



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

**Academic Session 2023-2025**

**Classification and Prediction (Risk Assessment Data)  
based on Cluster data**

**Machine Learning for Managers**

**FMG 32 Section B**

**Submitted to  
Prof. Amarnath Mitra**

**Submitted by:  
321105 – Samriddhi Pandey**

# Contents

<b>1. Project Objectives.....</b>	<b>3</b>
<b>2. Description of Data .....</b>	<b>4</b>
<b>3. Analysis of Data.....</b>	<b>9</b>
<b>4. Results   Observations.....</b>	<b>17</b>
<b>5. Managerial Insights .....</b>	<b>18</b>

# 1. Project Objectives

## 1.1. Classification of Consumer Data into Segments

The first objective is to segment the consumer data using supervised learning algorithms (Decision tree). Loan applicant data is segmented into groups based on shared characteristics. This helps assess risk by grouping similar applicants together.

## 1.2. Determination of an Appropriate Classification Model

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

## 1.3. Identification of Significant Variables and their thresholds for Classification

The third objective is to identify significant/contributing variables and their thresholds for classification.

## 2. Description of Data

### 2.1. Data Source, Size, Shape

**2.1.1. Data Source:** <https://www.kaggle.com/datasets/yaminh/applicant-details-for-loan-approve>

**2.1.2. Data Size** (2 MB)

**2.1.3. Data Shape** (Dimension: 14 columns | 1,00,000 rows)

### 2.2. Description of Variables

**2.2.1. Index Variable:** Applicant ID

**2.2.2. Outcome Variable:** Cluster

**2.2.3. Variables or Features having Categories**

#### 2.2.3.1. Categorical Variables or Features (Nominal Type):

- **Occupation:** Profession or occupation of the loan applicant.
- **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.
- **Residence City:** City where the loan applicant resides.
- **Residence State:** State where the loan applicant resides.

#### 2.2.3.2. Categorical Variables or Features (Ordinal Type):

- **Marital Status:** Marital status of the loan applicant.
- **House Ownership:** Ownership status of the applicant's residence.
- **Vehicle Ownership(car):** Ownership status of the applicant's vehicle.
- **Loan Default Risk:** Indicator of loan default risk, with values indicating whether the loan applicant is at risk of defaulting on the loan

#### 2.2.3. Non-Categorical Variables or Features:

- **Annual Income:** Annual income of the loan applicant.
- **Applicant Age:** Age of the loan applicant.
- **Work Experience:** Number of years of work experience of the loan applicant.
- **Years in Current Employment:** Number of years the applicant has been in their current job.
- **Years in Current Residence:** Number of years the applicant has been residing in their current residence.

## 2.3. Descriptive Statistics

### 2.3.1. Descriptive Statistics: Outcome Variable or Feature (Categorical)

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

S Cluster	I Count (Cluster)	D Relative Frequency (...)
cluster_0	33334	0.333
cluster_2	33333	0.333
cluster_1	33333	0.333

### 2.3.2. Descriptive Statistics: Input Categorical Variables or Features

#### Occupation

S Occupation	I Count (Occupation)	D Relative Frequency (Occupation)
Physician	2426	0.024
Statistician	2338	0.023
Fashion_Designer	2189	0.022
Psychologist	2188	0.022
Magistrate	2169	0.022
Computer_hardware_engineer	2169	0.022
Web_designer	2153	0.022
Drafter	2133	0.021
Comedian	2103	0.021
Mechanical_engineer	2097	0.021
Air_traffic_controller	2087	0.021
Chemical_engineer	2087	0.021
Industrial_Engineer	2086	0.021
Financial_Analyst	2079	0.021
Flight_attendant	2073	0.021
Technical_writer	2060	0.021
Graphic_Designer	2059	0.021
Hotel_Manager	2052	0.021
Secretary	2044	0.02
Biomedical_Engineer	2039	0.02
Petroleum_Engineer	2028	0.02
Software_Developer	2016	0.02
Police_officer	1988	0.02
Computer_operator	1966	0.02
Politician	1964	0.02
Microbiologist	1918	0.019
Technician	1916	0.019
Consultant	1911	0.019
Surgeon	1907	0.019
Lawyer	1906	0.019
Artist	1906	0.019
Dentist	1897	0.019
Technology_specialist	1880	0.019
Scientist	1875	0.019
Surveyor	1862	0.019
Army_officer	1859	0.019

