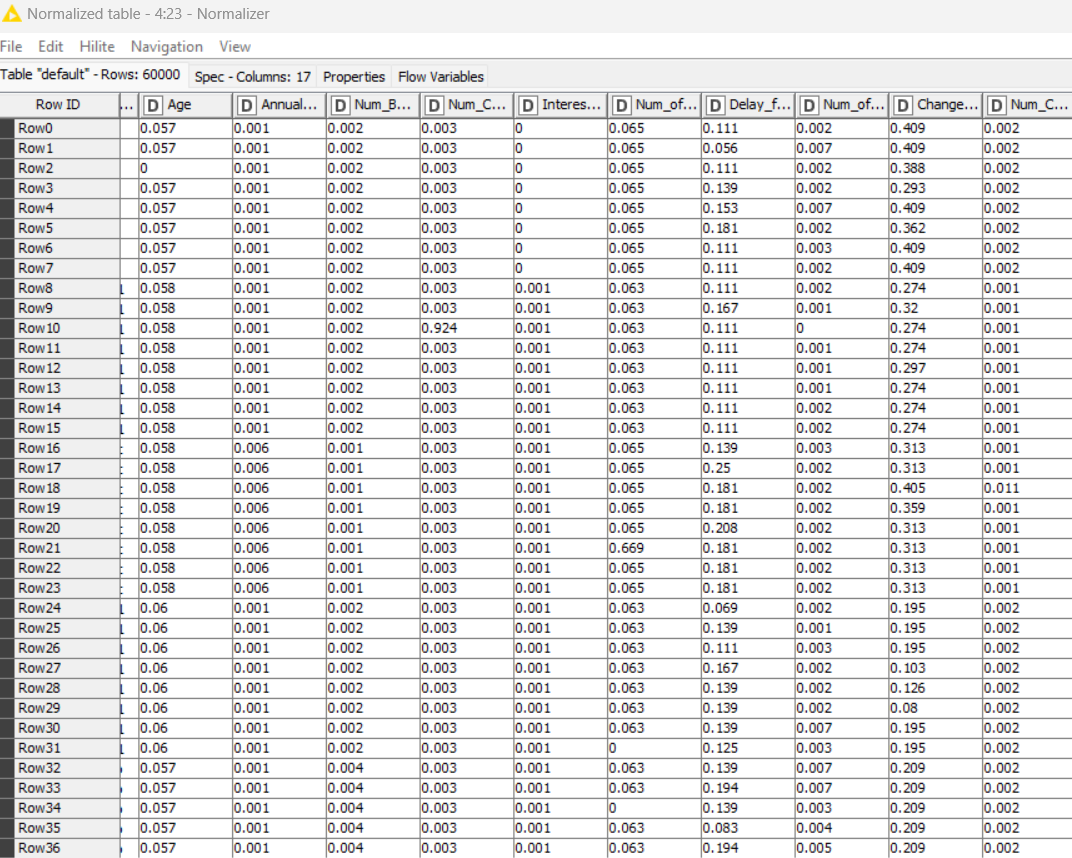
**3.1.3. Outlier Statistics and Treatment** (Scaling | Transformation)

