



**FOUNDATION FOR ORGANISATIONAL  
RESEARCH AND EDUCATION  
NEW DELHI**

**Academic Session 2023-2025**

**Customer Classification and Prediction (Loan Data)  
on the basis of Cluster data**

**Machine Learning for Managers**

**FMG 32 Section A**

**Submitted to:**

**Prof. Amarnath Mitra**

**Submitted by:**

**321035 - Prityush Agarwal**

## **Table of Contents**

<b>S. No</b>	<b>Title</b>	<b>Page Number</b>
<b>1</b>	<b>Project Objective</b>	<b>1</b>
<b>2</b>	<b>Data Description</b>	<b>2</b>
<b>3</b>	<b>Analysis</b>	<b>12</b>
<b>4</b>	<b>Results and observation</b>	<b>29</b>
<b>5</b>	<b>Managerial Insights</b>	<b>35</b>

## **1. Project Objectives**

→ The first objective is to segment the consumer (loan) data of the bank using supervised learning algorithms using Decision tree.

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

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

## **2. Data Description**

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

2.1.1. Link of the data: <https://www.kaggle.com/datasets/mrferozi/loan-data-for-dummy-bank>

2.1.2. The size of data is 30 MB.

2.1.3. Dimension of Data

- Number of Variables: The number of variables in the csv file is 30.
- Number of records: The number of records in the csv file is 1,06,485 (excluding naming column).

### **2.2. Description of Variables**

2.2.1 Index variables: id – gives the loan a unique identification (year, issue\_d and final\_d won't be used for evaluation purpose another variable term is being used to gauge how much time it took to repay the loan).

#### **2.2.2. Variables having categorical or non-categorical variables**

##### **2.2.2.1 Variables or Features having Nominal Categories:**

- cluster: This is the outcome variable. The results of the outcome variable I got from the previous project where we did unsupervised learning using K-means clustering.
- home\_ownership - home ownership status provided by the borrower during registration
- term – Term of the loan
- application\_type – Explains the status whether the account is individual or joint
- purpose – This variable tells the purpose why the loan was taken
- loan\_condition – This variable tells the status of the loan whether the loan is good or bad
- region – The region the loan was taken from

##### **2.2.2.2 Variables or Features having Ordinal Categories:**

- income\_category – This variable tells the bracket under which the person earns
- interest\_payments – This variable tells whether the interest payments on the loan is low or high
- grade – This variable tells the assigned grade of the loan

##### **2.2.2.3. Non-Categorical Variables:**

- emp\_length\_int – The number of years the person is employed
- annual\_inc – This variable tells the annual income the person earns
- loan\_amount – The variable tells the amount of loan that has been taken by the person

- interest\_rate – The variable tells the interest rate at which the loan needs to be paid
- dti - A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested loan, divided by the borrower's self-reported monthly income.
- total\_pymnt – This variable explains the total payment done against the loan
- total\_rec\_prncp – This variable explains the total received principal gotten from the loan
- recoveries – This variable explains the recoveries made from the bad loan
- installment – This variable explains the instalment made against the loan

## 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 (in %).

Row ID	I count	D Relative Frequency (in %)
cluster_0	77618	72.891
cluster_1	1183	1.111
cluster_2	27684	25.998

### 2.3.2. Descriptive Statistics of Input Categorical Variables

2.3.2.1. It provides the statistics of input variable (categorical variable) by giving frequency (count) as well as relative frequency (in %).

home\_ownership

Row ID	I count	D Relative Frequency (in %)
MORTGAGE	53079	49.846
NONE	5	0.005
OTHER	24	0.023
OWN	10502	9.862
RENT	42875	40.264

## Term

Row ID	I count	D Relative Frequency in %
36 months	74473	69.938
60 months	32012	30.062

## application\_type

Row ID	I count	D Relative Frequency in %
INDIVIDUAL	106427	99.946
JOINT	58	0.054

## Purpose

Row ID	I count	D Relative Frequency in %
car	1043	0.979
credit_card	24716	23.211
debt_consolidation	63092	59.25
educational	50	0.047
home_improvement	6193	5.816
house	433	0.407
major_purchase	2023	1.9
medical	1007	0.946
moving	661	0.621
other	5164	4.85
renewable_energy	67	0.063
small_business	1188	1.116
vacation	585	0.549
wedding	263	0.247

## loan\_condition

Row ID	I count	D Relative Frequency in %
Bad Loan	8084	7.592
Good Loan	98401	92.408

income\_category

Row ID	I count	D Relative Frequency in %
High	2020	1.897
Low	87400	82.077
Medium	17065	16.026

interest\_payments

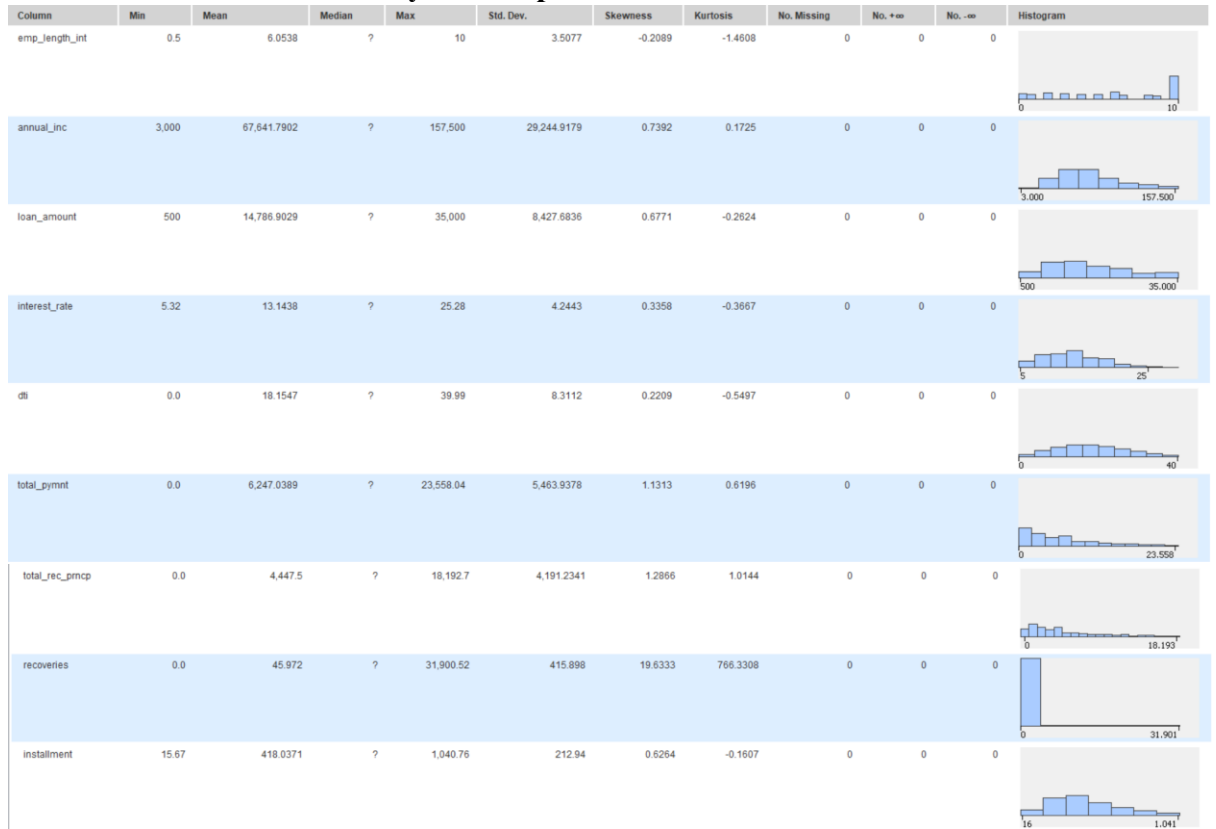
Row ID	I count	D Relative Frequency in %
High	50610	47.528
Low	55875	52.472

Grade

Row ID	I count	D Relative Frequency in %
A	17753	16.672
B	30642	28.776
C	29537	27.738
D	16703	15.686
E	8487	7.97
F	2729	2.563
G	634	0.595

## 2.3.3. Descriptive Statistics: Non-Categorical Variables

### 2.3.3.1. Measures of Central Tendency and Dispersion



Name ↓	Type	# Missi...	Minimum	Maximum	25% Quantile	50% Quantile (Median)	75% Quantile	Mean	Mean Absolut...	Standard Devi...	Variance	Skewness	Kurtosis	🔽
emp_length_int	Number (double)	0	0.5	10	3	6.05	10	6.054	3.115	3.508	12.304	-0.209	10.711	
annual_inc	Number (double)	0	3,000	157,500	45,000	65,000	85,000	67,641.79	23,007.866	29,244.918	855,265,222.1...	0.739	-1.265	
loan_amount	Number (double)	0	500	35,000	8,000	13,000	20,000	14,786.903	6,867.874	8,427.684	71,025,850.905	0.677	1.924	
interest_rate	Number (double)	0	5.32	25.28	9.99	12.99	15.88	13.144	3.417	4.244	18.014	0.336	2.689	
dti	Number (double)	0	0	39.99	11.9	17.71	23.97	18.155	6.819	8.311	69.076	0.221	4.031	
total_pymnt	Number (double)	0	0	23,558.04	1,918.7	4,868.8	8,905.36	6,247.039	4,282.876	5,463.938	29,854,616.597	1.131	-4.543	
total_rec_prncp	Number (double)	0	0	18,192.7	1,203.92	3,200	6,140.795	4,447.5	3,235.237	4,191.234	17,566,443.385	1.287	-7.438	
recoveries	Number (double)	0	0	31,900.52	0	0	0	45.972	89.481	415.898	172,971.14	19.633	-5,618.971	
installment	Number (double)	0	15.67	1,040.76	261.65	382.87	547.84	418.037	171.352	212.94	45,343.458	0.626	1.178	

