

**Academic Year 2023-2025**

**Classification of Consumer Data**

**into Segments (Loan Data)**

**Machine Learning for Managers-2**

**Submitted to:**

**Prof. Amarnath Mitra**

**Submitted by:**

**321066 - Akriti Agarwal**

***Contents of the Report***

|  |  |
| --- | --- |
|  | Objective of the Project |
| 2. | Description of Data |
| 3. | Analysis of Data |
| 4. | Observations |
| 5. | Managerial Insights |

1. **Objectives of the Project**

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

1.2. Determination of an Appropriate Classification Model

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

1. **Description of Data**

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

2.1.1. Data Source - <https://www.kaggle.com/datasets/rounak02/financial-data>

2.1.2. Data Size- 57.3 MB

2.1.3. Data Shape

Dimensions: Number of Variables- 43

Number of Records- 212999

* 1. **Description of Variables**

**2.2.1. Index Variable(s**): 2- Id, Member Id

**2.2.2. Outcome Variable**- Cluster

**2.2.3. Input Variables having Categories** | Input Categorical Variables (ICV)

**2.2.3.1. Input Variables having Nominal Categories**

* Term
* Issue\_d
* Last\_pymnt\_d
* Last\_credit\_pull
* Home Ownership
* Loan status
* Initial list
* Application type

**2.2.3.2. Input Variables having Ordinal Categories**

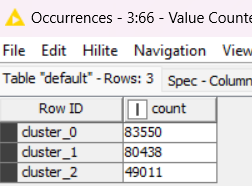
* Grade
* sub grade
* inq\_last\_6mths
* verification

**2.2.3. Input Non-Categorical Variables**

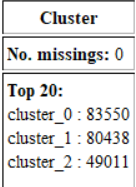
* Loan Amount
* Annual Income
* Delinq\_2yrs
* Emp\_length
* Purpose
* dti
* Installment
* Earliest\_cr\_line
* Interest Rate
* Mnths\_since\_last\_delinq
* Open\_acc
* Pub\_rec
* Employee title
* Revol\_bal
* Revol\_util
* total acc
* total payment
* total recovery principal
* total recovery interest
* late fees
* recoveries
* collection
* emp\_title
* last\_pymnt\_amnt
* mnths\_since\_last\_major\_derog
* total\_coll\_amt
* total\_rev\_hi\_lim
* tot\_cur\_bal
  1. **Descriptive Statistics**

**2.3.1. Descriptive Statistics: Outcome Variable (Categorical)**

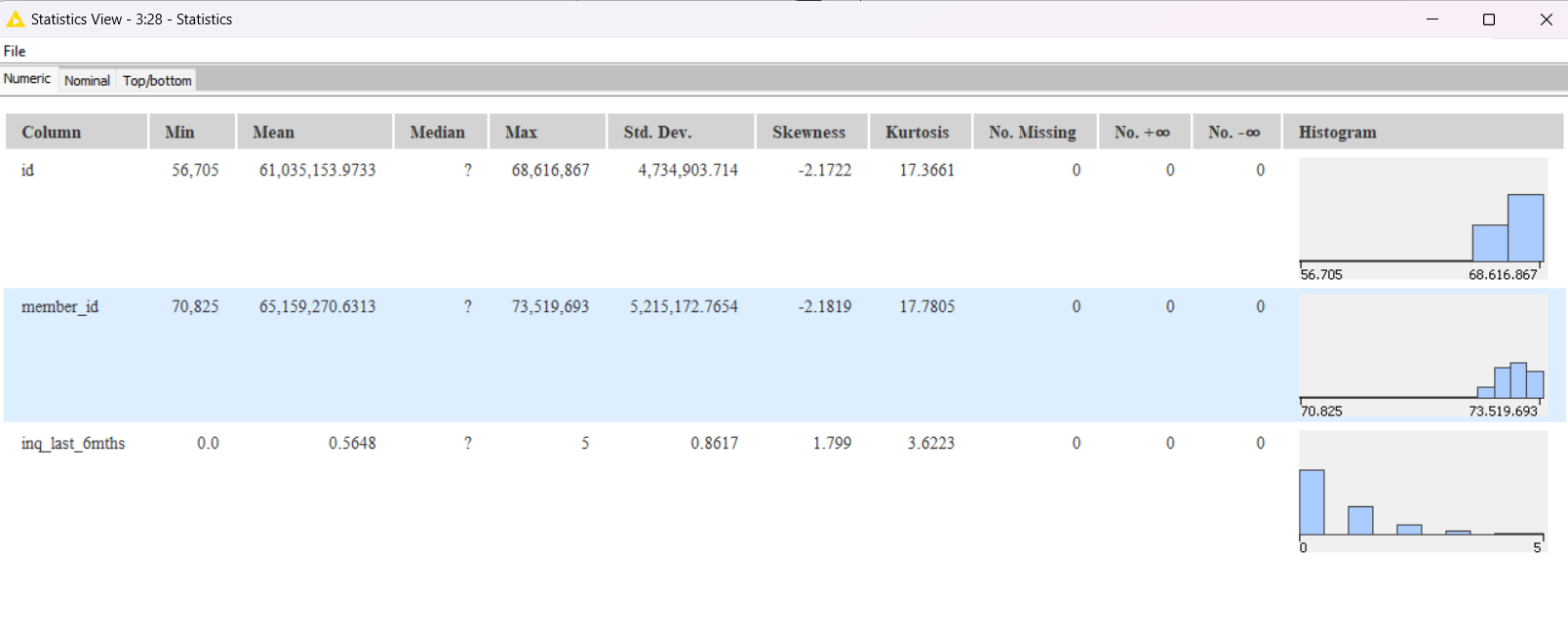
**2.3.1.1. Count | Frequency Statistics**

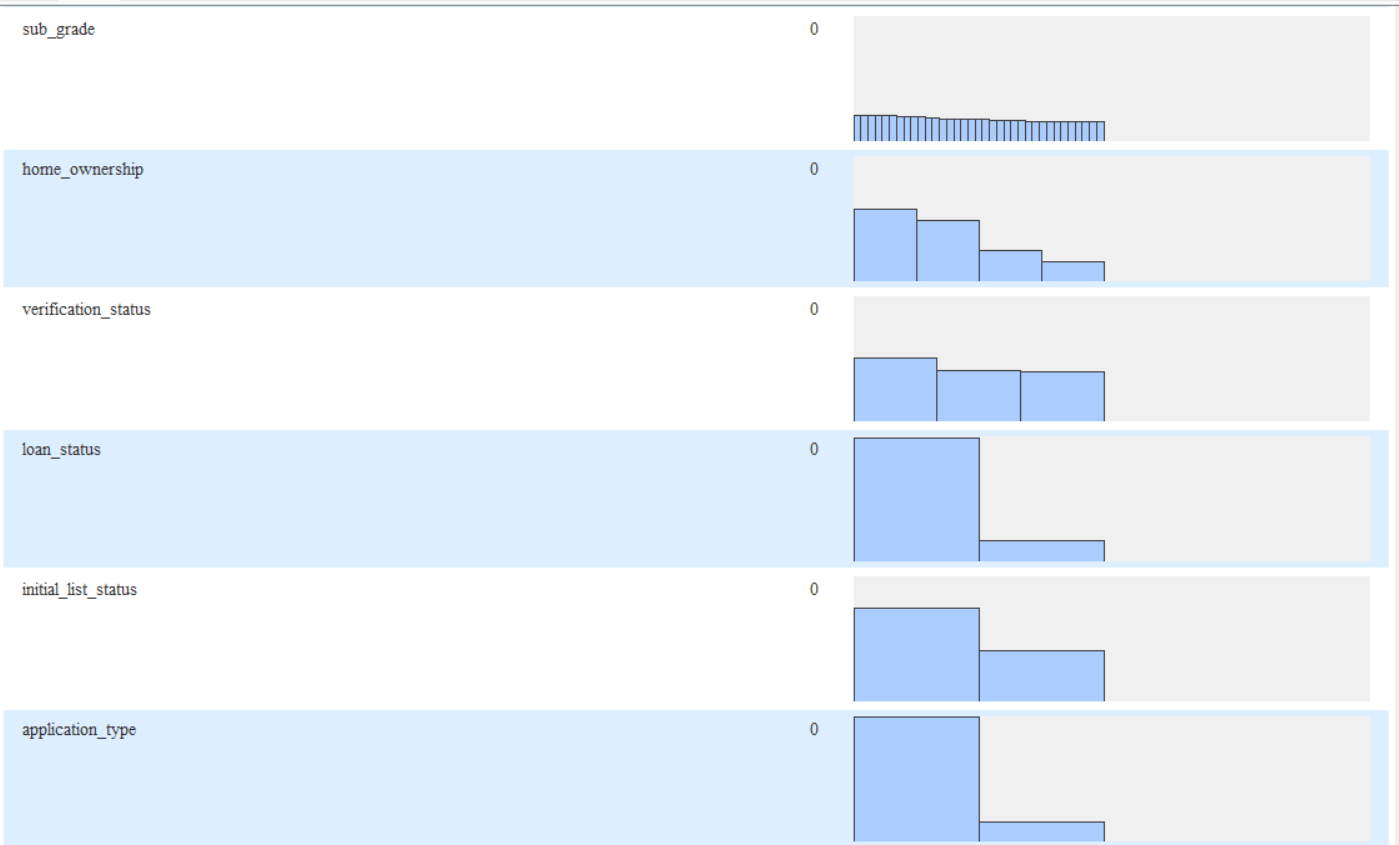
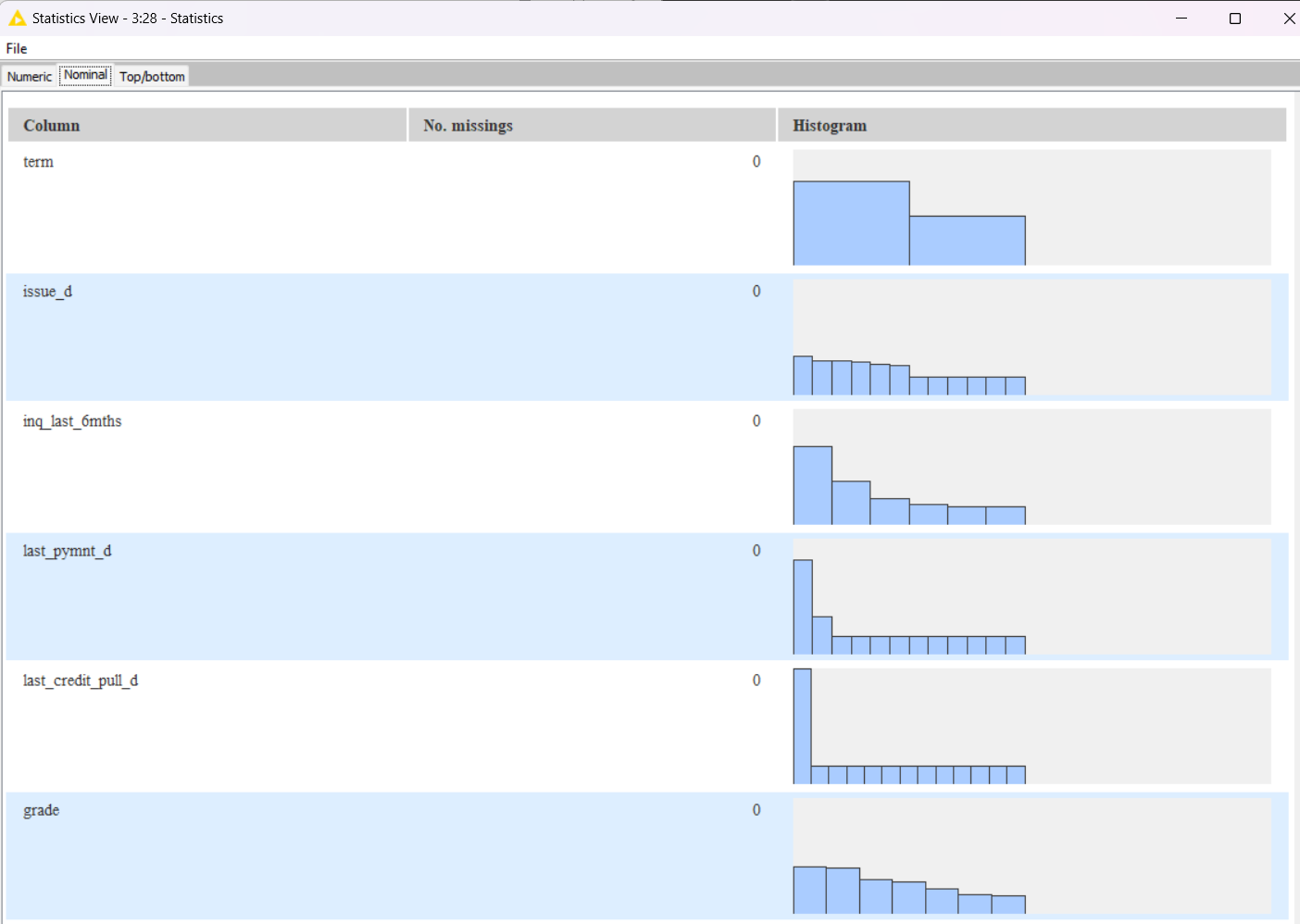