Geologist	1855	0.019
Aviator	1853	0.019
Architect	1848	0.018
Design_Engineer	1840	0.018
Analyst	1822	0.018
Civil_engineer	1818	0.018
Chef	1813	0.018
Designer	1801	0.018
Librarian	1795	0.018
Economist	1775	0.018
Firefighter	1765	0.018
Chartered_Accountant	1756	0.018
Civil_servant	1703	0.017
Engineer	1580	0.016
Official	1546	0.015

### Residence State

S Residence_State	I Count (Residence_State)	D Relative Frequency (Residence_State)
Uttar_Pradesh	11255	0.113
Maharashtra	10158	0.102
Andhra_Pradesh	10045	0.1
West_Bengal	9327	0.093
Bihar	7867	0.079
Tamil_Nadu	6595	0.066
Madhya_Pradesh	5587	0.056
Karnataka	4687	0.047
Gujarat	4582	0.046
Jharkhand	3601	0.036
Rajasthan	3589	0.036
Haryana	3075	0.031
Telangana	2929	0.029
Assam	2849	0.028
Kerala	2316	0.023
Delhi	2183	0.022
Punjab	1886	0.019
Odisha	1833	0.018
Chhattisgarh	1513	0.015
Uttarakhand	758	0.008
Jammu_and_Kashmir	721	0.007
Puducherry	566	0.006
Mizoram	340	0.003
Manipur	338	0.003
Himachal_Pradesh	337	0.003
Tripura	312	0.003
Uttar_Pradesh[5]	287	0.003
Chandigarh	255	0.003
Sikkim	209	0.002

### Marital Status

S Marital_Status	I Count...	D Relative Frequency (Marital_Status)
single	89763	0.898
married	10237	0.102

## House Ownership

S House_Ownership	I Count (House_Ownership)	D Relative Frequency (House_Ownership)
rented	92088	0.921
owned	5081	0.051
norent_noown	2831	0.028

## Vehicle Ownership(car)

S Vehicle_Ownership(car)	I Count (Vehicle_Ownership(car))	D Relative Frequency (Vehicle_Ownership(car))
no	69665	0.697
yes	30335	0.303

## Loan Default Risk

I Loan_Default_Risk	I Count (Loan_Default_Risk)	D Relative Frequency (Loan_Default_Risk)
0	87003	0.87
1	12997	0.13

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

Name	Type	# Missing values	# Unique values	Mean	Mean Absolute Deviation ↑
Years_in_Current_Residence	Number (integer)	0	5	11.996	1.18
Years_in_Current_Employment	Number (integer)	0	15	6.343	3.058
Work_Experience	Number (integer)	0	21	10.111	5.185
Applicant_Age	Number (integer)	0	59	49.995	14.781
Annual_Income	Number (integer)	0	5999	5,001,617.026	2,494,478.873






















Row ID	n	D Min	D Max	D Mean	D Std. de...	D Variance	D Skewn...	D Kurtosis	D Overall...	I No. mis...	I No. NaNs	I No. +∞s	I No. -∞s	D Median	I Row C...	Histogram
Applicant_ID	1	100,000	50,000.5	28,867.658	833,341,666.667	0	-1.2	5,000,050,000	0	0	0	0	?	100000	1	
Annual_Income	...	10,310	9,999,180	5,001,617.026	2,876,393.521	8,273,639,688,...	0.003	-1.197	500,161,702,...	0	0	0	?	100000	10,310	
Applicant_Age	ge 21	79	49.995	17.056	290.909	-0.008	-1.204	4,999,540	0	0	0	0	?	100000	21	
Work_Experience	...	0	20	10.111	5.996	35.952	-0.018	-1.195	1,011,075	0	0	0	?	100000	0	
Years_in_Curre...	...	0	14	6.343	3.645	13.286	0.275	-0.787	634,299	0	0	0	?	100000	0	
Years_in_Curre...	...	10	14	11.996	1.397	1.951	0.009	-1.271	1,199,602	0	0	0	?	100000	10	