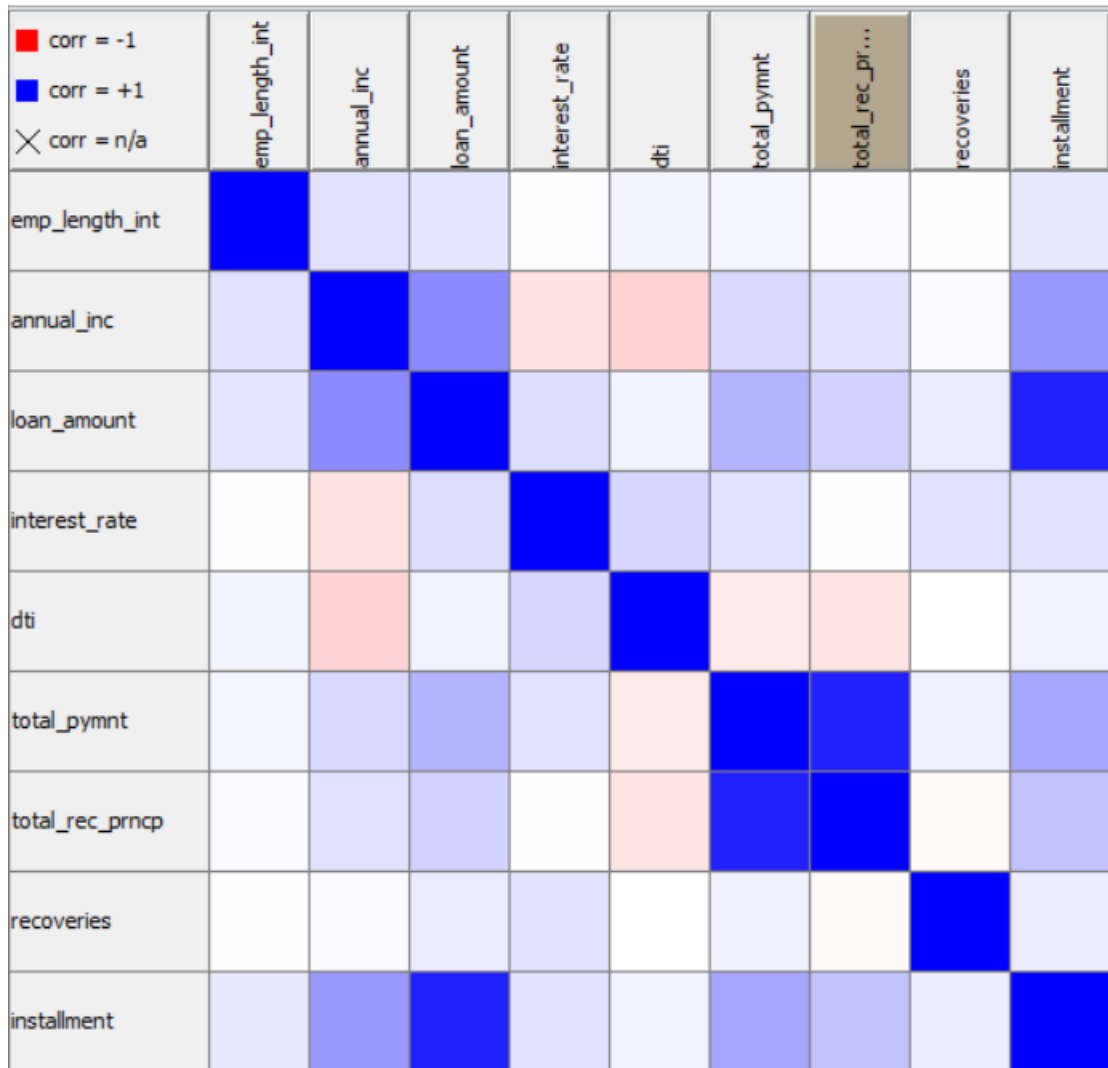


### 2.3.3.2. Correlation Statistics (using Test of Correlation)

Row ID	S First column name	S Second column name	D Correlation value	D p value	I Degrees of freedom
Row0	emp_length_int	annual_inc	0.11911012653120...	0.0	106483
Row1	emp_length_int	loan_amount	0.10171130496082...	0.0	106483
Row2	emp_length_int	interest_rate	0.00675005509605...	0.027617319551575648	106483
Row3	emp_length_int	dti	0.0445547463056736	0.0	106483
Row4	emp_length_int	total_pymnt	0.03741256187103...	0.0	106483
Row5	emp_length_int	total_rec_prncp	0.01743498543779...	1.2722030318101929E-8	106483
Row6	emp_length_int	recoveries	0.00435276338086...	0.1554954763978429	106483
Row7	emp_length_int	installment	0.08826361219589...	0.0	106483
Row8	annual_inc	loan_amount	0.45639069673328...	0.0	106483
Row9	annual_inc	interest_rate	-0.10890036507563...	3.130798239529113E-278	106483
Row10	annual_inc	dti	-0.1734362843491869	0.0	106483
Row11	annual_inc	total_pymnt	0.14931826101783...	0.0	106483
Row12	annual_inc	total_rec_prncp	0.11262203460731...	0.0	106483
Row13	annual_inc	recoveries	0.01508971149248...	8.466150787089788E-7	106483
Row14	annual_inc	installment	0.4045213669226771	0.0	106483
Row15	loan_amount	interest_rate	0.13267165502297...	0.0	106483
Row16	loan_amount	dti	0.04671651876506...	0.0	106483
Row17	loan_amount	total_pymnt	0.29787380746640...	0.0	106483
Row18	loan_amount	total_rec_prncp	0.1768913105828052	0.0	106483
Row19	loan_amount	recoveries	0.07681736222226...	0.0	106483
Row20	loan_amount	installment	0.8727387314553934	0.0	106483
Row21	interest_rate	dti	0.1625930381940522	0.0	106483
Row22	interest_rate	total_pymnt	0.10899088797054...	0.0	106483
Row23	interest_rate	total_rec_prncp	-0.00264481012310...	0.3881117024827133	106483
Row24	interest_rate	recoveries	0.11364039053627...	0.0	106483
Row25	interest_rate	installment	0.11244970536063...	0.0	106483
Row26	dti	total_pymnt	-0.08140112488851...	5.679030877555581E-156	106483
Row27	dti	total_rec_prncp	-0.10606357586467...	5.918196961921463E-264	106483
Row28	dti	recoveries	5.40189616100556...	0.8600796044771082	106483
Row29	dti	installment	0.04974955052132...	0.0	106483
Row30	total_pymnt	total_rec_prncp	0.8707573781921869	0.0	106483
Row31	total_pymnt	recoveries	0.06023104760486...	0.0	106483
Row32	total_pymnt	installment	0.3504988922257118	0.0	106483
Row33	total_rec_prncp	recoveries	-0.02247225026477...	2.2339385071952356E-13	106483
Row34	total_rec_prncp	installment	0.23863177319903...	0.0	106483
Row35	recoveries	installment	0.07088416491241...	0.0	106483

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

Row ID	D emp_length_int	D annual_inc	D loan_amount	D interest_rate	D dti	D total_pymnt	D total_rec_prncp	D recoveries	D installment
emp_length_int	1.0	0.11911012653120064	0.10171130496082201	0.0067500550960573145	0.0445547463056736	0.037412561871036316	0.017434985437797604	0.004352763380868477	0.08826361219589225
annual_inc	0.11911012653120064	1.0	0.45639069673328797	-0.10890036507563082	-0.1734362843491869	0.14931826101783496	0.11262203460731779	0.01508971149248958	0.4045213669226771
loan_amount	0.10171130496082201	0.45639069673328797	1.0	0.13267165502297726	0.04671651876506626	0.29787380746640885	0.1768913105828052	0.07681736222226498	0.8727387314553934
interest_rate	0.0067500550960573145	-0.10890036507563082	0.13267165502297726	1.0	0.1625930381940522	0.10899088797054506	-0.0026448101231001395	0.11364039053627975	0.11244970536063416
dti	0.0445547463056736	-0.1734362843491869	0.04671651876506626	0.1625930381940522	1.0	-0.08140112488851041	-0.10606357586467281	5.401896161005566E-4	0.04974955052132508
total_pymnt	0.037412561871036316	0.14931826101783496	0.29787380746640885	0.10899088797054506	-0.08140112488851041	1.0	0.8707573781921869	0.06023104760486053	0.3504988922257118
total_rec_prncp	0.017434985437797604	0.11262203460731779	0.1768913105828052	-0.0026448101231001395	-0.10606357586467281	0.8707573781921869	1.0	-0.022472250264771516	0.23863177319903192
recoveries	0.004352763380868477	0.01508971149248958	0.07681736222226498	0.11364039053627975	5.401896161005566E-4	0.06023104760486053	-0.022472250264771516	1.0	0.07088416491241371
installment	0.08826361219589225	0.4045213669226771	0.8727387314553934	0.11244970536063416	0.04974955052132508	0.3504988922257118	0.23863177319903192	0.07088416491241371	1.0