**2.3.1.2. Proportion (Relative Frequency) Statistics**



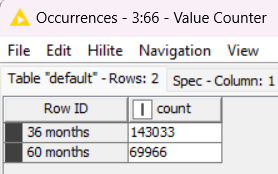
**2.3.2. Descriptive Statistics: Input Categorical Variables**



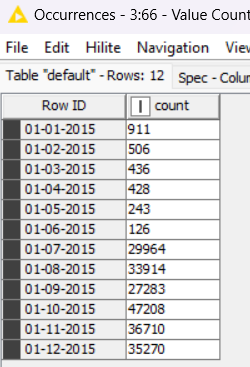


**2.3.2.1. Count | Frequency Statistics**

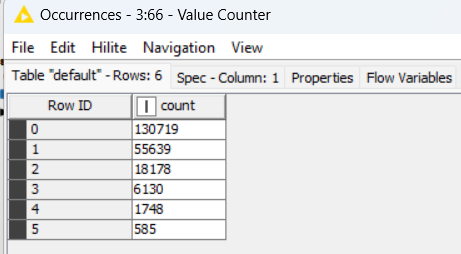
Term



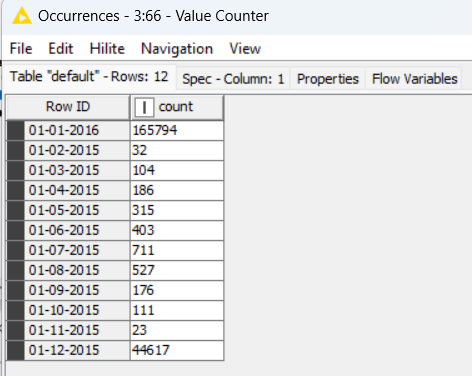
Issue\_d



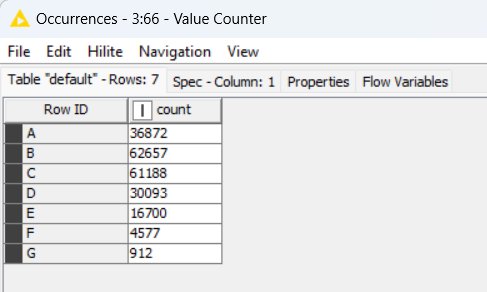
Inq\_last\_6mths



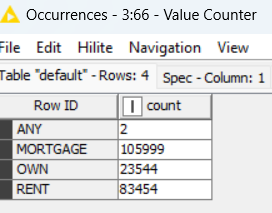
Last\_pymnt\_d



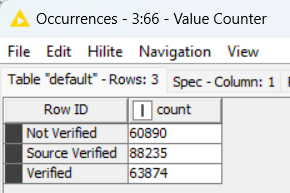
Grade



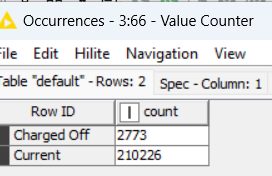
Home ownership



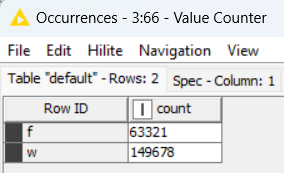
Verification



Loan status



Initial List



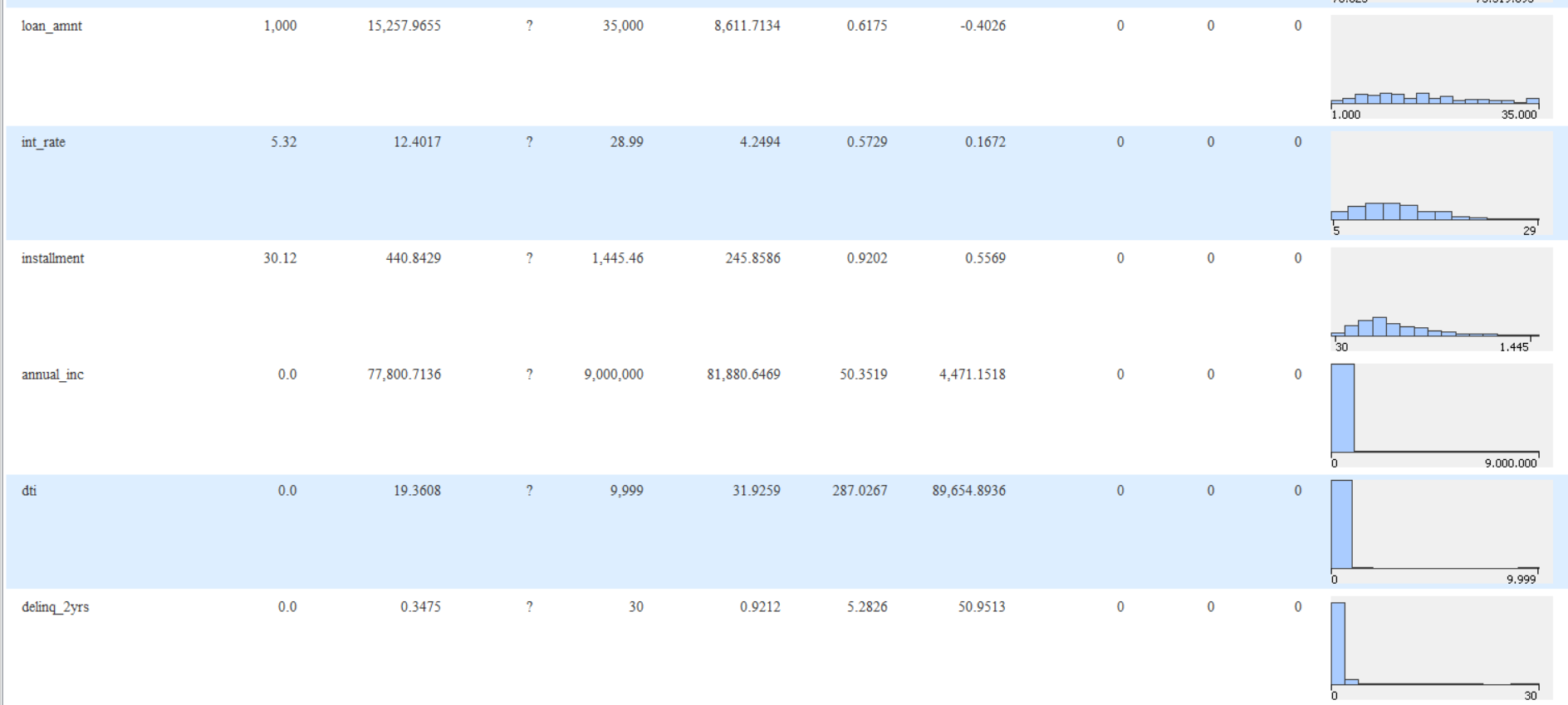
Application type

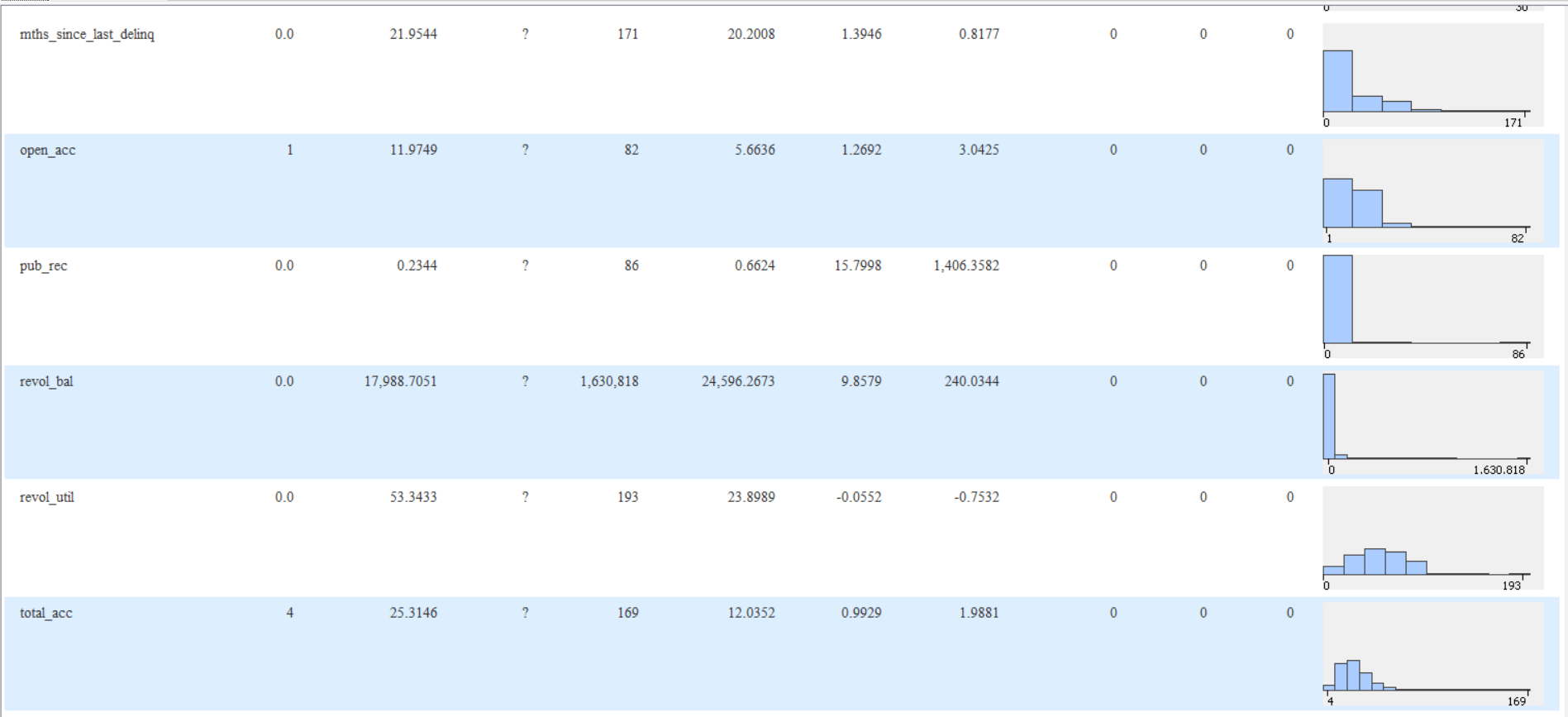


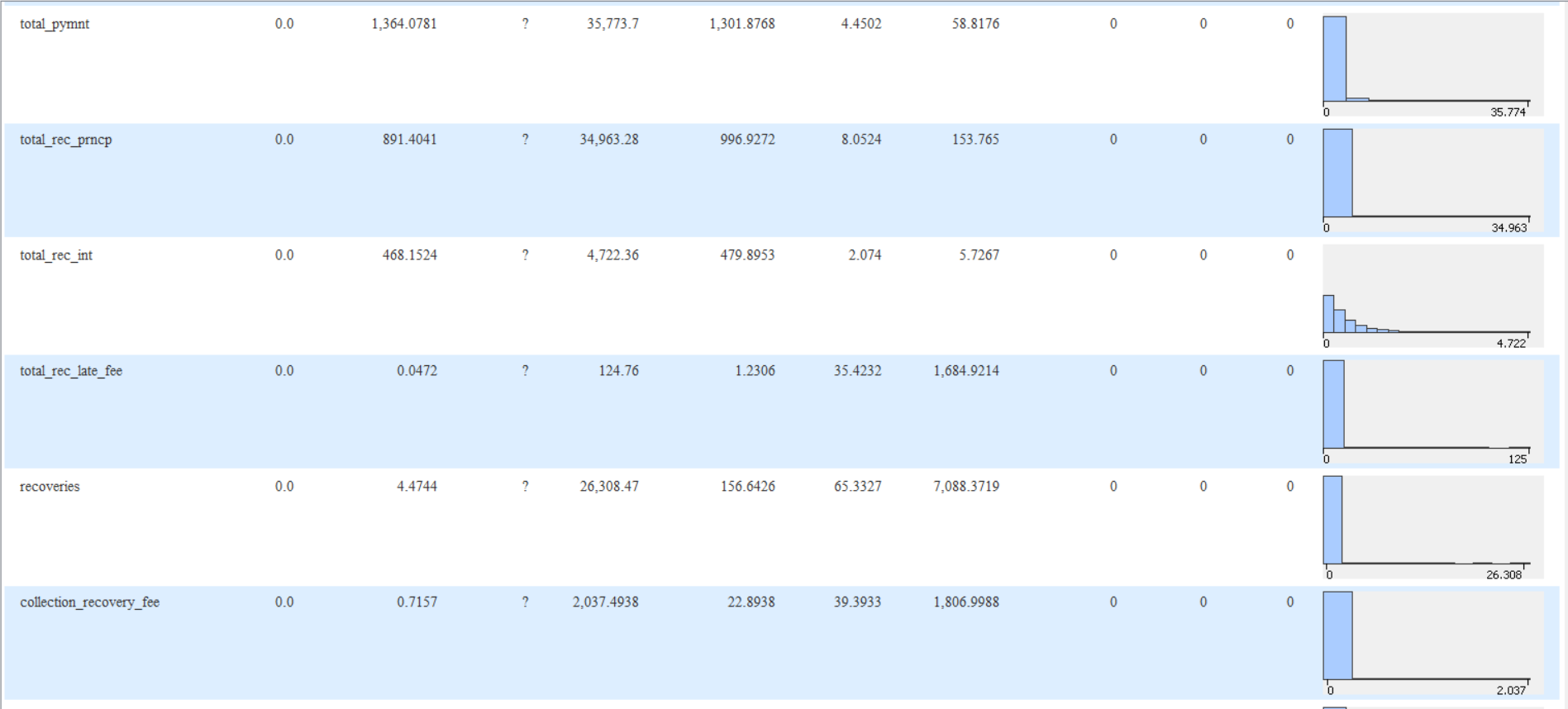
**2.3.3. Descriptive Statistics: Input Non-Categorical Variables**