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

S	First column name	S	Second column name	D	Correlation value	D	p value	I	Degrees of freedom
	Annual_Income		Applicant_Age		-0.00102507706105...		0.74582155...		99998
	Annual_Income		Work_Experience		0.00723317873968404		0.02217637...		99998
	Annual_Income		Years_in_Current_Emplo...		0.009235446338099...		0.00349436...		99998
	Annual_Income		Years_in_Current_Resid...		-0.00354887204131...		0.26175982...		99998
	Applicant_Age		Work_Experience		2.771118672768097...		0.93017125...		99998
	Applicant_Age		Years_in_Current_Emplo...		0.004642890073130...		0.14204981...		99998
	Applicant_Age		Years_in_Current_Resid...		-0.02612725186252...		1.41498103...		99998
	Work_Experience		Years_in_Current_Emplo...		0.6415113862467119		0.0		99998
	Work_Experience		Years_in_Current_Resid...		0.022818389287550...		5.32907051...		99998
	Years_in_Current_E...		Years_in_Current_Resid...		0.005722527153471...		0.07035564...		99998

The variables are correlated if the value of p is less than 0.05.

Row ID	D	Annual_Income	D	Applicant_Age	D	Work_Experience	D	Residence_City	D	Years_in_Current_Employment	D	Years_in_Current_Residence
Annual_Income		1.0		-0.001025077061...		0.00723317873968...		?		0.009235446338099971		-0.003548872041317275
Applicant_Age		-0.0010250770610...		1.0		2.77111867276809...		?		0.004642890073130738		-0.026127251862525777
Work_Experience		0.0072331787396...		2.7711186727680...		1.0		?		0.6415113862467119		0.022818389287550544
Residence_City		?		?		?		1.0		?		?
Years_in_Current_Em...		0.0092354463380...		0.0046428900731...		0.6415113862467119		?		1.0		0.005722527153471957
Years_in_Current_Re...		-0.0035488720413...		-0.026127251862...		0.02281838928755...		?		0.005722527153471957		1.0

 corr = -1	Annual_Income	Applicant_Age	Work_Experience	Residence_City	Years_in_Cu...	Years_in_Cu...
 corr = +1						
 corr = n/a						
Annual_Income						
Applicant_Age						
Work_Experience						
Residence_City						
Years_in_Current...						
Years_in_Current...						



## 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 or Features

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

- **Marital Status**

Single-0, Married-1

- **House Ownership**

Norent\_noown- 2, Rented-1, Owned-0

- **Vehicle Ownership(car)**

Yes-1, No-0

- **Residence State**

Punjab-0, West Bengal-1, Madhya Pradesh-2, Andhra Pradesh-3, Assam-4, Bihar-5, Rajasthan-6, Yttar Pradesh-7, Chandigarh-8, Karnataka-9, Delhi-10, Haryana-11, Gujarat-12, Maharashtra-13, Chhattisgarh-14, Kerala-15, Tripura-16, Tamil Nadu-17, Sikkim-18, Mizoram-19, Jharkhand-20, Uttar Pradesh-21, Odisha-22, Telangana-23, J&K-24, Himanchal Pradesh-25, Uttarakhand-26, Punducherry-27, Manipur-28

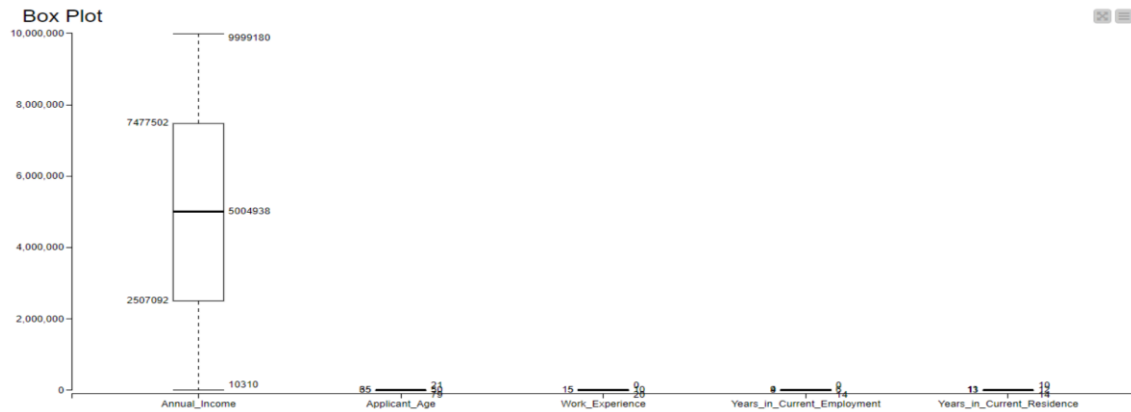
#### 3.1.3. Outlier Statistics and Treatment