### 3. 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 Categorical Variables**

In this case, category to number node will be used to encode the categorical variables.

home\_ownership

mortgage - 3, none - 5, other - 4, own - 2, rent - 1

Term

36 months - 1, 60 months - 2

application\_type

Individual - 1, Joint - 2

Purpose

Credit card - 1, car - 2, small business - 3, other - 4, wedding - 5, debt consolidation - 6, home improvement - 7, major purchase - 8, medical - 9, moving - 10, vacation - 11, house - 12, renewable energy - 13, educational - 14

loan\_condition

Good Loan - 1, Bad Loan - 2

Region

Munster - 1, Leinster - 2, Cannught - 3, Ulster - 4, Northern-Irl - 5

income\_category

Low - 1, Medium - 2, High - 3

interest\_payments

Low - 1, High - 2

Grade

B - 1, C - 2, A - 3, E - 4, F - 5, D - 6, G - 7

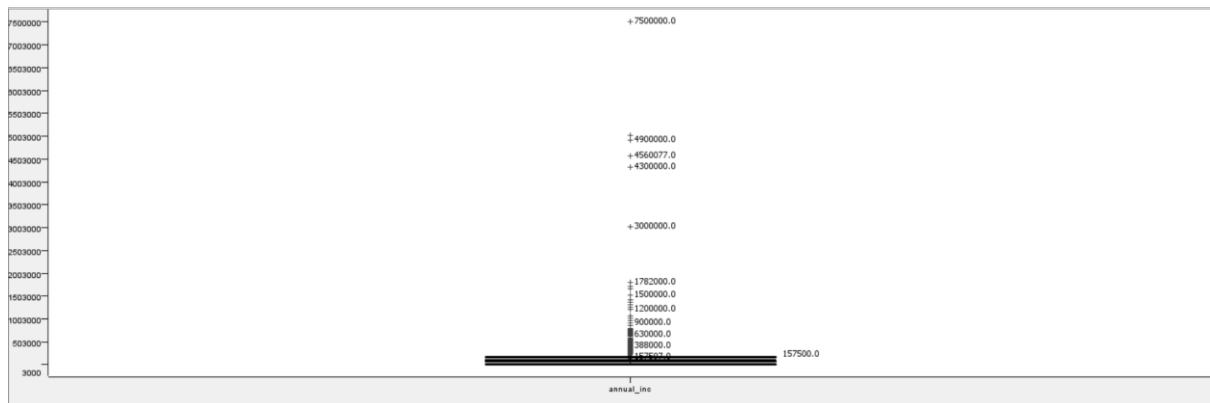
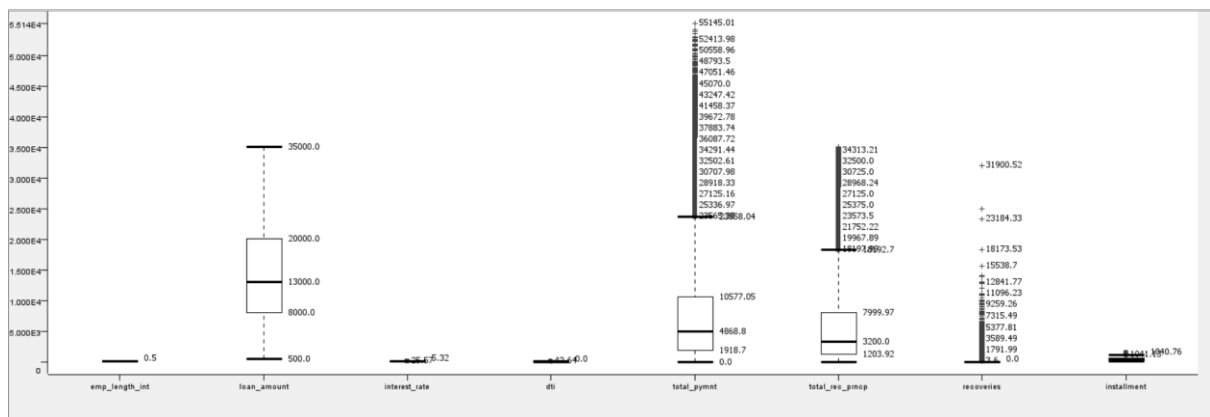
### 3.1.3. Outlier Statistics and Treatment

#### 3.1.3.1.Outlier Statistics: Non-Categorical Variables

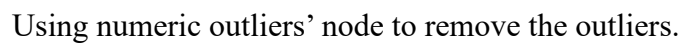
Row ID	emp_length_int	annual_inc	loan_amount	interest_rate	dti	total_pymnt	total_rec_pncp	recoveries	installment
Minimum	0.5	3,000	500	5.32	0	0	0	0	15.67
Smallest	0.5	3,000	500	5.32	0	0	0	0	15.67
Lower Quartile	3	45,000	8,000	9.99	11.9	1,918.7	1,203.92	0	261.65
Median	6.05	65,000	13,000	12.99	17.71	4,868.8	3,200	0	382.87
Upper Quartile	10	90,000	20,000	16.2	23.98	10,577.05	7,999.97	0	573.35
Largest	10	157,500	35,000	25.28	42.1	23,558.04	18,192.7	0	1,040.76
Maximum	10	7,500,000	35,000	28.99	104	55,145.01	35,000.03	31,900.52	1,445.46

#### 3.1.3.2.Normalization using Min-Max Scaler

Before Normalization



Min-Max Scaler Normalization (between 0 and 1) for variables: annual\_inc, interest\_rate, dti, total\_pymnt, total\_rec\_prncp, recoveries, installment



#### 3.1.4. Data Bifurcation

The bifurcation schema used is stratified sampling on the basis of outcome variable cluster variable with 80% (training data) and 20% (testing data).

### 3.2. Data Analysis

#### 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

Row ID	I TruePositives	I FalsePositives	I TrueNegatives	I FalseNegatives	D Recall	D Precision	D Sensitivity	D Specificity	D F-measure	D Accuracy	D Cohen's kappa
cluster_0	15499	21	5745	23	0.999	0.999	0.999	0.996	0.999	?	?
cluster_2	5434	120	15634	100	0.982	0.978	0.982	0.992	0.98	?	?
cluster_1	135	79	20977	97	0.582	0.631	0.582	0.996	0.605	?	?
Overall	?	?	?	?	?	?	?	?	?	0.99	0.974

Row ID	I cluster_0	I cluster_2	I cluster_1
cluster_0	15499	23	0
cluster_2	21	5434	79
cluster_1	0	97	135

##### With Pruning

Row ID	I TruePositives	I FalsePositives	I TrueNegatives	I FalseNegatives	D Recall	D Precision	D Sensitivity	D Specificity	D F-measure	D Accuracy	D Cohen's kappa
cluster_0	15497	25	5749	26	0.998	0.998	0.998	0.996	0.998	?	?
cluster_2	5441	98	15662	96	0.983	0.982	0.983	0.994	0.982	?	?
cluster_1	165	71	20989	72	0.696	0.699	0.696	0.997	0.698	?	?
Overall	?	?	?	?	?	?	?	?	?	0.991	0.977

Row ID	I cluster_0	I cluster_2	I cluster_1
cluster_0	15497	26	0
cluster_2	25	5441	71
cluster_1	0	72	165

#### Cluster 0

- 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 1**

- 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 2**

- This cluster has a relatively high recall and precision, indicating that the model performs well in identifying high-income and low-risk borrowers.
- The number of false positives is relatively low suggesting that the model effectively minimizes misclassifications within this cluster.
- Both sensitivity and specificity scores are high indicating that the model correctly identifies both true positives and 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 increased with pruning but the recall decreased, indicating that the model becomes more conservative in predicting positive cases resulting in fewer false positives but also fewer true positives.
- In cluster\_0 and cluster\_2, pruning led to a slight decrease in precision and specificity but an increase in recall and sensitivity indicating that the model becomes less conservative resulting in more true positives but also more 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