3.1.3.1.1. Outlier Statistics: Non-Categorical Variables or Features



3.1.3.1.2. Outlier Treatment: Non-Categorical Variables or Features

3.1.3.1.2.2. Normalization using Min-Max Scaler:

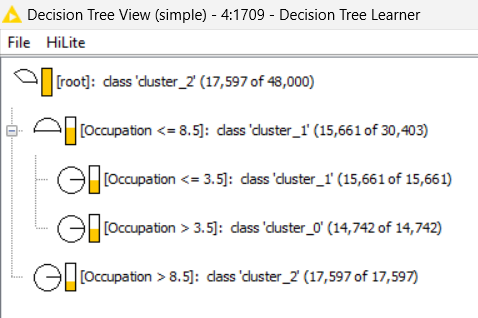


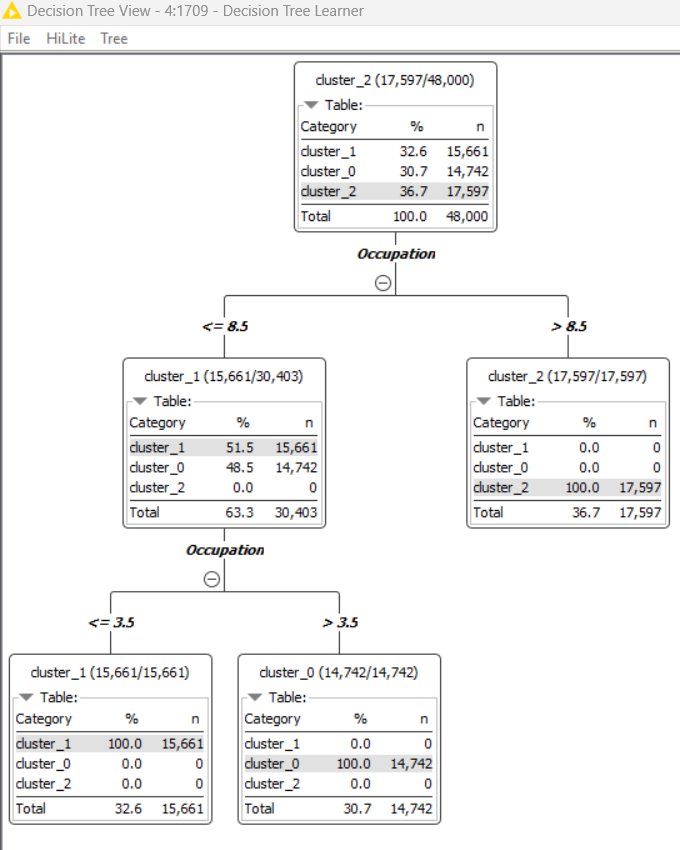
We have done normalization using MIN-MAX scaler where minimum value is 0 and maximum value is 1.

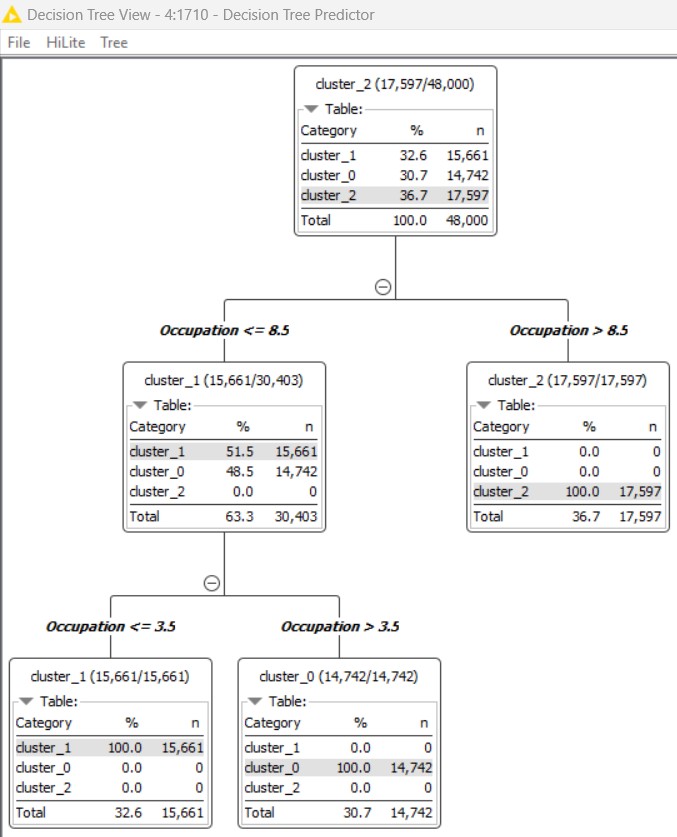
**3.1.4. Data Bifurcation: Training & Testing Sets** [Bifurcation Schema: Random Sampling or Stratified Sampling (Based on Outcome Variable or Feature) with 80% Data in Training Set and 20% Data in Testing Set]