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

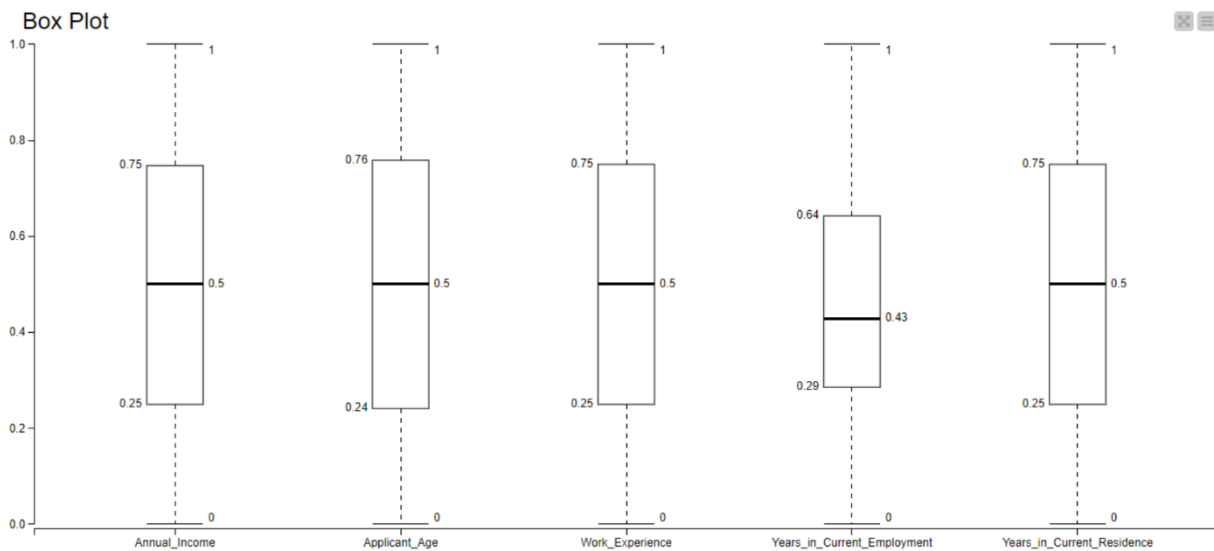
S Outlier column	I Membe...	I Outlier count	D Lower bound	D Upper bound
Annual_Income	100000	0	-4,948,523	14,933,117
Applicant_Age	100000	0	-10	110
Work_Experience	100000	0	-10	30
Years_in_Current_Employment	100000	0	-3.5	16.5
Years_in_Current_Residence	100000	0	8	16