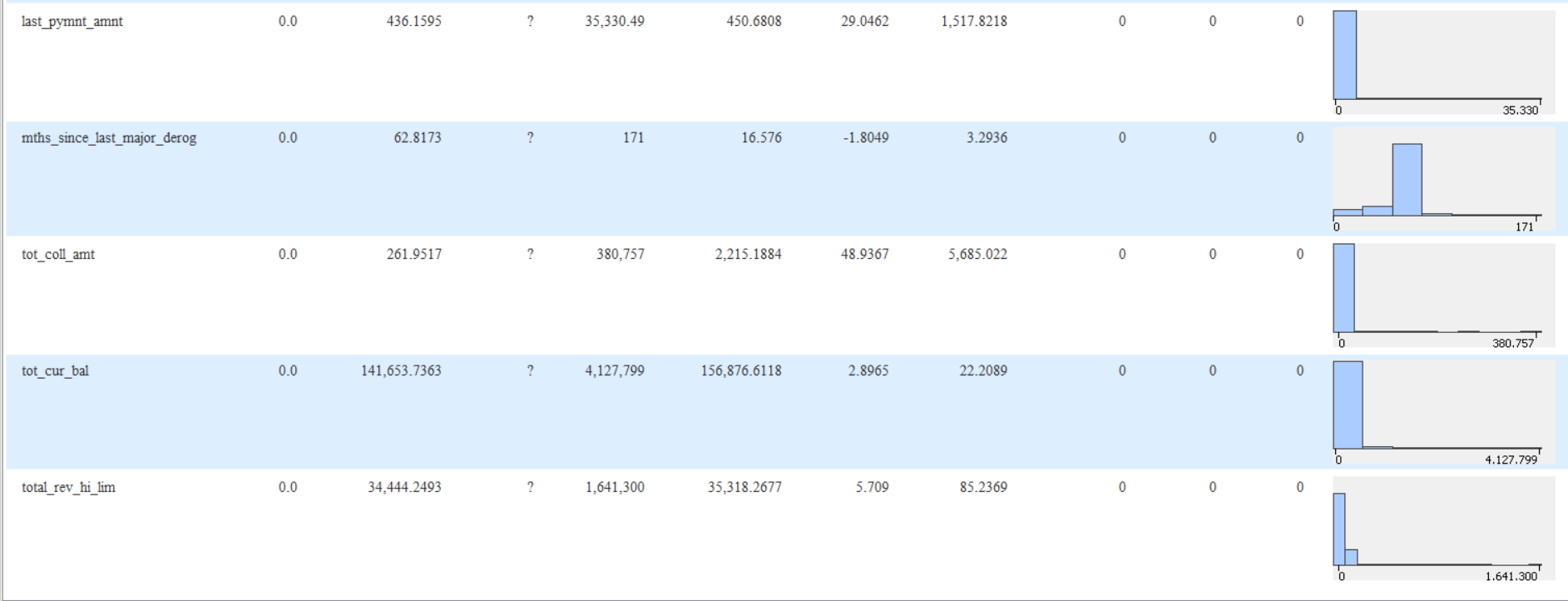
2.3.3.1. Measures of Central Tendency

2.3.3.2. Measures of Dispersion

****

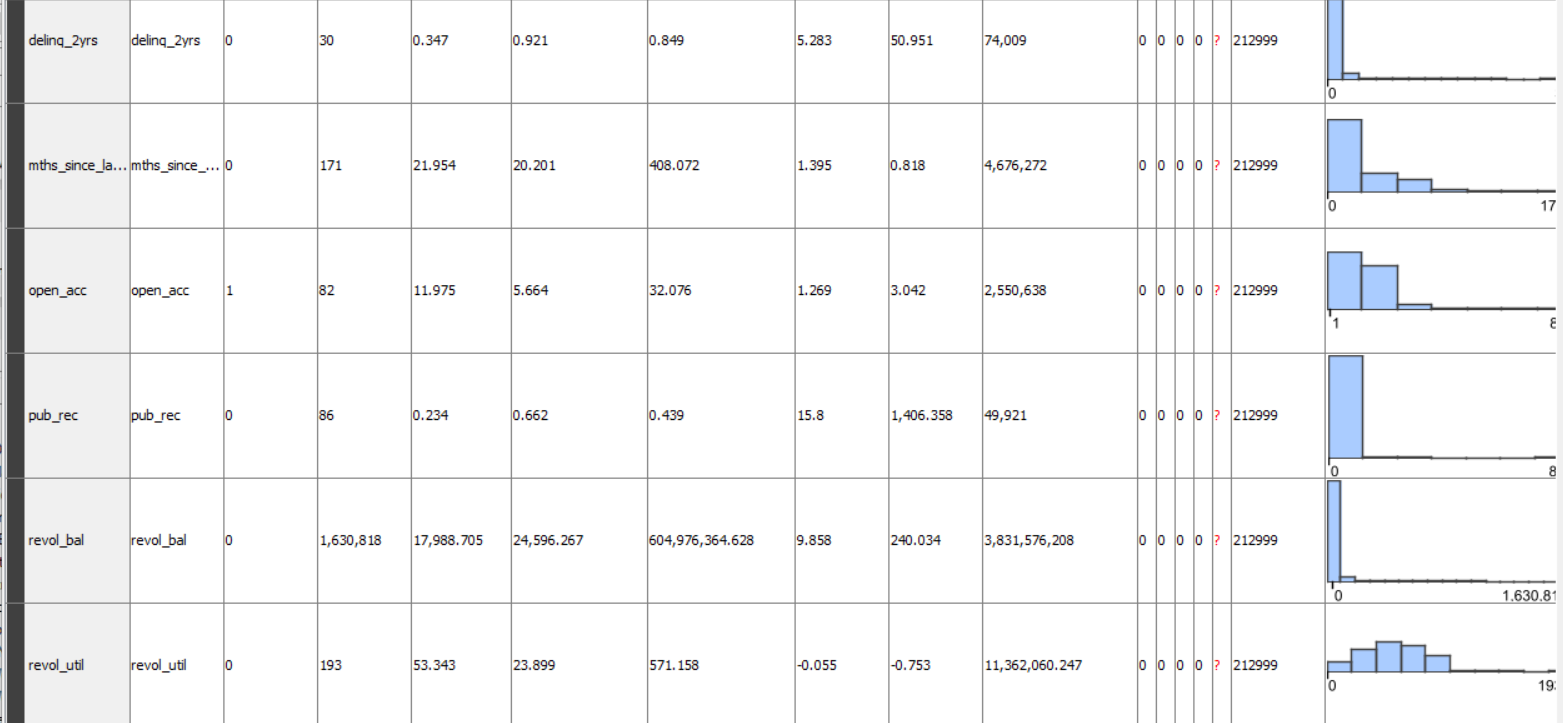
****

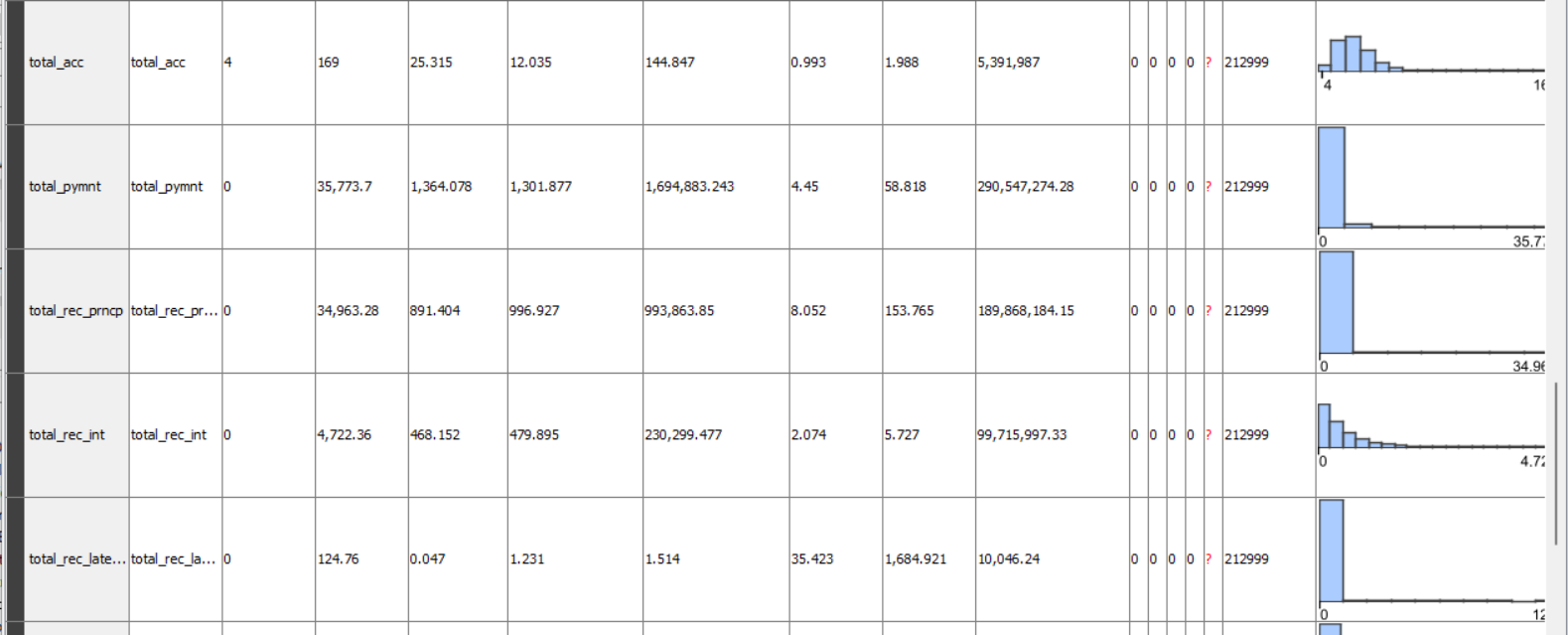




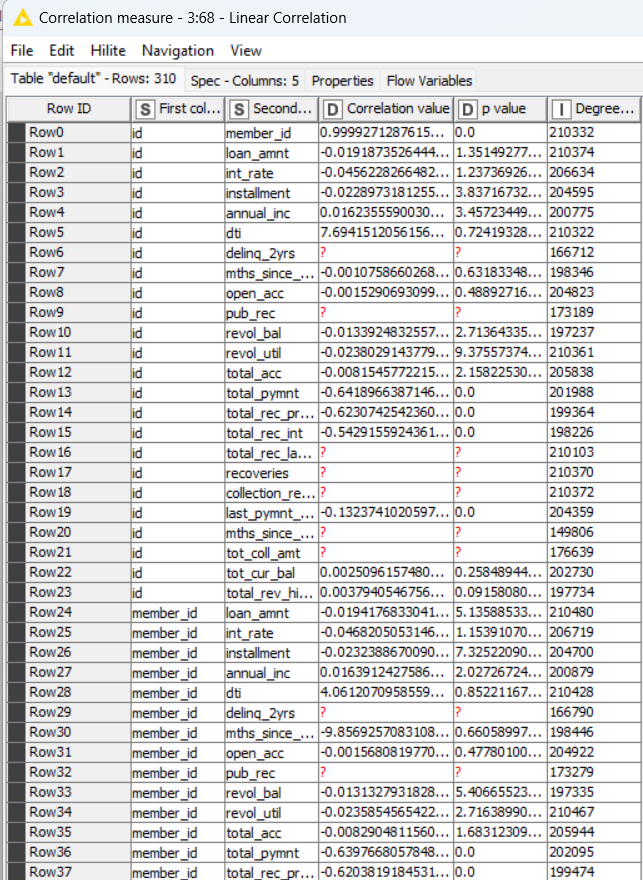




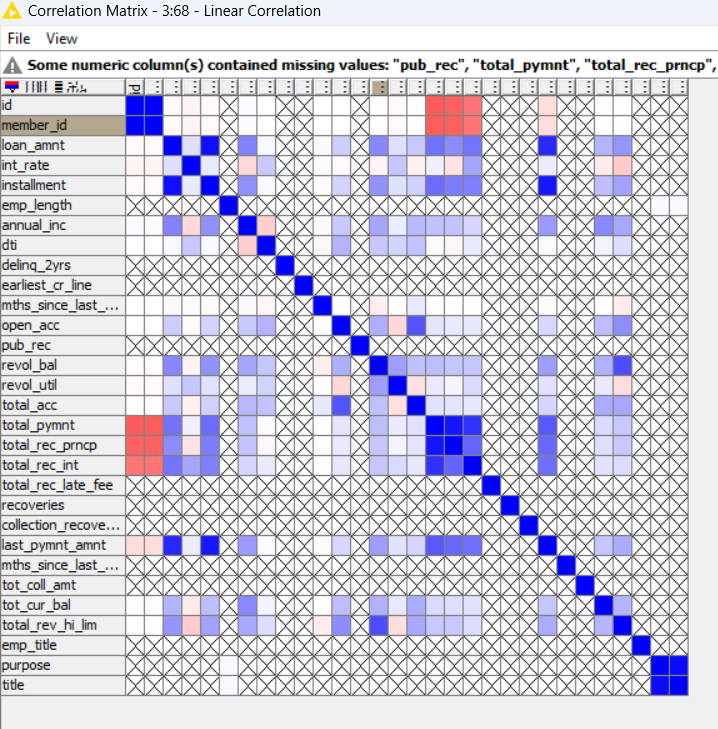


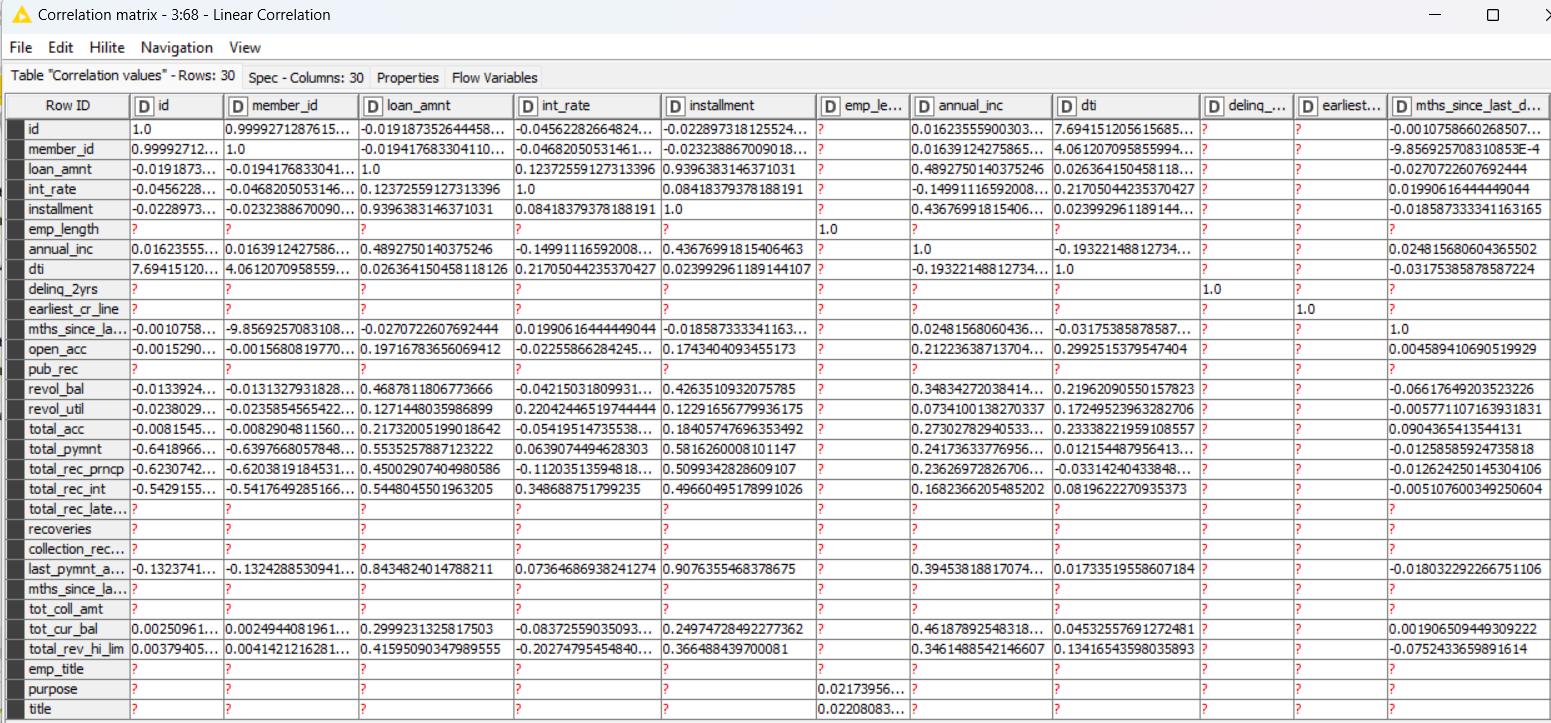


**2.3.3.3. Correlation Statistics (with Test of Correlation)**



The variables are correlated if the value of p is less than 0.05. The variables that are not correlated are emp\_lenght\_int and recoveries, interest\_rate and total\_rec\_prncp, and dti and recoveries because the p-value is less than 0.05.