Row ID	S Logit	S Variable	D Coeff.	D Std. Err.	D z-score	D P> z
Row1	cluster_1	home_ownership=MORTGAGE	5.483	1,446,932.624	0	1
Row2	cluster_1	home_ownership=NONE	0.001	187,671,46...	0	1
Row3	cluster_1	home_ownership=OTHER	0.004	9,255,151,7...	0	1
Row4	cluster_1	home_ownership=OWN	5.357	1,446,908.16	0	1
Row5	cluster_1	home_ownership=RENT	5.352	1,447,015.439	0	1
Row6	cluster_1	income_category=Low	-87.618	230,092.592	-0	1
Row7	cluster_1	income_category=Medium	-23.125	81,424,888....	-0	1
Row8	cluster_1	term=60 months	-0.372	0.249	-1.494	0.135
Row9	cluster_1	application_type=JOINT	-0	57,588,969,...	-0	1
Row10	cluster_1	purpose=credit_card	-0.085	1.025	-0.083	0.934
Row11	cluster_1	purpose=debt_consolidation	0.226	1.014	0.223	0.823
Row12	cluster_1	purpose=educational	-0	1,059,333,2...	-0	1
Row13	cluster_1	purpose=home_improvement	0.461	1.057	0.436	0.663
Row14	cluster_1	purpose=house	-1.682	1.833	-0.918	0.359
Row15	cluster_1	purpose=major_purchase	1.155	1.185	0.974	0.33
Row16	cluster_1	purpose=medical	1.658	1.36	1.219	0.223
Row17	cluster_1	purpose=moving	-0.178	1.434	-0.124	0.901
Row18	cluster_1	purpose=other	0.531	1.082	0.49	0.624
Row19	cluster_1	purpose=renewable_energy	0.001	272,357,37...	0	1
Row20	cluster_1	purpose=small_business	0.631	1.244	0.507	0.612
Row21	cluster_1	purpose=vacation	-1.68	1.815	-0.926	0.355
Row22	cluster_1	purpose=wedding	25.103	57,954.513	0	1
Row23	cluster_1	interest_payments=Low	0.031	0.404	0.077	0.938
Row24	cluster_1	loan_condition=Good Loan	0.685	0.431	1.587	0.113
Row25	cluster_1	grade=B	-1.209	0.377	-3.211	0.001
Row26	cluster_1	grade=C	-2.357	0.568	-4.148	0
Row27	cluster_1	grade=D	-2.925	0.793	-3.69	0
Row28	cluster_1	grade=E	-4.908	1.002	-4.9	0
Row29	cluster_1	grade=F	-4.729	1.222	-3.868	0
Row30	cluster_1	grade=G	-2.58	1.158	-2.228	0.026
Row31	cluster_1	region=cannught	-0.258	0.297	-0.869	0.385
Row32	cluster_1	region=leinster	-0.581	0.268	-2.172	0.03
Row33	cluster_1	region=munster	-0.344	0.325	-1.058	0.29
Row34	cluster_1	region=ulster	0.022	0.265	0.082	0.934
Row35	cluster_1	emp_length_int	0.019	0.258	0.073	0.942
Row36	cluster_1	annual_inc	2.486	4,055,625.828	0	1
Row37	cluster_1	loan_amount	14.359	0.819	17.523	0
Row38	cluster_1	interest_rate	6.909	1.474	4.686	0
Row39	cluster_1	dti	-3.214	0.534	-6.019	0
Row40	cluster_1	total_pymnt	1.314	0.708	1.856	0.063
Row41	cluster_1	total_rec_prncp	2.637	0.751	3.513	0
Row42	cluster_1	recoveries	-2.334	7.154	-0.326	0.744
Row43	cluster_1	installment	0.688	0.695	0.99	0.322
Row44	cluster_1	Constant	16.162	1,372,935.443	0	1
Row45	cluster_2	home_ownership=MORTGAGE	-31.106	522,982.131	-0	1

Row46	cluster_2	home_ownership=NONE	-0	579,998,09...	-0	1
Row47	cluster_2	home_ownership=OTHER	-38.607	522,981.961	-0	1
Row48	cluster_2	home_ownership=OWN	-31.479	522,982.131	-0	1
Row49	cluster_2	home_ownership=RENT	-31.324	522,982.131	-0	1
Row50	cluster_2	income_category=Low	-88.308	43,778.853	-0.002	0.998
Row51	cluster_2	income_category=Medium	23.124	81,424,917....	0	1
Row52	cluster_2	term=60 months	-0.47	0.209	-2.247	0.025
Row53	cluster_2	application_type=JOINT	0.644	16.119	0.04	0.968
Row54	cluster_2	purpose=credit_card	-0.454	0.841	-0.54	0.589
Row55	cluster_2	purpose=debt_consolidation	-0.075	0.832	-0.091	0.928
Row56	cluster_2	purpose=educational	-3.907	82.469	-0.047	0.962
Row57	cluster_2	purpose=home_improvement	0.222	0.875	0.254	0.8
Row58	cluster_2	purpose=house	-1.625	1.625	-1	0.317
Row59	cluster_2	purpose=major_purchase	0.592	0.971	0.61	0.542
Row60	cluster_2	purpose=medical	0.34	1.104	0.308	0.758
Row61	cluster_2	purpose=moving	-0.944	1.15	-0.82	0.412
Row62	cluster_2	purpose=other	-0.031	0.892	-0.034	0.972
Row63	cluster_2	purpose=renewable_energy	-6.591	48,312.976	-0	1
Row64	cluster_2	purpose=small_business	0.258	1.054	0.245	0.806
Row65	cluster_2	purpose=vacation	-0.82	1.249	-0.657	0.511
Row66	cluster_2	purpose=wedding	1.368	1.887	0.725	0.469
Row67	cluster_2	interest_payments=Low	0.047	0.322	0.146	0.884
Row68	cluster_2	loan_condition=Good Loan	0.514	0.346	1.486	0.137
Row69	cluster_2	grade=B	-1.13	0.317	-3.567	0
Row70	cluster_2	grade=C	-2.241	0.471	-4.757	0
Row71	cluster_2	grade=D	-2.917	0.66	-4.422	0
Row72	cluster_2	grade=E	-4.712	0.841	-5.6	0
Row73	cluster_2	grade=F	-4.632	1.012	-4.576	0
Row74	cluster_2	grade=G	-2.588	0.945	-2.739	0.006
Row75	cluster_2	region=cannught	-0.332	0.235	-1.411	0.158
Row76	cluster_2	region=leinster	-0.345	0.222	-1.553	0.12
Row77	cluster_2	region=munster	-0.091	0.272	-0.334	0.738
Row78	cluster_2	region=ulster	0.377	0.222	1.703	0.089
Row79	cluster_2	emp_length_int	-0.131	0.214	-0.612	0.54
Row80	cluster_2	annual_inc	445.549	17.723	25.139	0
Row81	cluster_2	loan_amount	13.317	0.781	17.05	0
Row82	cluster_2	interest_rate	7.017	1.228	5.712	0
Row83	cluster_2	dti	-0.52	0.402	-1.294	0.196
Row84	cluster_2	total_pymnt	1.53	0.578	2.65	0.008
Row85	cluster_2	total_rec_prncp	2.352	0.612	3.843	0
Row86	cluster_2	recoveries	-6.741	5.585	-1.207	0.227
Row87	cluster_2	installment	0.741	0.631	1.174	0.24
Row88	cluster_2	Constant	-132.625	523,626.232	-0	1

**Cluster\_0 was used as the reference category**

### **Cluster\_1**

- Having a low-income category appears to have a strong negative impact on being in cluster\_1 emphasizing that low-income individuals are less likely to belong to this cluster.
- The grade of the loan (Grade B, C, D, E, F, G) also plays a significant role in distinguishing Cluster 1 from Cluster 0. Higher-grade loans are less likely to be in Cluster 1.

- Variables such as term, purpose and interest rate do not seem to have a significant impact on distinguishing Cluster 1 from Cluster 0.

## Cluster\_2

- Similar to cluster\_1, income category is a significant predictor of being in cluster\_2. Borrowers with higher income categories are more likely to belong to cluster\_2.
- Grade of the loan also significantly influences the likelihood of being in cluster\_2. Higher-grade loans are less likely to be in cluster\_2.
- Interestingly, some purposes of the loan (e.g., renewable energy, small business) appear to have no significant impact on distinguishing cluster\_2 from cluster\_0.
- Variables such as term, application type and region do not seem to have a significant impact on distinguishing cluster\_2 from cluster\_0.

Row ID	I TruePositives	I FalsePositives	I TrueNegatives	I FalseNegatives	D Recall	D Precision	D Sensitivity	D Specificity	D F-measure	D Accuracy	D Cohen's kappa
cluster_0	15492	32	5742	31	0.998	0.998	0.998	0.994	0.998	?	?
cluster_2	5412	81	15679	125	0.977	0.985	0.977	0.995	0.981	?	?
cluster_1	187	93	20967	50	0.789	0.668	0.789	0.996	0.723	?	?
Overall	?	?	?	?	?	?	?	?	?	0.99	0.976

Row ID	I cluster_0	I cluster_2	I cluster_1
cluster_0	15492	31	0
cluster_2	32	5412	93
cluster_1	0	50	187

## Cluster\_0

- True Positives: 15492, False Positives: 32, True Negatives: 5742, False Negatives: 31
- Recall: 0.998, Precision: 0.998, Sensitivity: 0.998, Specificity: 0.994
- F-measure: 0.998, Accuracy: 0.998

Cluster\_0 demonstrates very high performance across all metrics indicating that the logistic regression model is highly effective at correctly classifying instances into cluster\_0. The high values of recall, precision, sensitivity, specificity, F-measure and accuracy suggest that the model has strong predictive power for this cluster.

## Cluster\_1

- True Positives: 187, False Positives: 93, True Negatives: 20967, False Negatives: 50
- Recall: 0.789, Precision: 0.668, Sensitivity: 0.789, Specificity: 0.996
- F-measure: 0.723, Accuracy: 0.990

Cluster\_1 shows slightly lower performance compared to cluster\_0 and cluster\_2 with lower values of recall, precision, sensitivity and F-measure. However, the specificity remains high,

indicating that the model effectively identifies true negatives for cluster\_1. Despite the lower values for some metrics the overall accuracy is still high.