### 3.1.3.1.2. Normalization using Min-Max Scaler

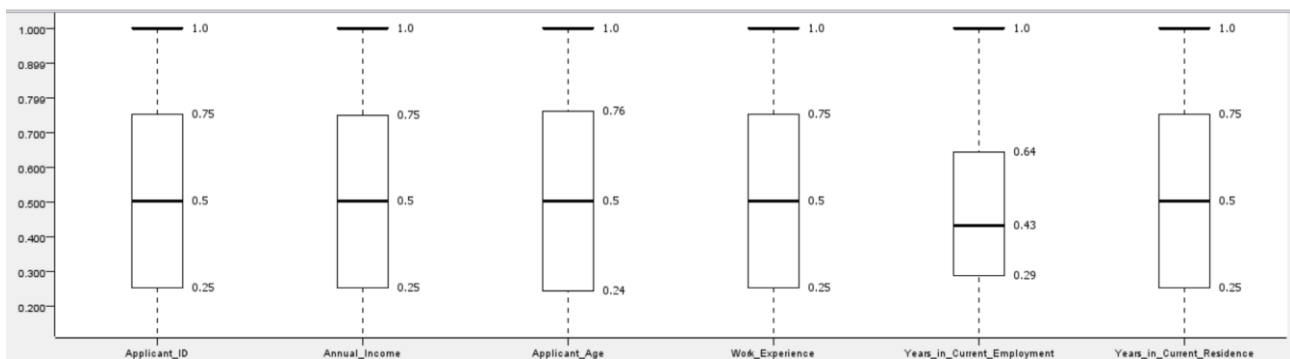
#### Before Normalization



#### After Normalization



Using numeric outliers' node to remove the outliers.



## De-normalizing the data



### 3.1.4. Data Bifurcation

The bifurcation schema used is stratified sampling based on outcome variable cluster variable with 70% (training data) and 30% (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 Algorithms: 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 Neighbors**

K-Nearest Neighbors (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 neighbors 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.

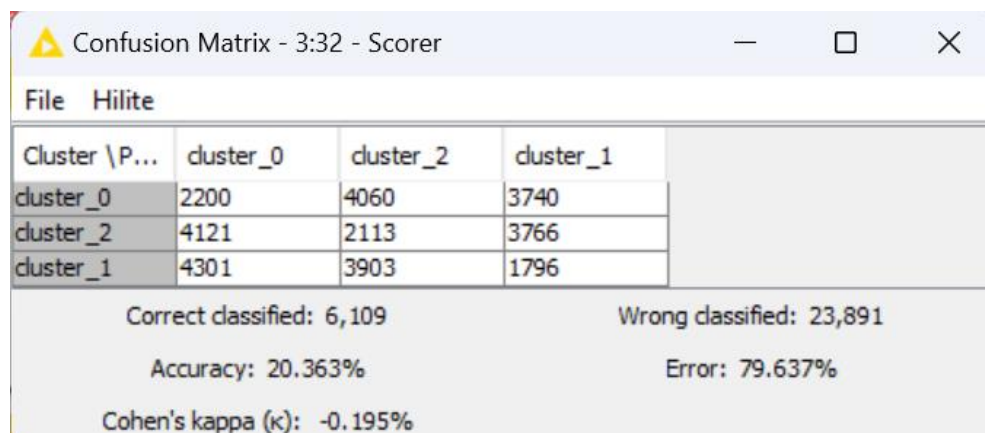
For comparison, we will be using k = 7 till k=11 in steps of 2 i.e. k=7,9,11

### **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: Confusion Matrix (Decision Tree)**



Cluster \ P...	cluster_0	cluster_2	cluster_1
cluster_0	2200	4060	3740
cluster_2	4121	2113	3766
cluster_1	4301	3903	1796

Correct classified: 6,109      Wrong classified: 23,891

Accuracy: 20.363%      Error: 79.637%

Cohen's kappa ( $\kappa$ ): -0.195%

### 3.2.2.2. Classification Model Performance Evaluation: Confusion matrix for other methods

#### Logistic Regression

Prediction ...	cluster_0	cluster_2	cluster_1
cluster_0	10000	0	0
cluster_2	0	10000	0
cluster_1	0	0	10000

Correct classified: 30,000      Wrong classified: 0

Accuracy: 100%      Error: 0%

Cohen's kappa ( $\kappa$ ): 1%

Row ID	I TruePo...	I FalsePo...	I TrueNe...	I FalseN...	D Recall	D Precision	D Sensitivity	D Specificity	D F-meas...	D Accuracy	D Cohen'...
cluster_0	10000	0	20000	0	1	1	1	1	1	?	?
cluster_2	10000	0	20000	0	1	1	1	1	1	?	?
cluster_1	10000	0	20000	0	1	1	1	1	1	?	?
Overall	?	?	?	?	?	?	?	?	?	1	1

The overall accuracy of the logistic regression model is very high at 100 and it effectively predicts the cluster labels for most instances. Additionally, the Cohen's Kappa coefficient suggests substantial agreement beyond chance among the predicted and actual cluster labels.