1. **Analysis**
   1. **Data Preprocessing**
      1. **Missing Data Treatment**
         1. *Stats: Categorical missing data variables*-

*Non-categorical missing data variables*-

* + - 1. *Proposed Treatment: Categorical Data*- Mode

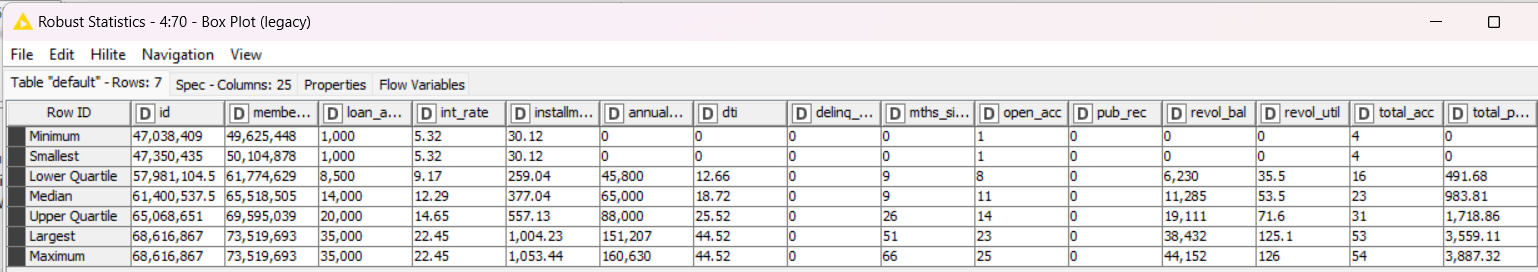
*Non-Categorical*- Mean

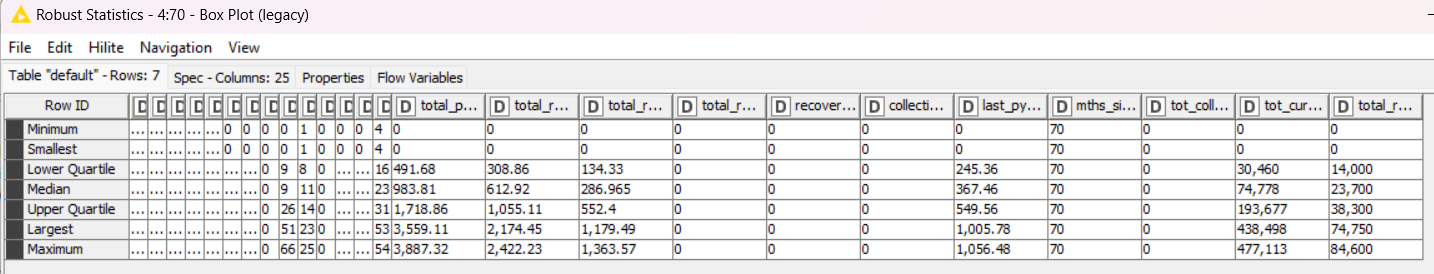
3.1.1.3 *Variables excluded due to more than 50% missing data*- None

* + 1. **Categorical Data**- Numerical Encoding

Encoding Schema- Alphanumeric

* + 1. **Non-Categorical Data**- Outlier Statistics and Treatment





Using Box Plot



**Normalisation for outliers using min/max scaler**



**3.1.4. Data Bifurcation**: Training & Testing Sets [Bifurcation Schema: Stratified Sampling (Based on Clusters) with 80% Data in Training Set and 20% Data in Testing Set

* 1. **Data Analysis**

**3.2.1.1. PO1 | Supervised Machine Learning Clustering Algorithm: Decision Tree (Base Model)| Metrics Used**

➔ 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

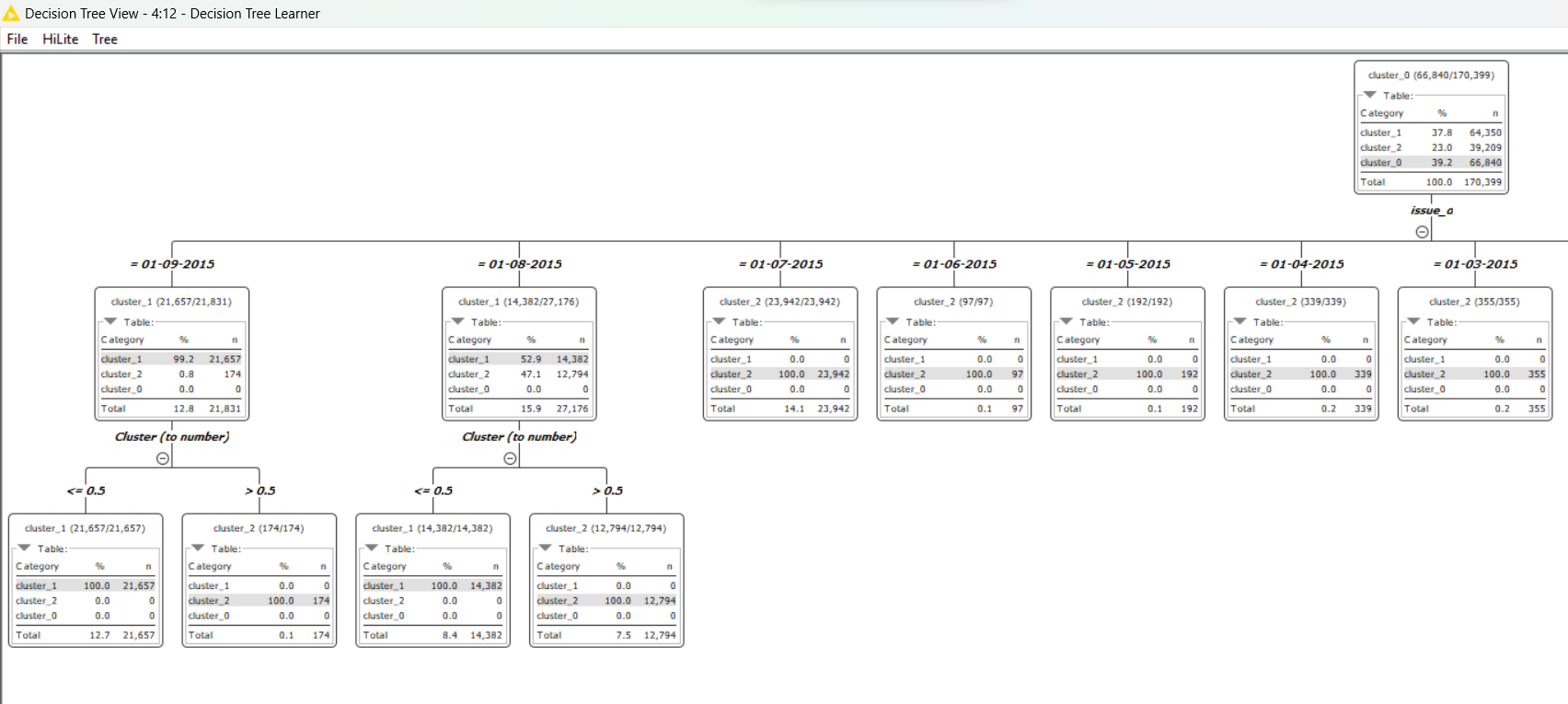
a specific feature, and the data is split into subsets. This process continues until

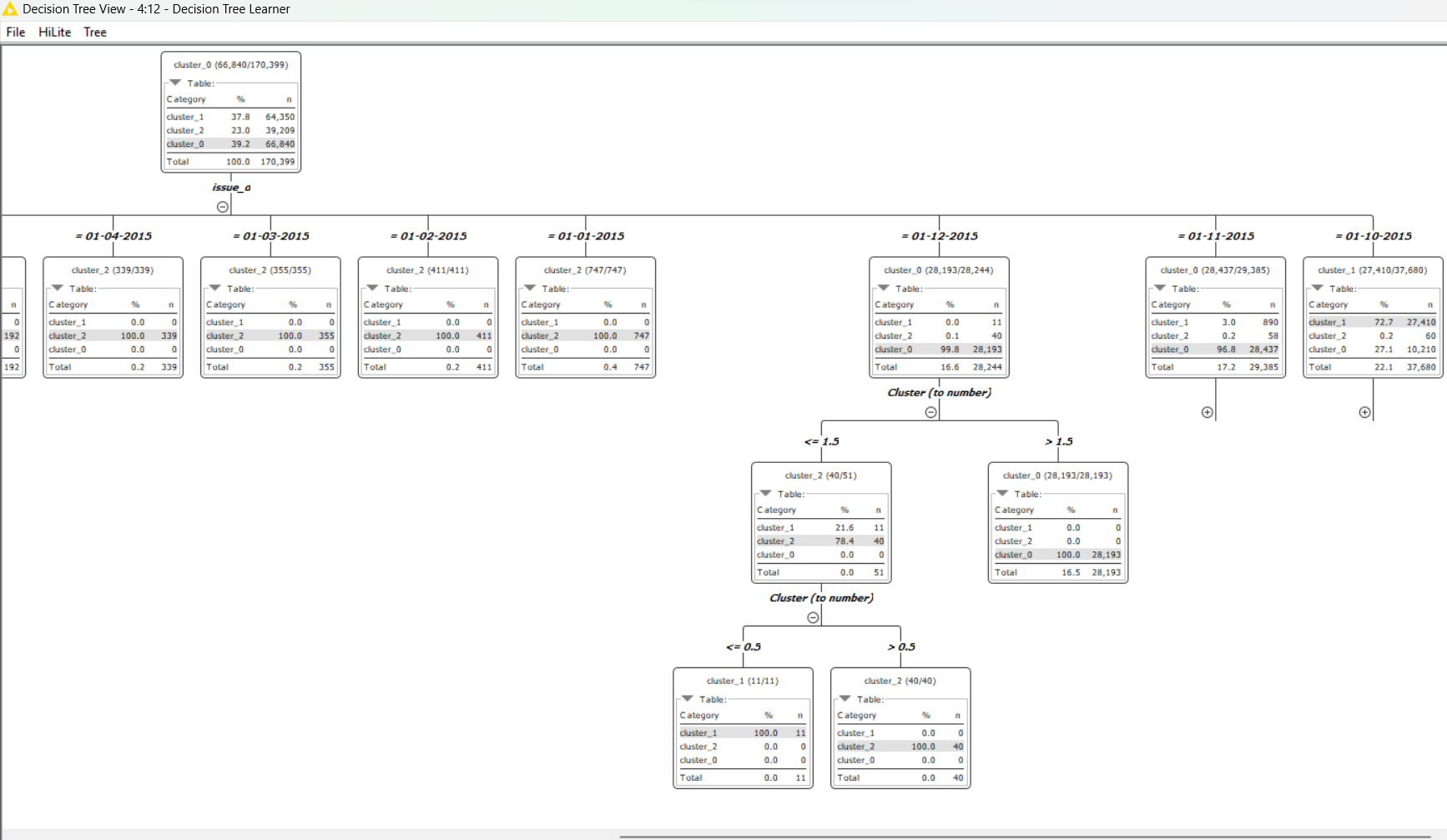
a stopping criterion is met, such as reaching a maximum depth or no further