## Cluster\_2

- True Positives: 5412, False Positives: 81, True Negatives: 15679, False Negatives: 125
- Recall: 0.977, Precision: 0.985, Sensitivity: 0.977, Specificity: 0.995
- F-measure: 0.981, Accuracy: 0.990

Cluster\_2 also exhibits high performance metrics with high values of recall, precision, sensitivity, specificity, F-measure and accuracy. This tells that the logistic regression model accurately classifies instances into cluster\_2, with relatively low false positive and false negative rates.

The overall accuracy of the logistic regression model is very high at 99 and it effectively predicts the cluster labels for the majority of instances. Additionally, the Cohen's Kappa coefficient suggests substantial agreement beyond chance among the predicted and actual cluster labels.

## K-Nearest Neighbour

K=7

Row ID	I TruePositives	I FalsePositives	I TrueNegatives	I FalseNegatives	D Recall	D Precision	D Sensitivity	D Specificity	D F-measure	D Accuracy	D Cohen's kappa
cluster_0	14256	5318	456	1267	0.918	0.728	0.918	0.079	0.812	?	?
cluster_2	445	1277	14483	5092	0.08	0.258	0.08	0.919	0.123	?	?
cluster_1	0	1	21059	237	0	0	0	1	?	?	?
Overall	?	?	?	?	?	?	?	?	?	0.69	-0.002

Row ID	I cluster_0	I cluster_2	I cluster_1
cluster_0	14256	1267	0
cluster_2	5091	445	1
cluster_1	227	10	0

K=9

Row ID	I TruePositives	I FalsePositives	I TrueNegatives	I FalseNegatives	D Recall	D Precision	D Sensitivity	D Specificity	D F-measure	D Accuracy	D Cohen's kappa
cluster_0	14571	5433	341	952	0.939	0.728	0.939	0.059	0.82	?	?
cluster_2	331	962	14798	5206	0.06	0.256	0.06	0.939	0.097	?	?
cluster_1	0	0	21060	237	0	?	0	1	?	?	?
Overall	?	?	?	?	?	?	?	?	?	0.7	-0.002

Row ID	I cluster_0	I cluster_2	I cluster_1
cluster_0	14571	952	0
cluster_2	5206	331	0
cluster_1	227	10	0

K=11

Row ID	I TruePositives	I FalsePositives	I TrueNegatives	I FalseNegatives	D Recall	D Precision	D Sensitivity	D Specificity	D F-measure	D Accuracy	D Cohen's kappa
cluster_0	14820	5521	253	703	0.955	0.729	0.955	0.044	0.826	?	?
cluster_2	245	711	15049	5292	0.044	0.256	0.044	0.955	0.075	?	?
cluster_1	0	0	21060	237	0	?	0	1	?	?	?
Overall	?	?	?	?	?	?	?	?	?	0.707	-0.002

Row ID	I cluster_0	I cluster_2	I cluster_1
cluster_0	14820	703	0
cluster_2	5292	245	0
cluster_1	229	8	0

K=13

Row ID	I TruePositives	I FalsePositives	I TrueNegatives	I FalseNegatives	D Recall	D Precision	D Sensitivity	D Specificity	D F-measure	D Accuracy	D Cohen's kappa
cluster_0	15038	5602	172	485	0.969	0.729	0.969	0.03	0.832	?	?
cluster_2	167	490	15270	5370	0.03	0.254	0.03	0.969	0.054	?	?
cluster_1	0	0	21060	237	0	?	0	1	?	?	?
Overall	?	?	?	?	?	?	?	?	?	0.714	-0.002

Row ID	I cluster_0	I cluster_2	I cluster_1
cluster_0	15038	485	0
cluster_2	5370	167	0
cluster_1	232	5	0

K=15

Row ID	I TruePositives	I FalsePositives	I TrueNegatives	I FalseNegatives	D Recall	D Precision	D Sensitivity	D Specificity	D F-measure	D Accuracy	D Cohen's kappa
cluster_0	15157	5659	115	366	0.976	0.728	0.976	0.02	0.834	?	?
cluster_2	111	370	15390	5426	0.02	0.231	0.02	0.977	0.037	?	?
cluster_1	0	0	21060	237	0	?	0	1	?	?	?
Overall	?	?	?	?	?	?	?	?	?	0.717	-0.005


Row ID	I cluster_0	I cluster_2	I cluster_1
cluster_0	15157	366	0
cluster_2	5426	111	0
cluster_1	233	4	0

K=17

Row ID	I TruePositives	I FalsePositives	I TrueNegatives	I FalseNegatives	D Recall	D Precision	D Sensitivity	D Specificity	D F-measure	D Accuracy	D Cohen's kappa
cluster_0	15236	5679	95	287	0.982	0.728	0.982	0.016	0.836	?	?
cluster_2	93	289	15471	5444	0.017	0.243	0.017	0.982	0.031	?	?
cluster_1	0	0	21060	237	0	?	0	1	?	?	?
Overall	?	?	?	?	?	?	?	?	?	0.72	-0.003

Row ID	I cluster_0	I cluster_2	I cluster_1
cluster_0	15236	287	0
cluster_2	5444	93	0
cluster_1	235	2	0

K=19

 Dialog - 3:68 - K Nearest Neighbor

File

Standard settings
Flow Variables
Job Manager Selection
Memory Policy

Column with class labels

S cluster

Number of neighbours to consider (k)


19

Weight neighbours by distance

☐

Output class probabilities

☐

OK
Apply
Cancel


Row ID	I TruePositives	I FalsePositives	I TrueNegatives	I FalseNegatives	D Recall	D Precision	D Sensitivity	D Specificity	D F-measure	D Accuracy	D Cohen's kappa
cluster_0	15301	5691	83	222	0.986	0.729	0.986	0.014	0.838	?	?
cluster_2	81	224	15536	5456	0.015	0.266	0.015	0.986	0.028	?	?
cluster_1	0	0	21060	237	0	?	0	1	?	?	?
Overall	?	?	?	?	?	?	?	?	?	0.722	0

Row ID	I cluster_0	I cluster_2	I cluster_1
cluster_0	15301	222	0
cluster_2	5456	81	0
cluster_1	235	2	0

In KNN, the number of neighbours 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\_0**

- True Positives: 15301, False Positives: 5691, True Negatives: 83, False Negatives: 222
- Recall: 0.986, Precision: 0.729, Sensitivity: 0.986, Specificity: 0.014
- F-measure: 0.838, Accuracy: 0.722

In cluster\_0, the KNN model achieved high recall indicating that it effectively identifies true positives within this cluster. However, the precision is relatively low emphasizing a higher 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\_1**

- True Positives: 0, False Positives: 0, True Negatives: 21060, False Negatives: 237
- Recall: 0, Precision: N/A, Sensitivity: 0, Specificity: 1
- F-measure: N/A, Accuracy: N/A

In cluster\_1, the KNN model correctly identifies true negatives but fails to identify any true positives. This results in a recall, precision and F-measure of 0. However, the specificity is 1 indicating that the model effectively identifies instances not belonging to Cluster 1.

### **Cluster\_2**

- True Positives: 81, False Positives: 224, True Negatives: 15536, False Negatives: 5456
- Recall: 0.015, Precision: 0.266, Sensitivity: 0.015, Specificity: 0.986
- F-measure: 0.028, Accuracy: 0.722

In cluster\_2 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

Row ID	I TruePositives	I FalsePositives	I TrueNegatives	I FalseNegatives	D Recall	D Precision	D Sensitivity	D Specificity	D F-measure	D Accuracy	D Cohen's kappa
cluster_1	237	21060	0	0	1	0.011	1	0	0.022	?	?
cluster_0	0	0	5774	15523	0	?	0	1	?	?	?
cluster_2	0	0	15760	5537	0	?	0	1	?	?	?
Overall	?	?	?	?	?	?	?	?	?	0.011	0

Row ID	I cluster_1	I cluster_0	I cluster_2
cluster_1	237	0	0
cluster_0	15523	0	0
cluster_2	5537	0	0

### Cluster\_0

- True Positives: 237, False Positives: 21060, 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: 15760, False Negatives: 5537
- 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: 5774, False Negatives: 15523
- 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.