#### K-Nearest Neighbor

**K=7**

Confusion Matrix - 3:22 - Scorer			
Cluster \ Cl...	cluster_0	cluster_2	cluster_1
cluster_0	9924	54	22
cluster_2	18	9010	972
cluster_1	14	1017	8969


Correct classified: 27,903      Wrong classified: 2,097

Accuracy: 93.01%      Error: 6.99%

Cohen's kappa ( $\kappa$ ): 0.895%

Row ID	I TruePo...	I FalsePo...	I TrueNe...	I FalseN...	D Recall	D Precision	D Sensitivity	D Specificity	D F-meas...	D Accuracy	D Cohen'...
cluster_0	9924	32	19968	76	0.992	0.997	0.992	0.998	0.995	?	?
cluster_2	9010	1071	18929	990	0.901	0.894	0.901	0.946	0.897	?	?
cluster_1	8969	994	19006	1031	0.897	0.9	0.897	0.95	0.899	?	?
Overall	?	?	?	?	?	?	?	?	?	0.93	0.895

K=9



Confusion Matrix - 3:23 - Scorer

File

Hilite

Class [kNN...	cluster_0	cluster_2	cluster_1
cluster_0	9927	61	21
cluster_2	42	9257	704
cluster_1	31	682	9275

Correct classified: 28,459

Wrong classified: 1,541


Accuracy: 94.863%

Error: 5.137%

Cohen's kappa ( $\kappa$ ): 0.923%

Row ID	I TruePo...	I FalsePo...	I TrueNe...	I FalseN...	D Recall	D Precision	D Sensitivity	D Specificity	D F-meas...	D Accuracy	D Cohen'...
cluster_0	9927	73	19918	82	0.992	0.993	0.992	0.996	0.992	?	?
cluster_2	9257	743	19254	746	0.925	0.926	0.925	0.963	0.926	?	?
cluster_1	9275	725	19287	713	0.929	0.927	0.929	0.964	0.928	?	?
Overall	?	?	?	?	?	?	?	?	?	0.949	0.923

K=11



Confusion Matrix - 3:24 - Scorer

File

Hilite

Cluster \ Cl...	cluster_0	cluster_2	cluster_1
cluster_0	9945	40	15
cluster_2	38	9642	320
cluster_1	27	353	9620

Correct classified: 29,207

Wrong classified: 793


Accuracy: 97.357%

Error: 2.643%

Cohen's kappa ( $\kappa$ ): 0.96%

Row ID	I TruePo...	I FalsePo...	I TrueNe...	I FalseN...	D Recall	D Precision	D Sensitivity	D Specificity	D F-meas...	D Accuracy	D Cohen'...
cluster_0	9945	65	19935	55	0.995	0.994	0.995	0.997	0.994	?	?
cluster_2	9642	393	19607	358	0.964	0.961	0.964	0.98	0.963	?	?
cluster_1	9620	335	19665	380	0.962	0.966	0.962	0.983	0.964	?	?
Overall	?	?	?	?	?	?	?	?	?	0.974	0.96

## Support Vector Machines



Confusion Matrix - 3:39 - Scorer

File

Hilite

Cluster \ P...	cluster_0	cluster_2	cluster_1
cluster_0	10000	0	0
cluster_2	10000	0	0
cluster_1	10000	0	0