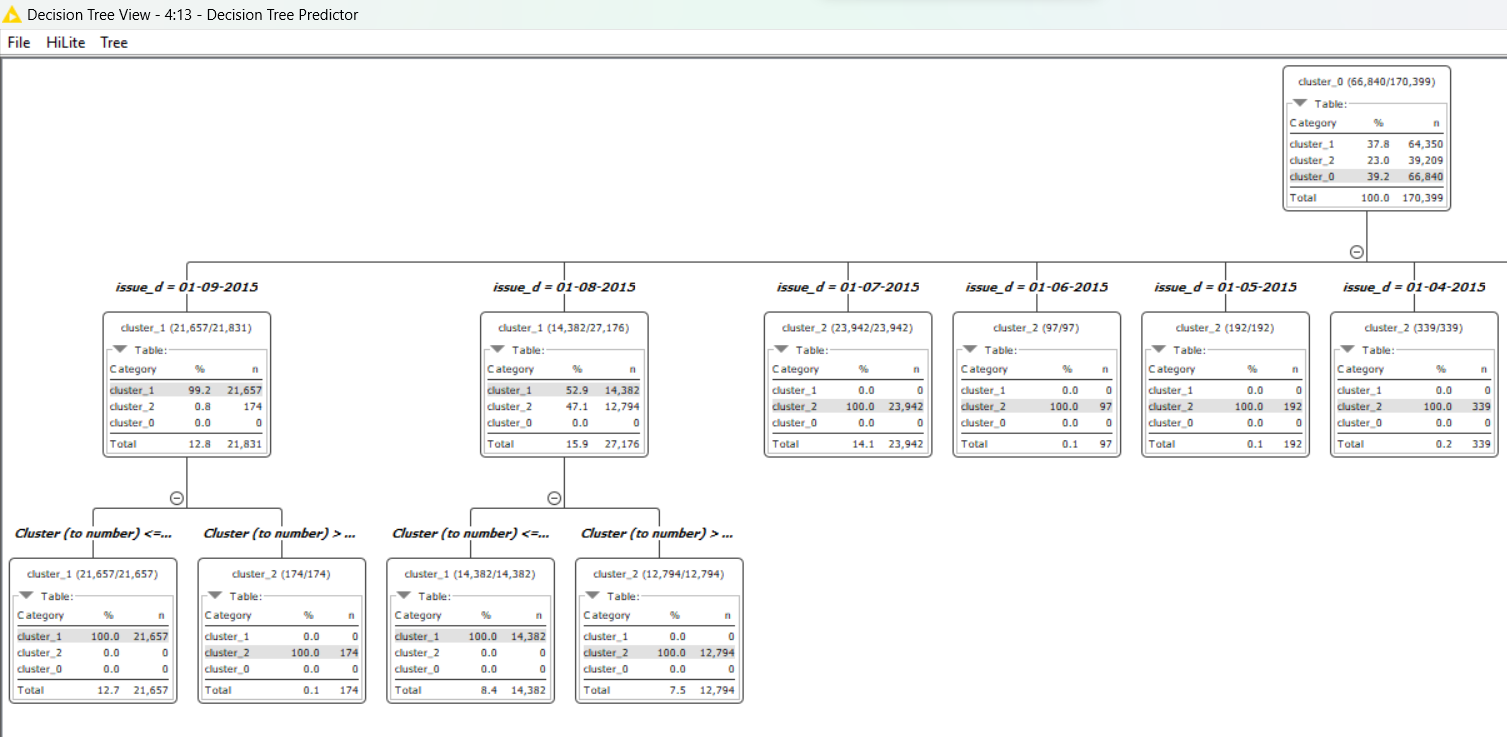
improvement in impurity reduction.

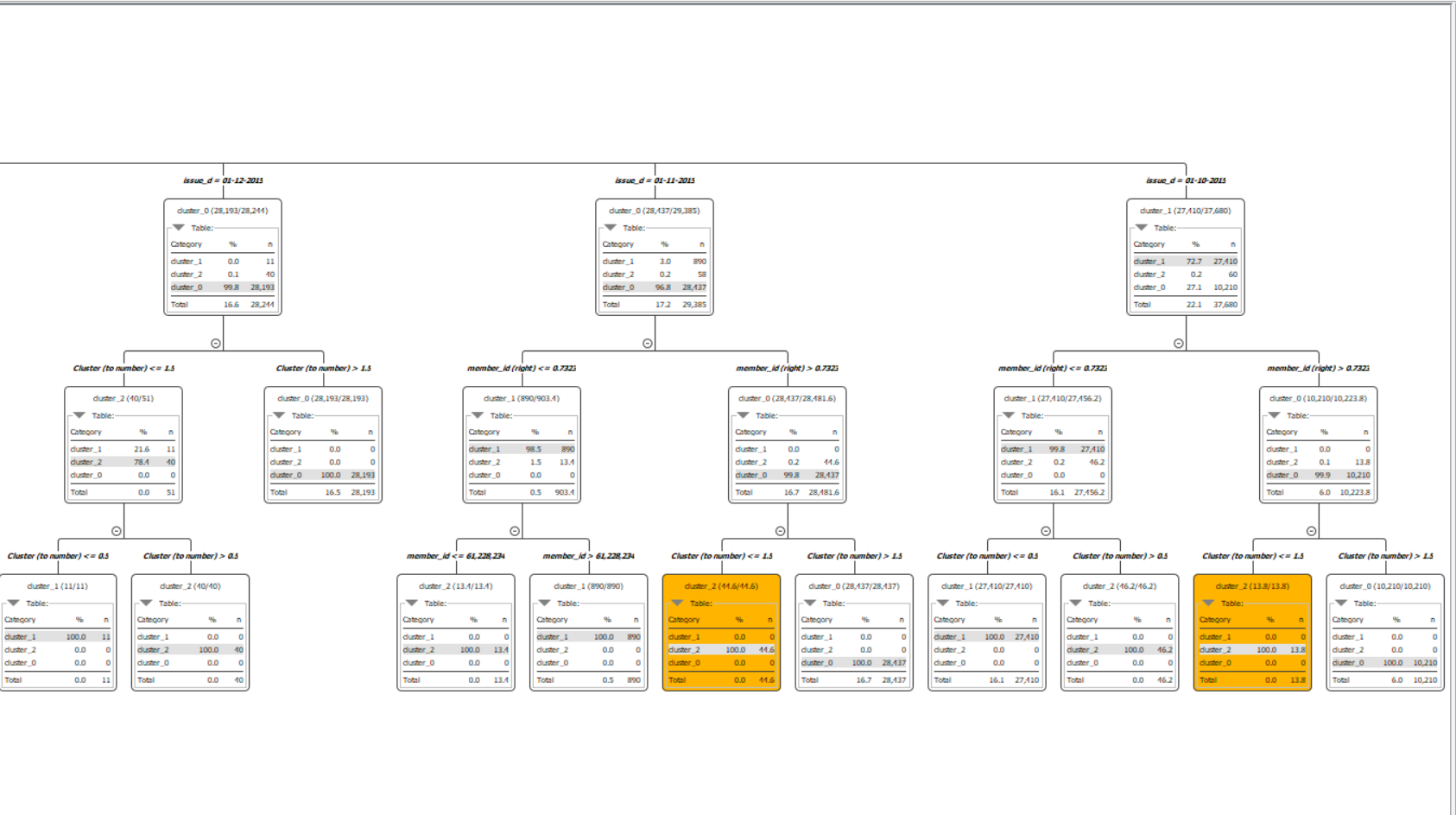
➔ In this project, decision tree will be the classification algorithm used for

unsupervised learning. The metrics used in decision tree is Gini coefficient.









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

**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).

**Support Vector Machines**

Support Vector Machine (SVM) is a powerful supervised learning algorithm usedfor classification and regression tasks. It works by finding the hyperplane that bestseparates the classes in the feature space, maximizing the margin between them.SVM can handle high-dimensional data and is effective even in cases where thenumber of features exceeds the number of samples.

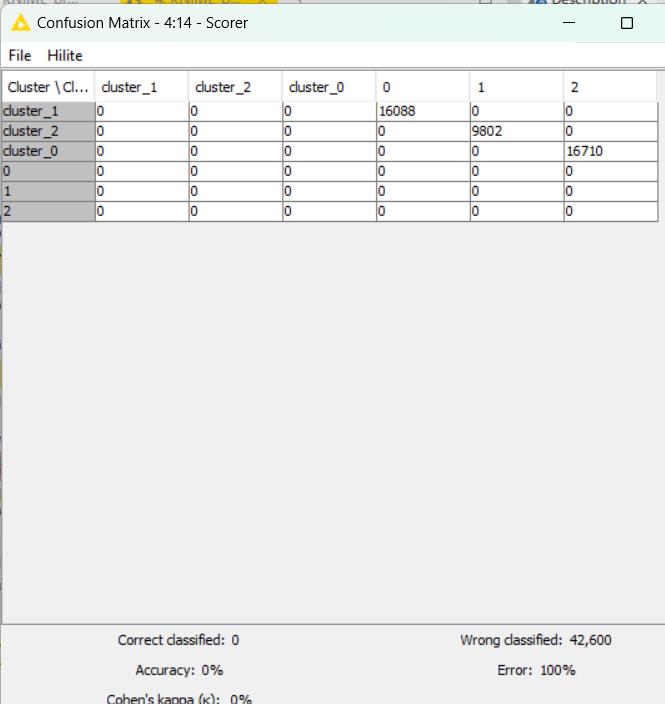
In this project, the kernel used will be polynomial and the parameters are power = 1, bias = 1 and gamma = 1.

**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.

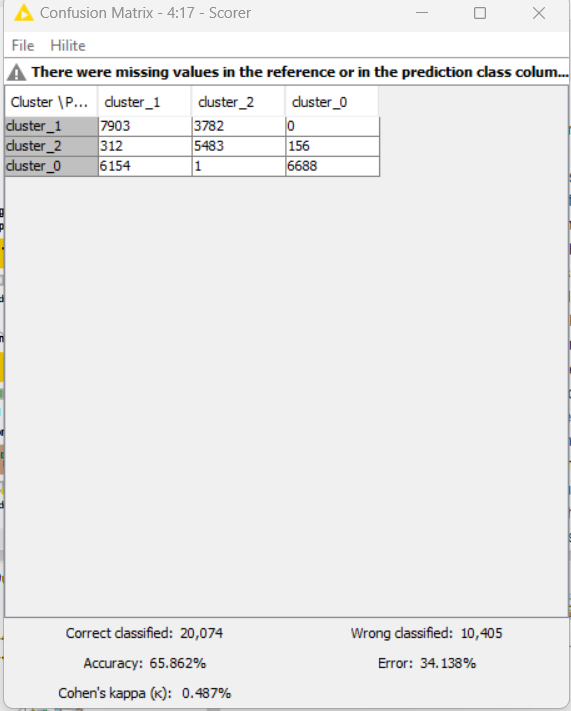
In this project, KNN will be used and the metric used is Euclidean distance.For comparison, we will be using k=7,9,11.