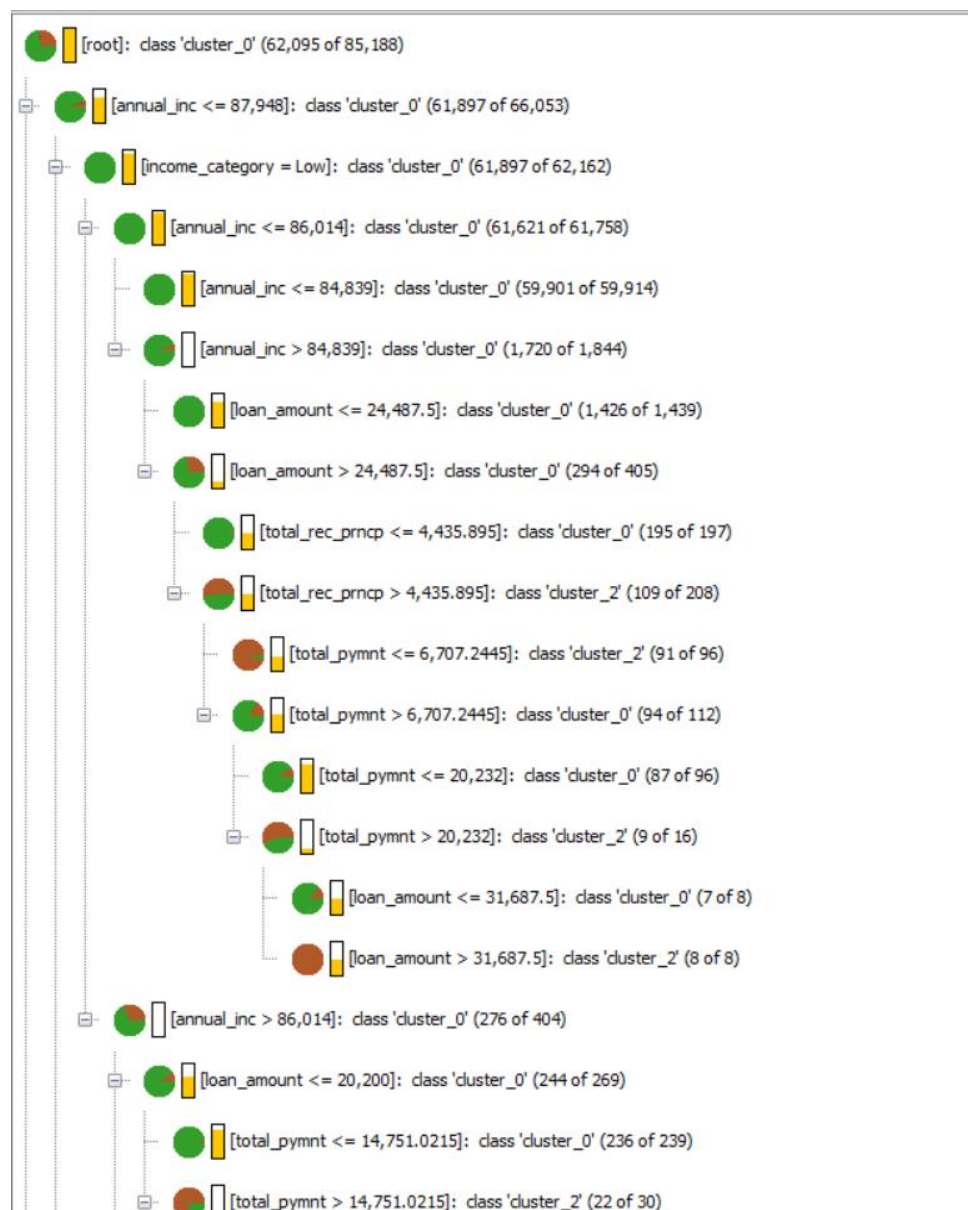


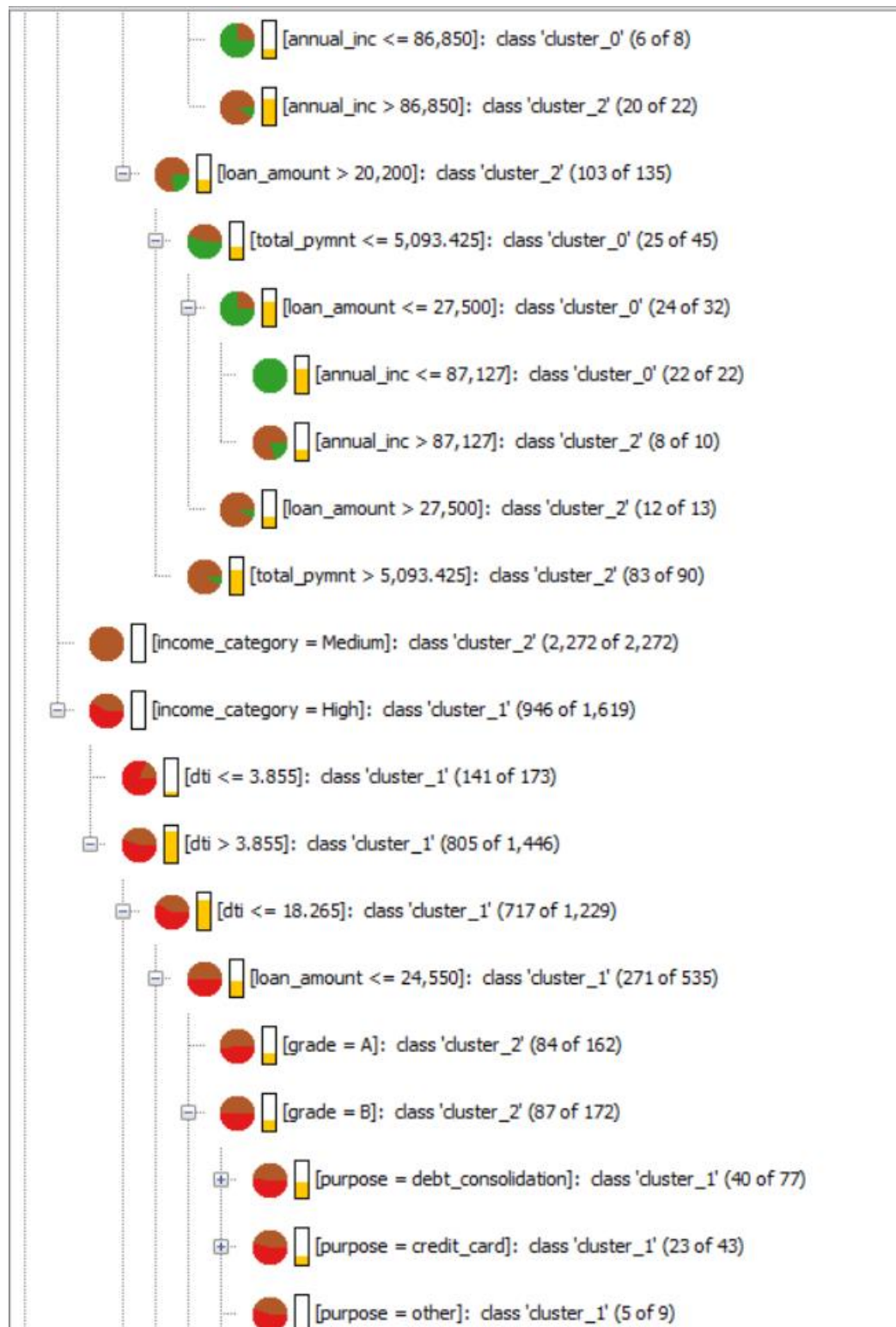
The overall performance of the SVM model is very poor with extremely low recall, precision and accuracy metrics. The absence of true positives in cluster\_2 and cluster\_1 severely impacts the model's ability to provide meaningful insights or make accurate predictions for these clusters.

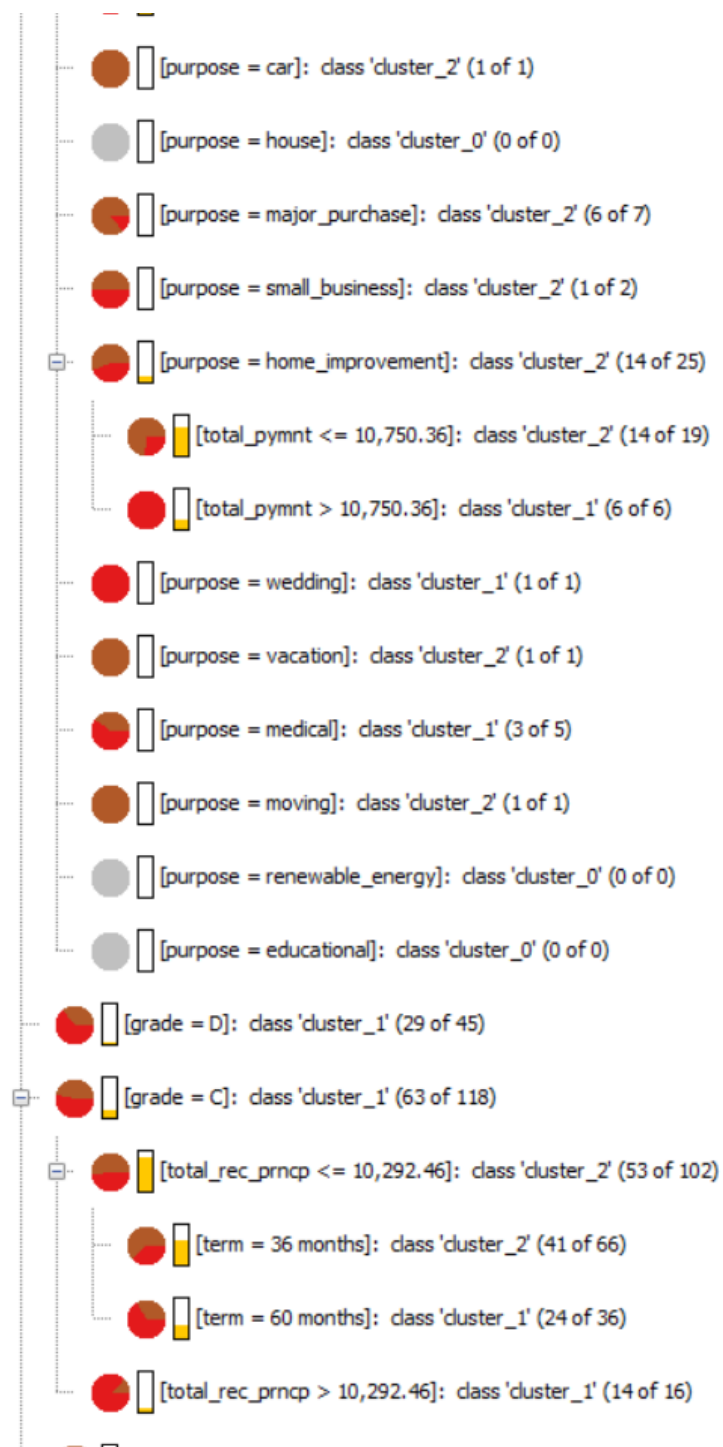
### 3.2.3.1. Variable or Feature Analysis for Decision Tree