Correct classified: 10,000

Wrong classified: 20,000

Accuracy: 33.333%

Error: 66.667%

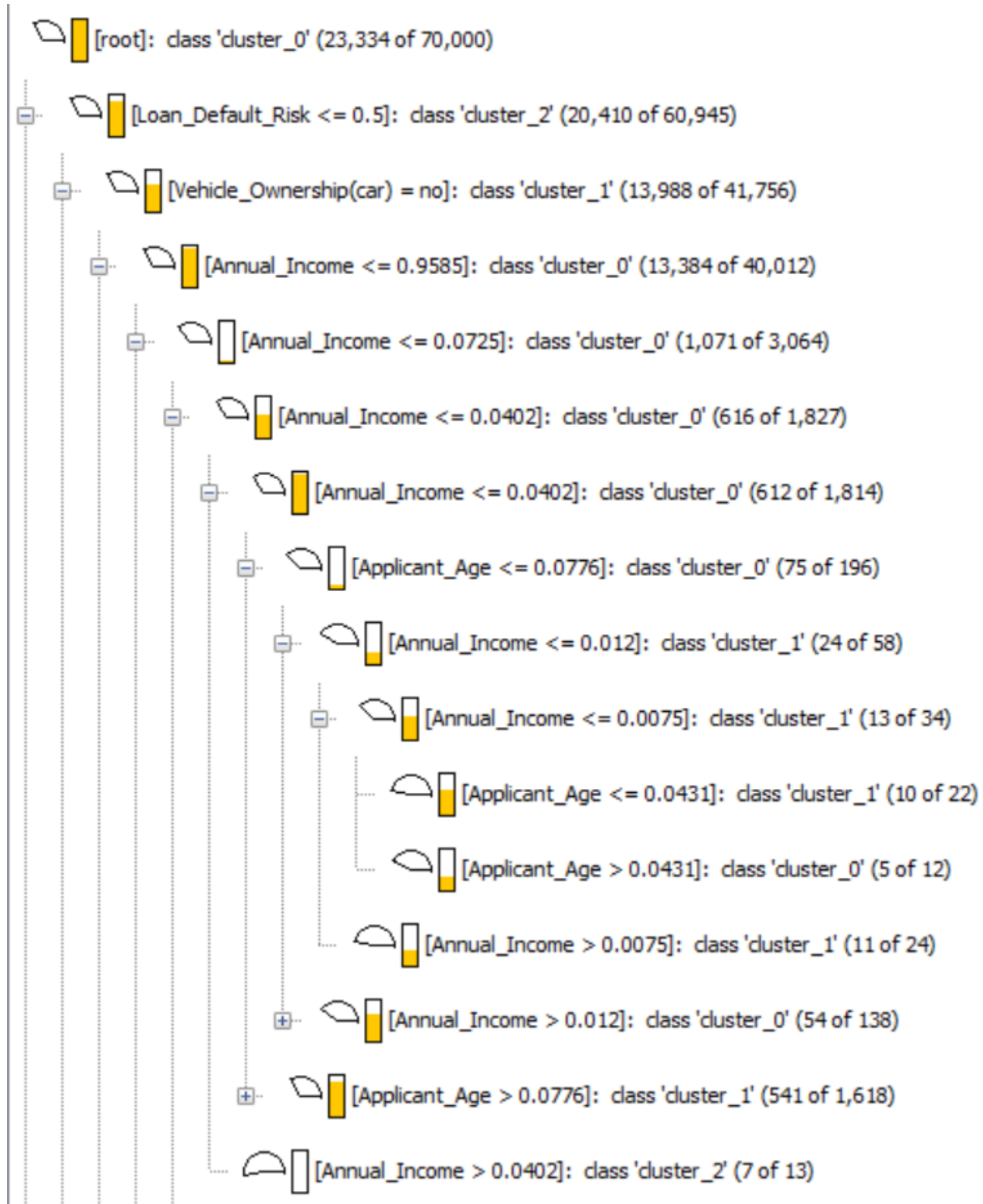
Cohen's kappa ( $\kappa$ ): 0%

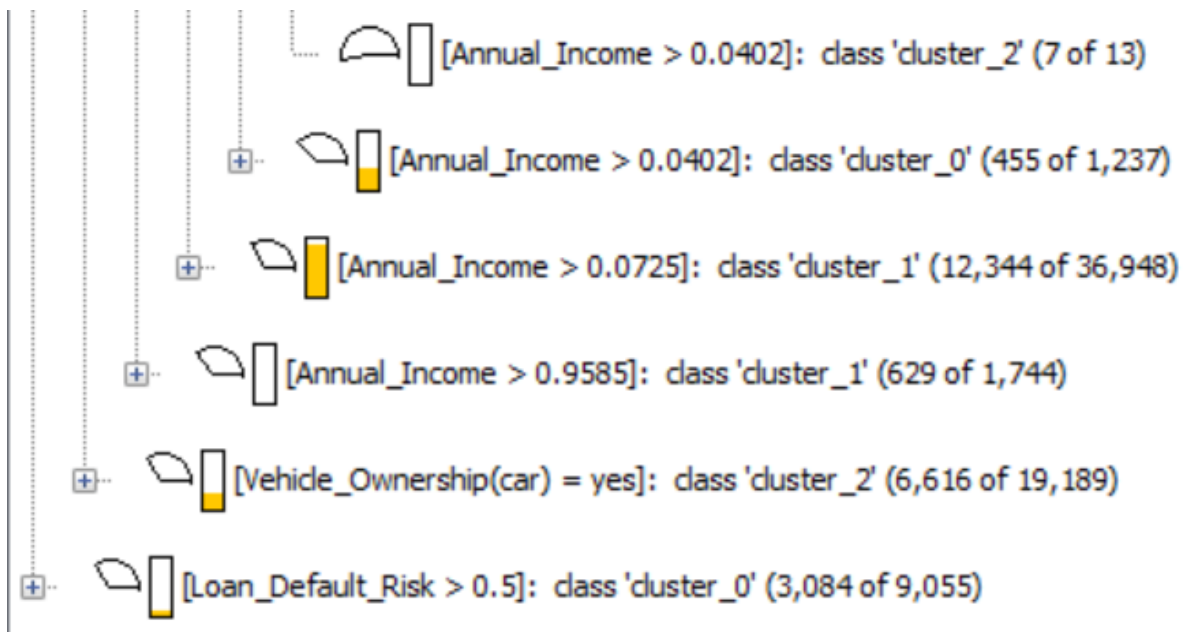
Row ID	I TruePo...	I FalsePo...	I TrueNe...	I FalseN...	D Recall	D Precision	D Sensitivity	D Specificity	D F-meas...	D Accuracy	D Cohen'...
cluster_0	10000	20000	0	0	1	0.333	1	0	0.5	?	?
cluster_2	0	0	20000	10000	0	?	0	1	?	?	?
cluster_1	0	0	20000	10000	0	?	0	1	?	?	?
Overall	?	?	?	?	?	?	?	?	?	0.333	0

### 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.





#### 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: -

Marital Status , House ownership, Work Experience, Residence City, Residence State etc.



## 4. Results | Observations

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

Metrics	Decision Tree	Logistic Regression	KNN (k=11)	SVM
Accuracy (in %)	20.363	100	97.357	33.333
Error (in %)	79.637	0	2.643	66.666
Cohen's Kappa (in %)	-0.195	1	0.96	0
Correctly Classified	6109	30000	29207	10000
Wrongly Classified	23891	0	793	20000

- KNN(k=11) and logistic regression demonstrate high accuracy and Cohen's Kappa values which shows a robust performance in classification.
- SVM shows a lower accuracy and Cohen's Kappa values compared to other models suggesting its limitations in handling this particular dataset effectively.
- Decision Tree demonstrates extremely poor performance with an accuracy of just over 20% and a negative Cohen's Kappa indicating a failure to effectively classify instances in this dataset.