**3.2.2.1.1 PO2 | Classification Model Performance Evaluation: Confusion Matrix (Base Model: Decision Tree)**

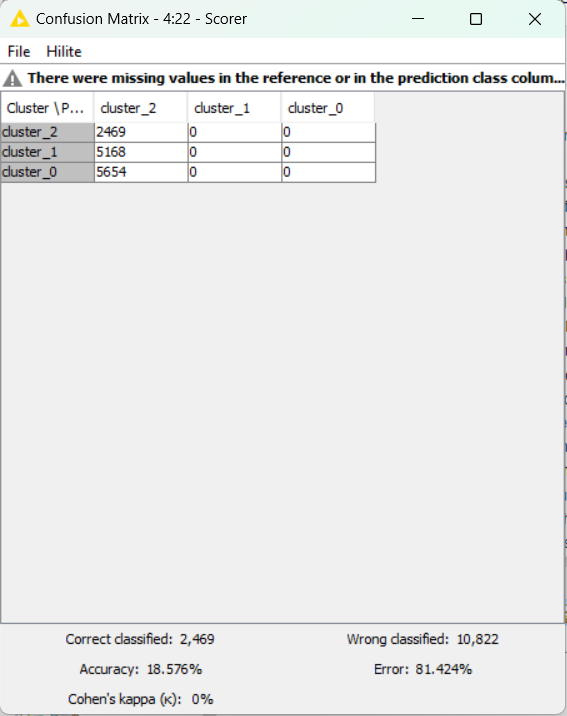


**3.2.2.2.1. PO2 | Classification Model Performance Evaluation: Confusion Matrix (Comparison Models: Logistic Regression | Support Vector Machine | K Nearest Neighbour)**

Logistic Regression

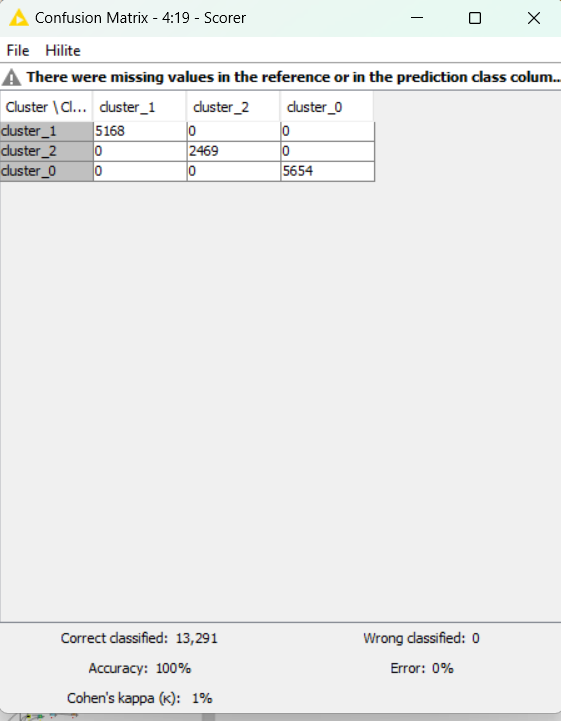


Support Vector Machine

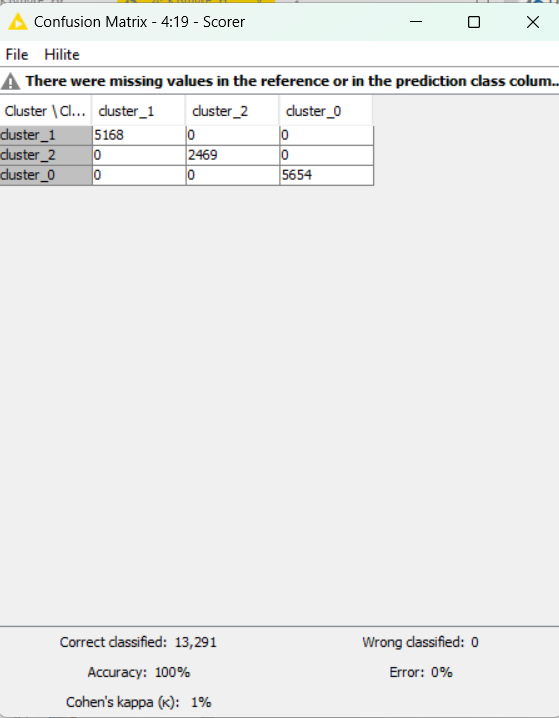


K nearest Neighbour

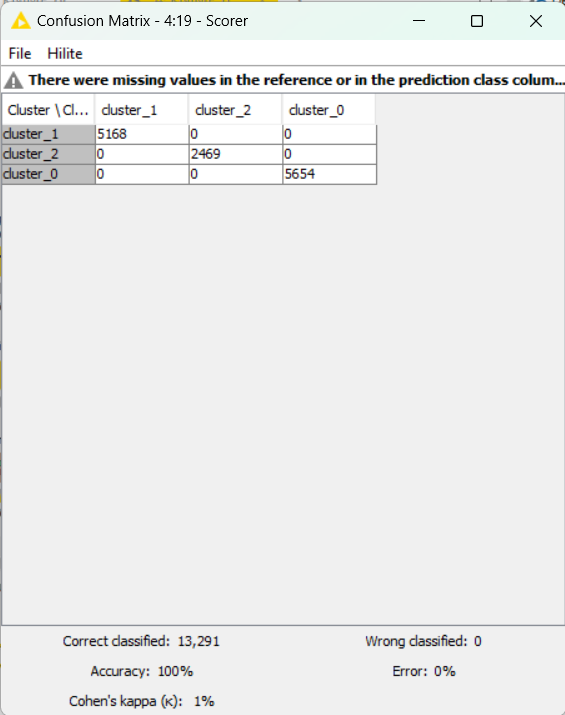
K=7



K=9

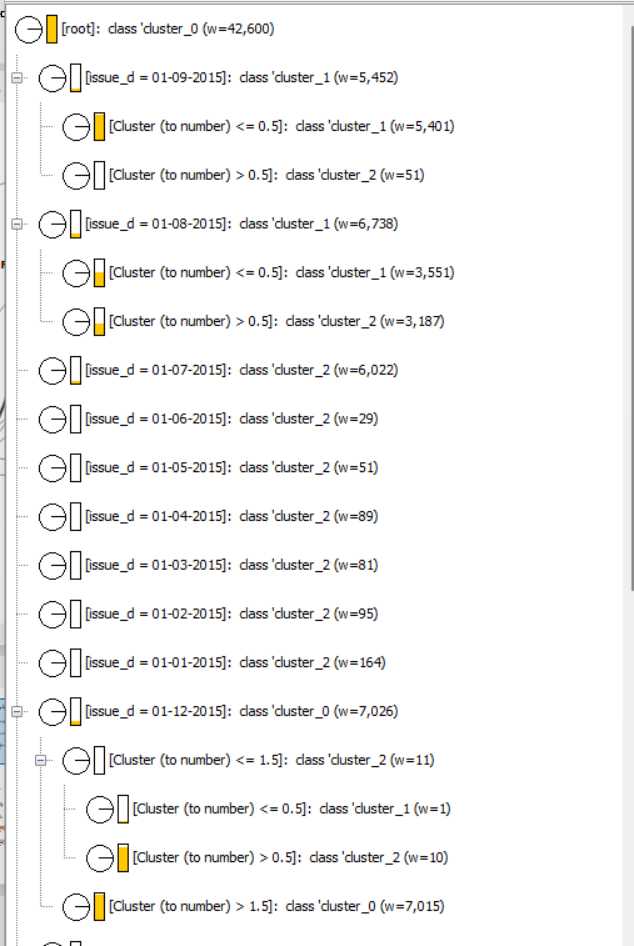


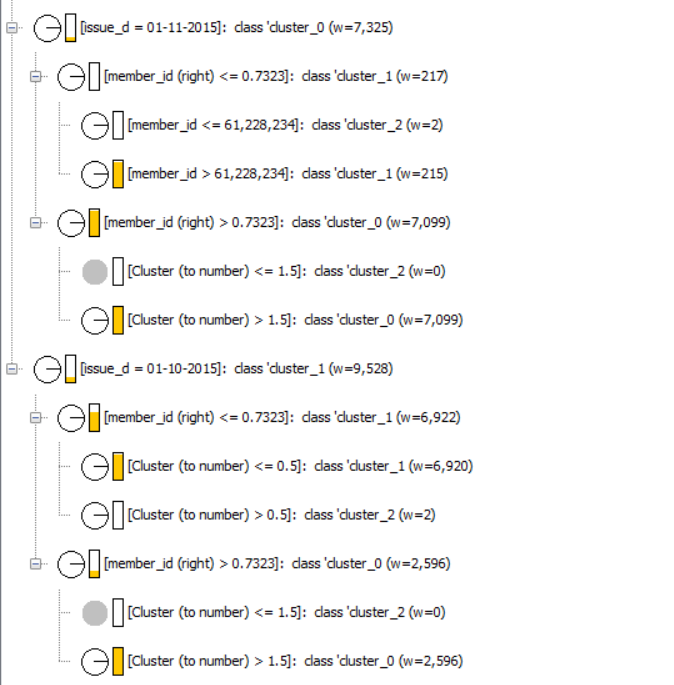
K=11



**3.2.3.1. PO3 | Variable Analysis: Base Model (Decision Tree)**

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





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

All the variables other than cluster, issue date and member id.

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

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

Grade, interest\_rate, dti (debt to income ratio), loan\_amount, total\_rec\_prncp (Total received principal)

1. Grade

• Variables related to the borrower's credit grade (grade=B, grade=C, grade=D, grade=E, grade=F, grade=G) have significant coefficients with p-values < 0.05. This indicates that the borrower's credit grade significantly influences the loan outcome.

2. Interest\_rate:

• The interest rate variable has a significant positive coefficient which suggests that higher interest rates are associated with higher odds of default or unfavourable loan conditions.

3. dti (Debt-to-Income Ratio):

• The debt-to-income ratio variable has a significant negative coefficient, implying that lower debt-to-income ratios are associated with better loan conditions.

4. loan\_amount:

• The loan amount variable has a significant positive coefficient indicating that larger loan amounts are associated with higher odds of default or unfavourable loan conditions.

5. total\_rec\_prncp (Total Received Principal):

• The total received principal variable has a significant positive coefficient meaning that higher amounts of principal received are associated with better loan conditions.