**3.2. Data Analysis**

3.2.1.1. PO1 | PS1:: Supervised Machine Learning Classification Algorithm: Decision Tree (Base Model) | Metrics Used - Gini Coefficient, Entropy

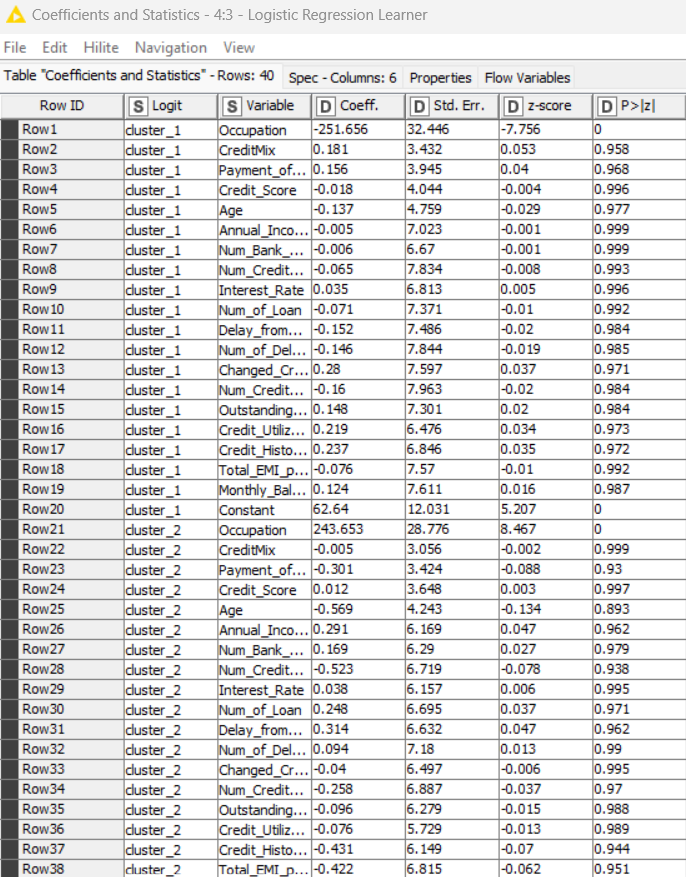






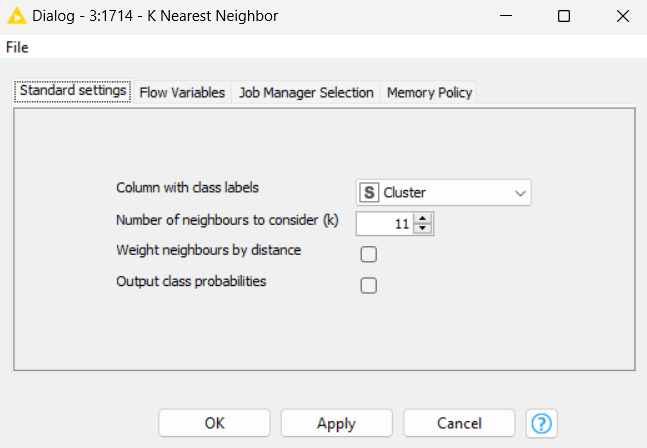
3.2.1.2. PO1 | PS1:: Supervised Machine Learning Classification Algorithms: {Logistic Regression | Support Vector Machine | K Nearest Neighbour} (Comparison Models) | Metrics Used

Logistic Regression:

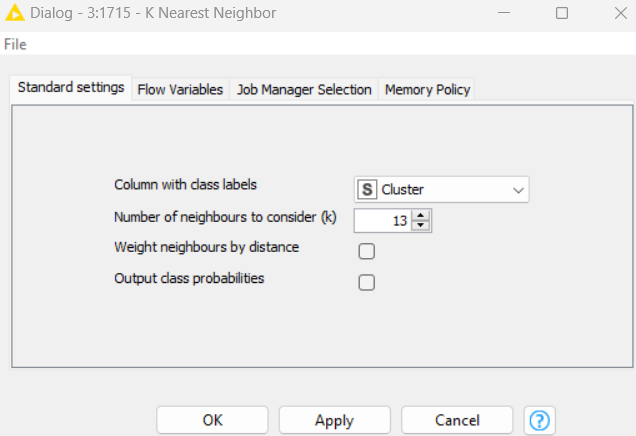


KNN:

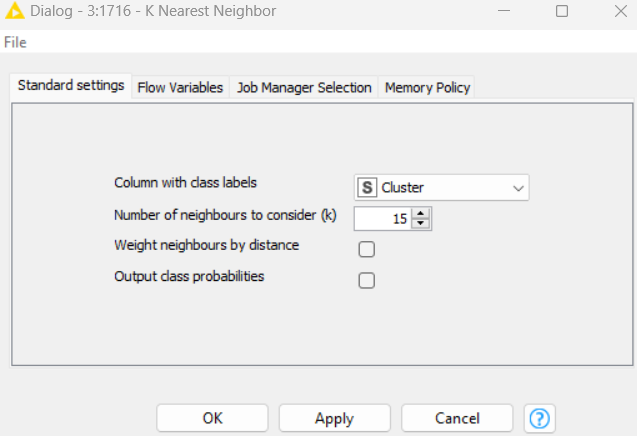
For k=11:



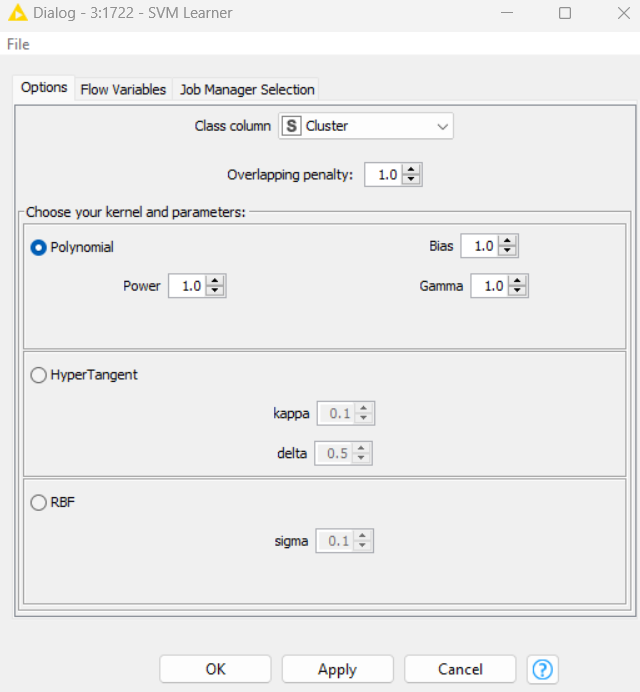
For k=13:



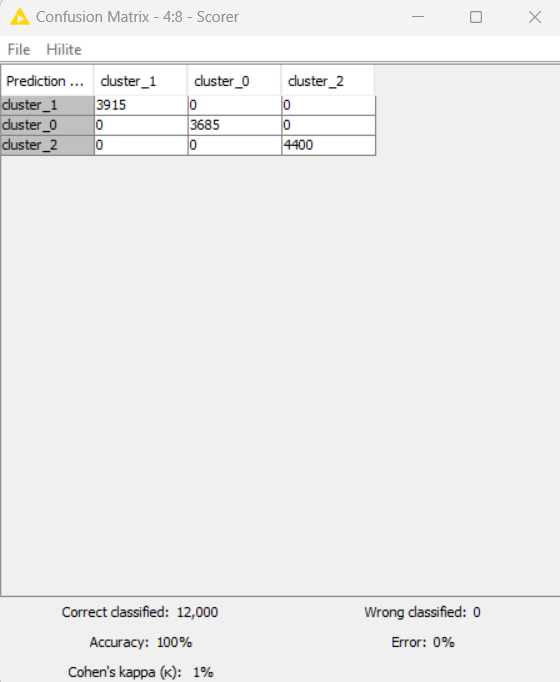
For K=15:



SVM:

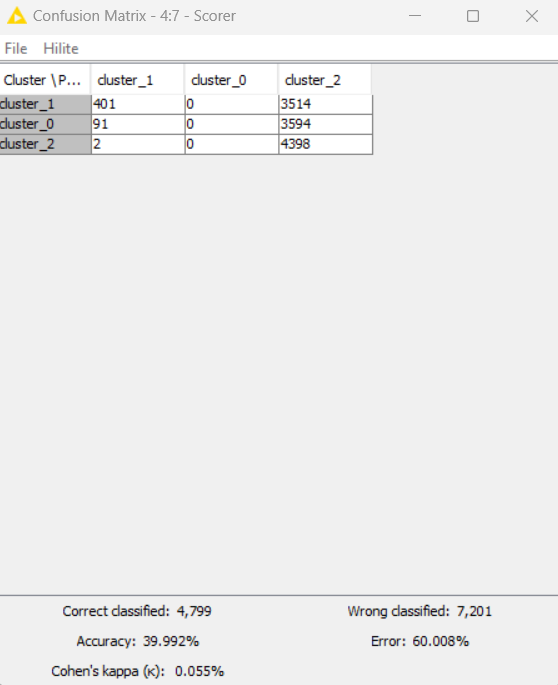


3.2.2.1.1. PO2 | PS2:: Classification Model Performance Evaluation: Confusion Matrix {Accuray, Recall, Precision, F1-Score} (Base Model: Decision Tree)

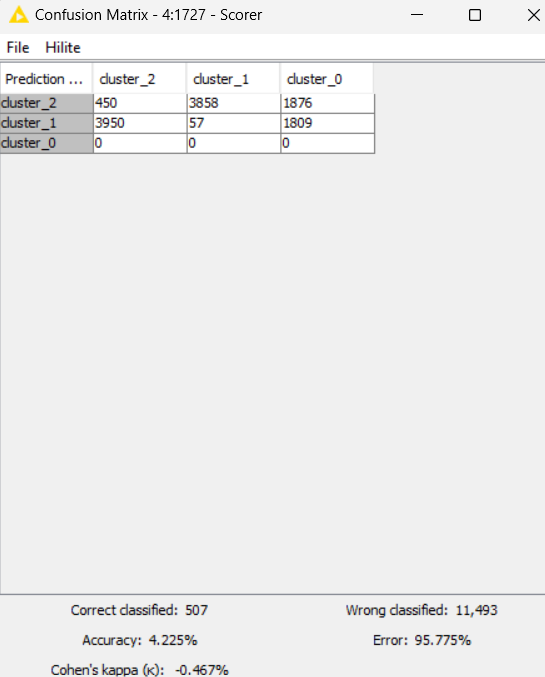


3.2.2.2.1. PO2 | PS2:: Classification Model Performance Evaluation: Confusion Matrix {Accuray, Recall, Precision, F1-Score} (Comparison Models: Logistic Regression | Support Vector Machine | K Nearest Neighbour)

Logistic Regression:

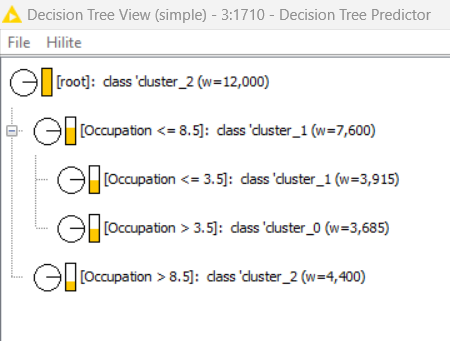


SVM:



3.2.3.1. PO3 | PS3:: Variable or Feature Analysis: Base Model (Decision Tree)

3.2.3.1.1. List of Relevant or Important Variables or Features and their Thresholds



Occupation is a relevant variable. The cluster is dependent on the customer’s occupation