- Overall, KNN (K=11) and logistic regression are recommended for this dataset due to their high accuracy and reliable performance. Decision Tree and SVM are not suitable for this dataset based on the provided results.

## 5. Managerial Insights

### 5.1. Appropriate Model: Compare and Contrast

Metrics	Decision Tree	Logistic Regression	KNN (k=11)	SVM
Accuracy (in %)	20.363	100	97.357	33.333

The Logistic Regression has the highest accuracy (100%) followed closely by KNN (97.357%). Decision Tree and SVM have significantly lower accuracies of 20.363% and 33.333% respectively.

Logistic Regression provides the highest accuracy of all the models according to the data and will be the appropriate model for the customer classification. It is ideal for understanding the impact of several independent variables.

### Managerial Insights

- **Risk Patterns:**

**Segmentation:** Use the models to segment applicants into risk categories. This can help in tailoring loan offers and interest rates.

**Anomalies:** Identify any anomalies or outliers in the predictions that may indicate data issues or exceptional cases.

- **Policy Implications:**

**Credit Policies:** Adjust credit policies based on the insights from the models. For instance, if employment history is a strong predictor, consider revising the employment requirements for loan eligibility.

**Risk Mitigation:** Develop strategies to mitigate risks identified by the models, such as requiring collateral or co-signers for high-risk applicants.

### 5.2. Relevant or Important Variables or Features

- **Loan Default Risk**
- **Annual Income**
- **Applicant Age**