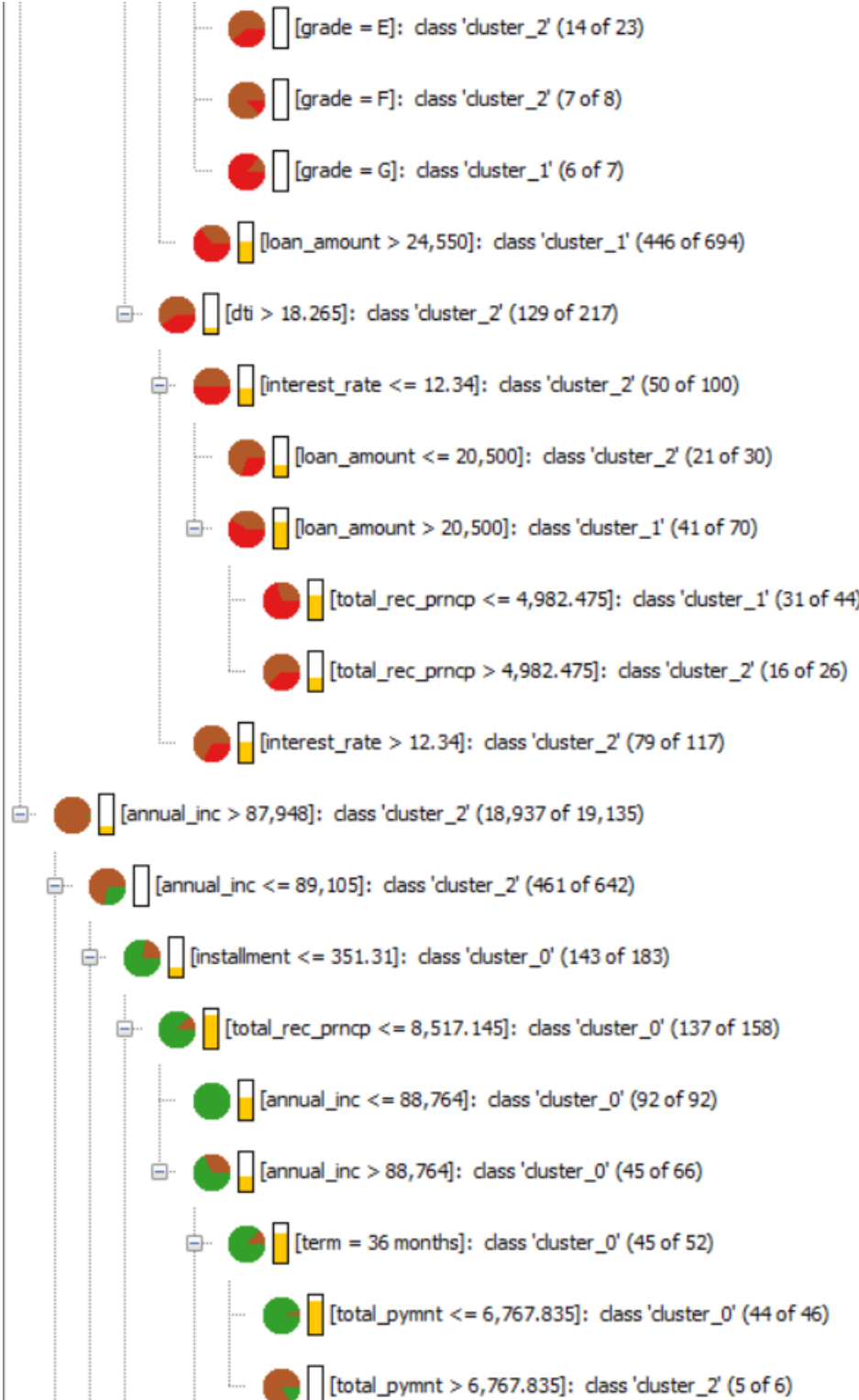
#### 3.2.3.1.1. List of Relevant or Important Variables

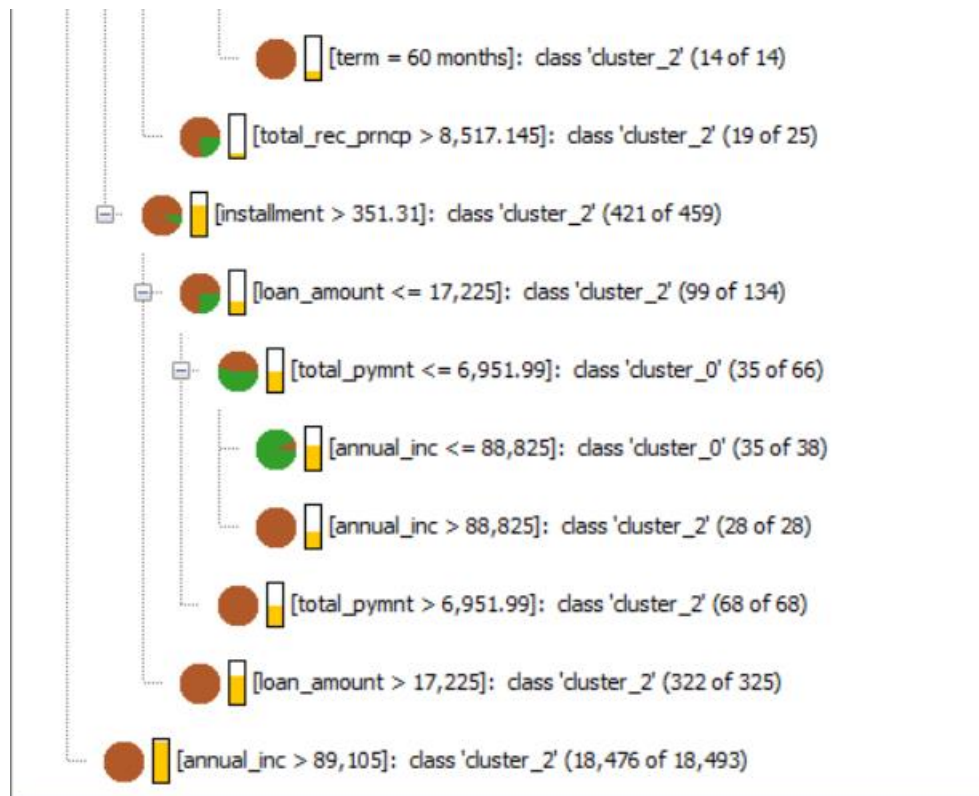
This image describes the variables that were important and contributed in the supervised learning algorithm to predict which cluster the record belonged to as well as the threshold onto which decision were made.











In the decision tree analysis, we see that these were the important variables that contributed in the supervised learning algorithm which are: -

Annual\_inc (annual income), loan\_amount (loan amount), total\_rec\_prncp (total received principal), total\_payment, dti (debt to income ratio), interest\_rate, installment, income\_category, grade, purpose and term.

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

In the decision tree analysis, we see that these were the non-important variables that did not contribute in the supervised learning algorithm which are: -

Recoveries, emp\_length\_int (number of years employee worked), home\_ownership, application\_type, interest\_payments, loan\_condition and region.

### 3.2.3.2. Variable or Feature Analysis for Logistic Regression, K-Nearest Neighbour and Support Vector Machine

#### 3.2.3.2.1. List of Relevant Variables

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-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

The above variables have value of  $p > 0.05$  which suggests potentially negligible impact on loan outcomes.