3.2.3.1.2. List of Non-Relevant or Non-Important Variables or Features

3.2.3.2. PO3 | PS3:: Variable or Feature Analysis: Comparison Models (Logistic Regression | Support Vector Machine | K Nearest Neighbour)

3.2.3.2.1. List of Relevant or Important Variables or Features and their Thresholds

3.2.3.2.2. List of Non-Relevant or Non-Important Variables or Features

**4. Results | Observations**

4.1. Classification Model Parameters: Base Model (Decision Tree) | Comparison Models (Logistic Regression | Support Vector Machine | K Nearest Neighbour)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Metric | Decision  Tree | Logistic  Regression | KNN (k=11) | KNN (k=13) | KNN (k=15) | SVM |
| Accuracy | 100% | 39.99% | 92.09% | 92.05% | 92.1% | 4.23% |
| Error | 0% | 60.01% | 7.91% | 7.95% | 7.9% | 95.77% |
| Cohen’s Kappa | 1% | 0.055% | 0.881% | 0.880% | 0.881% | -0.47% |
| Correctly Classified | 12000 | 4799 | 11051 | 11046 | 11052 | 507 |
| Wrongly Classified | 0 | 7201 | 949 | 954 | 948 | 11493 |

Based on the provided metrics, the decision tree model clearly outperforms the other models in terms of accuracy, error rate, and agreement beyond chance (Cohen's Kappa). Therefore, the recommendation would be to use the decision tree model for the classification task, as it demonstrates superior performance. However, it's also essential to consider factors such as model complexity, interpretability, and computational resources required for deployment when making the final decision.

**5. Managerial Insights**

5.1. Appropriate Model: Compare and Contrast {Decision Tree | Logistic Regression | Support Vector Machine | K Nearest Neighbour}

Decision Tree

Occupation can indirectly influence credit score classification through various factors such as income level, stability of employment, and overall financial behavior associated with specific professions. Here's how different occupations can contribute to credit score clustering:

1. Income Level: Certain occupations tend to offer higher salaries than others. Individuals with higher incomes may find it easier to manage debt and make timely payments, potentially leading to higher credit scores. Occupations such as doctors, lawyers, engineers, and executives often fall into this category.

2. Stability of Employment: Some occupations are more stable than others in terms of job security and consistent income. Stable employment provides lenders with confidence in the borrower's ability to repay loans, positively impacting credit scores. Jobs in sectors like healthcare, government, and education are typically considered more stable.

3. Debt-to-Income Ratio: Occupations with irregular income or freelance work may face challenges in maintaining a favorable debt-to-income ratio, which can affect credit scores. Industries like entertainment, arts, and hospitality often have fluctuating incomes, potentially leading to higher debt burdens.

4. Access to Credit: Certain professions may provide employees with perks such as corporate credit cards or loans with preferential terms, which can impact credit utilization and overall credit health. Occupations in finance, for example, may have easier access to credit-building tools compared to other sectors.

5. Financial Responsibility: The nature of one's occupation may also reflect on their financial responsibility. Individuals in professions requiring strong attention to detail, financial planning, and risk management may exhibit better financial behaviors, leading to higher credit scores. This could include occupations in accounting, financial advising, or banking.

6. Credit History: Over time, individuals in specific occupations may develop industry-specific credit patterns. For instance, individuals in real estate may have higher mortgage debts but also potentially higher property assets, influencing their credit profiles differently from individuals in other sectors.

7. Geographic Location: Certain occupations are more prevalent in specific geographic regions, where cost of living and economic conditions can vary. This can impact credit score classification due to differences in expenses, income levels, and overall financial circumstances.

It's important to note that while occupation can play a role in credit score classification, individual financial behavior and circumstances remain the primary factors influencing creditworthiness. Lenders assess creditworthiness based on a combination of factors including payment history, credit utilization, length of credit history, new credit accounts, and credit mix, in addition to income and employment stability.