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

Purpose, Annual\_Income, Home\_Ownership, recoveries, emp\_lenght\_int (number of years employee worked), application\_type, interest\_payments, loan\_condition, region, annual\_inc, installment, term.

1. **Observations**

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

* Decision Tree:

Maximum depth of the tree-3

Minimum number of samples required to split a node-

Splitting criteria - Gini impurity, entropy

* Logistic Regression:

Regularization parameter (controls the strength of the regularization)

* SVM

Kernel function

Regularization parameter

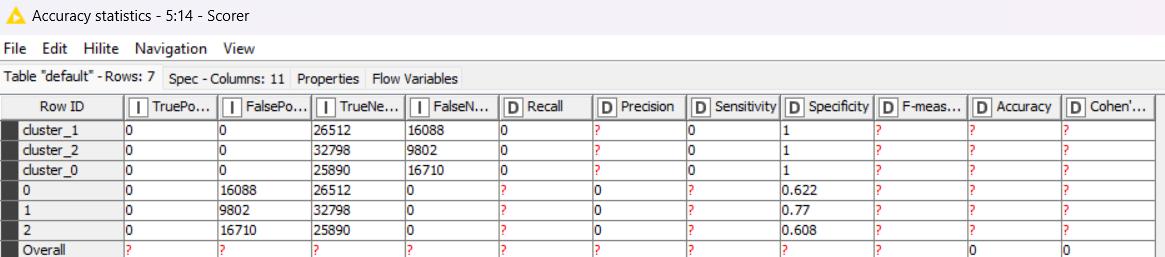
* KNN:

Number of neighbors (K)= 7,9,11

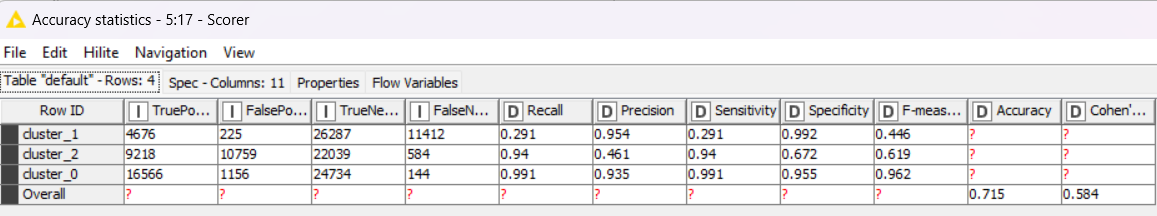
Distance metric: Euclidean Distance

**4.2. Classification Model Performance: [Base Model (Decision Tree) | Comparison Models (Logistic Regression | Support Vector Machine | K Nearest Neighbour)]**

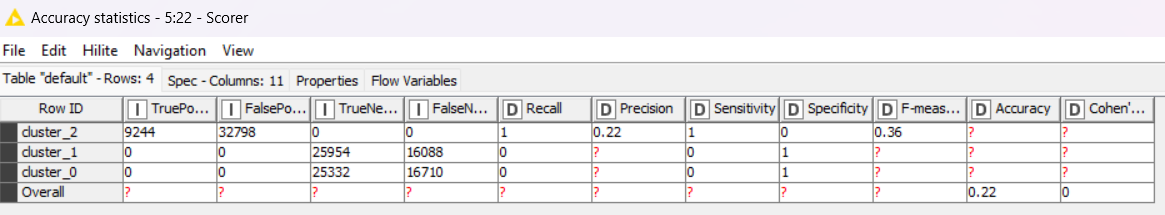
Decision Tree



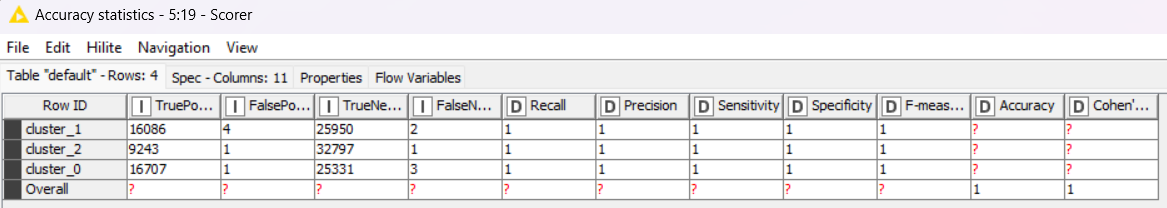
Logistic Regression



SVM



KNN



Decision Tree: This model appears to have the highest accuracy (0.622) but the lowest recall for some classes (duster 1 and duster 2). This suggests it is good at predicting non-defaults but not as good at predicting defaults.

Logistic Regression: This model has an overall accuracy of 0.715. It has a high precision for duster 0 (0.935) but a low precision for duster 1 (0.291). This means that Logistic Regression is good at predicting non-defaults for class duster 0 but not for duster 1.

SVM: This model has the lowest overall accuracy (0.22) out of the four models. It seems to have a very high precision for duster 2 (0.36) but poor performance on all other classes.

KNN: It has the highest overall accuracy (0.89) but some confusion between good and bad debtors in class Duster 0 (precision of 0.88). It perfectly classified both defaulters (Duster 1) and non-defaulters with good credit (Duster 2).

KNN has the highest overall accuracy (0.89) followed by Logistic Regression (0.715) and Decision Tree (0.622). SVM has the lowest overall accuracy (0.22).

KNN and SVM perfectly classified both defaulters (Cluster 1) and non-defaulters with good credit (Cluster 2). However, SVM performs poorly on the other classes.

Logistic Regression performs well in identifying good debtors in class Cluster 0 (recall of 0.991 and precision of 0.935) but has a lower performance for defaulters (Cluster 1) with a high precision (0.954) but low recall (0.291). This means it might often classify defaulters as non-defaulters.

Decision Tree has the lowest recall for defaulters (Cluster 1) but a perfect precision for Cluster 2. It also performs moderately on good debtors in Cluster 0.

**4.3. Variable Analysis: Base Model (Decision Tree) | Comparison Models (Logistic Regression | Support Vector Machine | K Nearest Neighbour)**

**4.3.1. List of Relevant or Important Variables and their Thresholds**

Grade, interest\_rate, dti (debt to income ratio), loan\_amount, total\_rec\_prncp (Total received principal)

1. Grade

• Variables related to the borrower's credit grade (grade=B, grade=C, grade=D, grade=E, grade=F, grade=G) have significant coefficients with p-values < 0.05. This indicates that the borrower's credit grade significantly influences the loan outcome.

2. Interest\_rate:

• The interest rate variable has a significant positive coefficient which suggests that higher interest rates are associated with higher odds of default or unfavourable loan conditions.

3. dti (Debt-to-Income Ratio):

• The debt-to-income ratio variable has a significant negative coefficient, implying that lower debt-to-income ratios are associated with better loan conditions.

4. loan\_amount:

• The loan amount variable has a significant positive coefficient indicating that larger loan amounts are associated with higher odds of default or unfavourable loan conditions.

5. total\_rec\_prncp (Total Received Principal):

• The total received principal variable has a significant positive coefficient meaning that higher amounts of principal received are associated with better loan conditions.

These variables had p <0.05 which shows its significance in the linear regression equation i.e. the impact of these variables is more in the classification of customers.

4.3.2. List of Non-Relevant or Non-Important Variables

Purpose, Annual\_Income, Home\_Ownership, recoveries, emp\_lenght\_int (number of years employee worked), application\_type, interest\_payments, loan\_condition, region, annual\_inc, installment, term.

Some of the variables had higher coefficients that should have impacted the regression equation but they have less significance due to P value > 0.05.

1. **Managerial Insights**

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Decision Tree | Logistic Regression | SVM | KNN |
| Accuracy | 0 | 0.715 | 0.22 | 1 |

K Nearest Neighbour has the highest accuracy of 100% followed by Logistic regression 71.5%. Support vector Machine has significantly lower accuracy of 22% whereas Decision tree if found to be absolutely inaccurate.

K Nearest Neighbour provides the highest accuracy of all the models according to the data and will be the appropriate model for customer classification. KNN is able to handle both numerical and categorical data which benefits the given dataset as it contains a combination of variables which are nominal, ordinal as well as non-categorical.

Advantages of KNN for Loan Default Prediction:

High Overall Accuracy: KNN has the highest overall accuracy (1) compared to the other models (Decision Tree: 0, Logistic Regression: 0.715, SVM: 0.22). This means it correctly classifies a higher proportion of loan applicants into defaulters and non-defaulters.

Balanced Performance Across Classes: KNN seems to perform well on both defaulters (Cluster 1) and non-defaulters with good credit (Cluster 2) based on achieving perfect recall (1.0) for these classes. This means it can effectively identify both high-risk and low-risk borrowers.

**Managerial Insights for Using KNN:**

*Improved Risk Assessment*: By using KNN, you can potentially improve the accuracy of your loan default risk assessment process. This can help you make better lending decisions by approving loans for qualified borrowers while mitigating the risk of defaults.

*Targeted Loan Products:* KNN's ability to classify borrowers across risk categories (Cluster 0, 1, 2) can be helpful for developing targeted loan products. You can design loan options with interest rates and terms tailored to the specific risk profiles identified by KNN.

*Focus on Data Quality*: KNN's performance is highly dependent on the quality and relevance of the training data. It's crucial to ensure your loan data is accurate, complete, and includes features that are most predictive of default risk.

*Model Monitoring and Refinement*: Like any machine learning model, KNN's performance can degrade over time as loan market conditions change. Regularly monitor the model's performance and retrain it with new data to maintain its accuracy.

**Limitations to Consider:**

*Interpretability*: KNN is a non-parametric model, meaning it can be difficult to understand the specific reasons behind its predictions. This might be a drawback if you need to explain or justify the model's decisions.

*Feature Selection*: KNN can be sensitive to the features used for training. It's important to carefully select relevant features that are most predictive of default risk.

*Computational Cost*: Classifying new loan applicants using KNN can be computationally expensive, especially for large datasets.

**5.2. Relevant or Important Variables – KNN Model**

* Loan Amount
* Interest Rate
* Loan Term (in years)
* Annual Income
* Debt-to-Income Ratio

Identify Key Drivers of Default: By analyzing the features with the highest weight or importance in the KNN model, the key factors most predictive of loan defaults can be identified. This can help to focus your loan approval criteria and risk assessment on these critical variables.

For instance, if credit score and debt-to-income ratio have high importance in the KNN model, managers can ensure these are prioritized during loan applications.

Targeted Customer Segments: KNN's ability to classify borrowers based on risk categories (predicted by these key variables) can be used to develop targeted customer segments. Marketing campaigns or loan products tailored to specific risk profiles can be created.

For example, borrowers with a high credit score and low debt-to-income ratio (predicted as low risk by KNN) might be offered lower interest rates or loan options with fewer restrictions.

Early Intervention for At-Risk Borrowers: The model can help identify borrowers with a higher predicted risk of default (based on KNN's classification). This allows for early intervention strategies like personalized financial counseling or loan restructuring options.

This proactive approach can help reduce defaults and improve loan recovery rates.