## 4. Results and Observations

### 4.1. Comparing Supervised Learning models: Decision Tree VS Logistic Regression, KNN and SVM

#### Decision Tree (without pruning)

cluster \ Prediction (cluster)	cluster_0	cluster_2	cluster_1
cluster_0	15499	23	0
cluster_2	21	5434	79
cluster_1	0	97	135

Correct classified: 21,068	Wrong classified: 220
Accuracy: 98.967%	Error: 1.033%
Cohen's kappa ( $\kappa$ ): 0.974%	

#### Decision Tree (with pruning)

cluster \ Prediction (cluster)	cluster_0	cluster_2	cluster_1
cluster_0	15497	26	0
cluster_2	25	5441	71
cluster_1	0	72	165

Correct classified: 21,103	Wrong classified: 194
Accuracy: 99.089%	Error: 0.911%
Cohen's kappa ( $\kappa$ ): 0.977%	

#### Logistic Regression

cluster \ Prediction (cluster)	cluster_0	cluster_2	cluster_1
cluster_0	15492	31	0
cluster_2	32	5412	93
cluster_1	0	50	187

Correct classified: 21,091	Wrong classified: 206
Accuracy: 99.033%	Error: 0.967%
Cohen's kappa ( $\kappa$ ): 0.976%	

## KNN

K=7

cluster \ Class [kNN]	cluster_0	cluster_2	cluster_1
cluster_0	14256	1267	0
cluster_2	5091	445	1
cluster_1	227	10	0

Correct classified: 14,701

Wrong classified: 6,596

Accuracy: 69.029%

Error: 30.971%

Cohen's kappa ( $\kappa$ ): -0.002%

K=9

cluster \ Class [kNN]	cluster_0	cluster_2	cluster_1
cluster_0	14571	952	0
cluster_2	5206	331	0
cluster_1	227	10	0

Correct classified: 14,902

Wrong classified: 6,395

Accuracy: 69.972%

Error: 30.028%

Cohen's kappa ( $\kappa$ ): -0.002%

K=11

cluster \ Class [kNN]	cluster_0	cluster_2	cluster_1
cluster_0	14820	703	0
cluster_2	5292	245	0
cluster_1	229	8	0

Correct classified: 15,065

Wrong classified: 6,232

Accuracy: 70.738%

Error: 29.262%

Cohen's kappa ( $\kappa$ ): -0.002%



K=13

cluster \ Class [kNN]	cluster_0	cluster_2	cluster_1
cluster_0	15038	485	0
cluster_2	5370	167	0
cluster_1	232	5	0

Correct classified: 15,205	Wrong classified: 6,092
Accuracy: 71.395%	Error: 28.605%
Cohen's kappa ( $\kappa$ ): -0.002%	

K=15

cluster \ Class [kNN]	cluster_0	cluster_2	cluster_1
cluster_0	15157	366	0
cluster_2	5426	111	0
cluster_1	233	4	0

Correct classified: 15,268	Wrong classified: 6,029
Accuracy: 71.691%	Error: 28.309%
Cohen's kappa ( $\kappa$ ): -0.005%	

K=17

cluster \ Class [kNN]	cluster_0	cluster_2	cluster_1
cluster_0	15236	287	0
cluster_2	5444	93	0
cluster_1	235	2	0

Correct classified: 15,329	Wrong classified: 5,968
Accuracy: 71.977%	Error: 28.023%
Cohen's kappa ( $\kappa$ ): -0.003%	

K=19

cluster \ Class [kNN]	cluster_0	cluster_2	cluster_1
cluster_0	15301	222	0
cluster_2	5456	81	0
cluster_1	235	2	0

Correct classified: 15,382	Wrong classified: 5,915
Accuracy: 72.226%	Error: 27.774%
Cohen's kappa ( $\kappa$ ): 0%	

## SVM

cluster \ Prediction (cluster)	cluster_1	cluster_0	cluster_2
cluster_1	237	0	0
cluster_0	15523	0	0
cluster_2	5537	0	0

Correct classified: 237	Wrong classified: 21,060
Accuracy: 1.113%	Error: 98.887%
Cohen's kappa ( $\kappa$ ): 0%	

Metrics	Decision Tree (without pruning)	Decision Tree (with pruning)	Logistic Regression	KNN (k=19)	SVM
Accuracy (in %)	98.967	99.089	99.033	72.226	1.113
Error (in %)	1.033	0.911	0.967	27.774	98.887
Cohen's Kappa (in %)	0.974	0.977	0.976	0	0
Correctly classified	21068	21103	21091	15382	237
Wrongly classified	220	194	206	5915	21060

- Decision Tree models (with and without pruning) 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 are recommended for this dataset due to their high accuracy and reliable performance. KNN and SVM are not suitable for this dataset based on the provided results.

#### 4.3. Variable or Feature Analysis

##### Variables important for Decision Tree

###### 1. **Annual\_inc (Annual Income)**

- Annual income is a fundamental measure of an individual's financial status. Customers with higher annual incomes are likely to be classified as more affluent as they have greater financial resources and purchasing power helping to make financial products according to what strata they belong to.

###### 2. **Loan\_amount (Loan Amount)**

- The loan amount indicates the financial commitment undertaken by the customer. Higher loan amounts may suggest larger purchases or investments, which is indicative of greater financial stability.

###### 3. **Total\_rec\_prncp (Total Received Principal)**

- This variable represents the total principal amount received by the customer over the loan term. It reflects the customer's repayment behaviour and their ability to manage debt responsibly. Higher total received principal amounts may indicate a stronger financial position and reliability in meeting loan obligations.

###### 4. **Total\_payment**

- Total payment encompasses all payments made by the customer over the loan period including principal, interest and any fees. Customers who consistently make full payments on time are likely to be classified as more affluent and financially responsible.

###### 5. **Dti (Debt-to-Income Ratio)**

- The debt-to-income ratio measures the proportion of a customer's monthly income that goes towards debt repayment. A lower DTI ratio indicates better financial health and a lower risk of default. Affluent customers typically have lower DTI ratios as they have more disposable income after meeting their financial obligations.

###### 6. **Interest\_rate**

- Interest rates on loans directly impact the cost of borrowing for customers. Lower interest rates are typically offered to borrowers with higher creditworthiness as they present lower risk to lenders.

###### 7. **Installment**

- Installment payments represent the periodic payments made by the customer to repay the loan. The size of the instalment payment may provide insights into the customer's financial capacity and willingness to meet loan obligations.

## 8. Income\_category

- Income category categorizes customers based on their income level (e.g., low, medium, high). Affluent customers are likely to fall into higher income categories indicating greater financial means.

## 9. Grade

- The grade assigned to the customer reflects their creditworthiness and risk profile. Higher grades are typically assigned to customers with better credit histories and financial stability.

## 10. Purpose

- The purpose of the loan can provide insights into the customer's financial goals and priorities. Certain loan purposes such as investment or home improvement may be more indicative of affluence than others and creating certain financial products that may serve different customers for what purposes they want the money to loaned creating an opportunity of different services.

## 11. Term

- The loan term refers to the duration over which the loan is repaid. Longer loan terms may indicate larger loan amounts or lower monthly payments.

### **Variables not important for decision tree analysis**

Recoveries, emp\_length\_int (number of years employee worked), home\_ownership, application\_type, interest\_payments, loan\_condition and region

The above variables or features were not included in the decision tree view which indicates that they had no influence on the classification of customers and these features can be ignored to save time and money when making financial products.

### **Variables important for linear regression**

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

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.

### **Variables not important for linear regression**

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 the p value being greater than 0.05.

## 5. **Managerial Insights**

### 5.1. **Appropriate Model**

Metrics	Decision Tree (without pruning)	Decision Tree (with pruning)	Logistic Regression	KNN (k=19)	SVM
Accuracy (in %)	98.967	99.089	99.033	72.226	1.113

The decision tree with pruning has the highest accuracy (99.089%) followed closely by logistic regression (99.033%). KNN and SVM have significantly lower accuracies of 72.226% and 1.113% respectively

Decision tree provides the highest accuracy of all the models according to the data and will be the appropriate model for the customer classification. Decision tree is able to handle both numerical and categorical which does benefit in this data as the data contains a combination of variables which are categorical and continuous in nature.

### **Managerial insights according to the appropriate model (Decision Tree)**

#### 1. **Risk Assessment**

Decision trees will help to assess the risk profile of customers applying for loans or other financial products. By analyzing customer attributes such as income, income category, purpose of the loan, debt-to-income ratio etc., decision trees can help identify high-risk customers who may default on loans or pose a credit risk to the bank or give benefit to those customers that are paying loans on time.

#### 2. **Customer Segmentation**

Decision trees can aid in segmenting customers based on their financial behaviour, demographics and banking preferences. This segmentation can help banks tailor their marketing strategies, product offerings and customer service initiatives to different customer segments helping to improve overall customer satisfaction and retention.

We had three clusters cluster\_0 (Debt-Ridden customers who are struggling for solvency), cluster\_1 (Affluent Purchasers) and cluster\_2 (Middle-Class Consumers). Accordingly, we can classify or predict according to these clusters and make financial products that will benefit the cluster as well as cut losses to the bank by making specific and specialized products for those clusters.

### **3. Loan Approval Process**

In the loan approval process, decision trees can assist in automating the decision-making process by evaluating customer attributes against predefined criteria. Banks can use decision trees to streamline the loan application process, reduce manual intervention and expedite loan approvals for eligible customers.

### **4. Product Recommendation**

Decision trees can help banks recommend suitable financial products to customers based on their financial goals, preferences and risk tolerance. By analyzing customer attributes and transactional data, decision trees will suggest personalized product recommendations such as investment options, savings accounts or credit cards.

#### **5.2. Relevant or Important Variables or Features**

The relevant or important variables that are used in the decision tree supervised learning algorithm are: -

Annual\_inc (annual income)

Loan\_amount (loan amount)

Total\_rec\_prncp (total received principal)

Total\_payment, dti (debt to income ratio)

Interest\_rate

Installment

Income\_category

Grade

Purpose

term