|  |  |
| --- | --- |
| **Student Name:** | Xihai Ren |
| **Student No.:** | 041127486 |
| **Due Date:** | June 21, 2024 |
| **Course:** | CST8390 BI |
| **Professor:** | Grace Pauly |

**CST8390 Assignment 2 Report**

**Table of Contents**

[1 Introduction 3](#_Toc170457893)

[2 Business Understanding 3](#_Toc170457894)

[3 Data Understanding 3](#_Toc170457895)

[3.1 Collect Initial Data 3](#_Toc170457896)

[3.2 Describe Data 3](#_Toc170457897)

[3.3 Explore Data 4](#_Toc170457898)

[3.4 Verify Data Quality 7](#_Toc170457899)

[4 Data Preparation 8](#_Toc170457900)

[4.1 Select Data 8](#_Toc170457901)

[4.2 Clean Data 8](#_Toc170457902)

[4.3 Construct Data 8](#_Toc170457903)

[4.4 Integrate Data 8](#_Toc170457904)

[4.5 Format Data 8](#_Toc170457905)

[5 Modeling 9](#_Toc170457906)

[5.1 Decision Tree 9](#_Toc170457907)

[5.2 Cluster 12](#_Toc170457908)

[5.3 Outlier Detect 14](#_Toc170457909)

[6 Conclusion 15](#_Toc170457910)

[7 Reference 16](#_Toc170457911)

Tables

[Table 1: Dataset Characteristics 3](#_Toc170459645)

[Table 2 Attributes Table 4](#_Toc170459646)

[Table 3 Attributes and Instances statistic 4](#_Toc170459647)

[Table 4 Attributes Data Distribution 5](#_Toc170459648)

[Table 5: Missing values 7](#_Toc170459649)

[Table 6: Binning Age Comparison 8](#_Toc170459650)

[Table 7: Decision tree split data performance 11](#_Toc170459651)

Figures

[Figure 1: Relevant between BIL and Category 5](#_Toc170459662)

[Figure 2: Relevant between PROT and Category 6](#_Toc170459663)

[Figure 3: Relevant between ALB and Category 7](#_Toc170459664)

[Figure 4: Detect Outlier (Distance) 8](#_Toc170459665)

[Figure 5: Decision tree cross-validation 9](#_Toc170459666)

[Figure 6: Decision tree cross validation performance 10](#_Toc170459667)

[Figure 7: Decision tree split data 11](#_Toc170459668)

[Figure 8: Cluster model 12](#_Toc170459669)

[Figure 9: Centroid Table 13](#_Toc170459670)

[Figure 10: Clustering performance 13](#_Toc170459671)

[Figure 11: Outlier detection pivot table 14](#_Toc170459672)

[Figure 12: LOF vs. Distance Outlier detection comparison 15](#_Toc170459673)

# Introduction

In this report, we explore and analyze a dataset related to HCV. The dataset includes laboratory values of blood donors and Hepatitis C patients, along with demographic information like age and sex. The goal of this analysis is to perform various machine learning tasks including classification, clustering, and outlier detection using RapidMiner.

The key objectives of this analysis are:

1. Classification: To categorize the patients into different stages of Hepatitis C progression using a Decision Tree algorithm.
2. Clustering: To group similar instances together using the k-Means algorithm, thereby identifying underlying patterns in the data.
3. Outlier Detection: To identify anomalies in the dataset using Local Outlier Factor (LOF) and distance-based methods.

These machine learning techniques help in understanding the structure of the dataset, identifying significant attributes, which can be valuable for medical diagnosis and treatment planning.

# Business Understanding

The goal of this report is to explore and analyze HCV dataset and perform

classification using Decision Trees, clustering using k-Means and outlier detection using LOF and

Distances approaches in RapidMiner.

# Data Understanding

## Collect Initial Data

This dataset was downloaded from the UCI ML Repository: https://archive.ics.uci.edu/dataset/571/hcv+data

## Describe Data

* The dataset contains laboratory values of blood donors and Hepatitis C patients and demographic values like age.
* Instances are patients
* The target attribute for classification is Category (blood donors vs. Hepatitis C, including its progress: 'just' Hepatitis C, Fibrosis, Cirrhosis).

|  |  |
| --- | --- |
| **Dataset Characteristics** | Multivariate |
| **Subject Area** | Health and Medicine |
| **Associated Tasks** | Classification, Clustering |
| **Feature Type** | Integer, Real |
| **Instances** | 615 |
| **Features** | 12 |

Table 1: Dataset Characteristics

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable Name | Role | Type | Demographic | Description | Units | Missing Values |
| ID | ID | Integer |  | Patient ID |  | no |
| Age | Feature | Integer | Age |  | years | no |
| Sex | Feature | Binary | Sex | (f,m) |  | no |
| ALB | Feature | Continuous |  | Albumin |  | yes |
| ALP | Feature | Continuous |  | Alkaline Phosphatase |  | yes |
| AST | Feature | Continuous |  | Aspartate Aminotransferase |  | yes |
| BIL | Feature | Continuous |  | Bilirubin |  | no |
| CHE | Feature | Continuous |  | Cholinesterase |  | no |
| CHOL | Feature | Continuous |  | Cholesterol |  | yes |
| CREA | Feature | Continuous |  | Creatinine |  | no |
| CGT | Feature | Continuous |  | Gamma-glutamyl Transferase |  | no |
| PROT | Feature | Continuous |  | Total Protein |  | yes |
| Category | Target | Categorical |  | values: '0=Blood Donor', '0s=suspect Blood Donor', '1=Hepatitis', '2=Fibrosis', '3=Cirrhosis' |  | no |
| ALT | Feature | Continuous |  |  |  | no |

Table 2 Attributes Table

## Explore Data

**Attribute:**

|  |  |
| --- | --- |
| **Attributes** | 14 |
| **Instances** | 615 |

Table 3 Attributes and Instances statistic

|  |  |  |
| --- | --- | --- |
| Attribute | Missing | Distribution |
| Patient ID | 0 | Patient ID is sequence No. |
| Category | 0 | Most of the data is concentrated in the 0=Blood Donor category. |
| Age | 0 | The age distribution is relatively even, with more samples around the age of 50. |
| Sex | 0 | There are more males than females. |
| ALB | 1 | ALB is concentrated between 40 and 50. |
| ALP | 18 | ALP is concentrated between 0 and 100, with most values between 51.83 and 92.36. |
| ALT | 1 | ALT is concentrated between 0.9 and 33.34. |
| AST | 0 | AST is concentrated between 10.6 and 41.94 |
| BIL | 0 | BIL is concentrated between 0.8 and 26.12 |
| CHE | 0 | Normal Distribution, the peak is between 7.416 and 8.915 |
| CHOL | 10 | Normal Distribution, the peak is between 4.726 and 5.55 |
| CREA | 0 | CREA is concentrated between 8 and 115.1 |
| GGT | 0 | GGT is concentrated between 4.5 and 69.14 |
| PROT | 1 | PROT is concentrated between 67.4 and 76.44 |

Table 4 Attributes Data Distribution

**Relevant:**

A screenshot of a graph

Description automatically generated

Figure 1: Relevant between BIL and Category

There is a significant correlation between BIL values and cirrhosis.

A graph of a diagram

Description automatically generated with medium confidence

Figure 2: Relevant between PROT and Category

PROT has a higher correlation with Suspect Blood Donor (0s) in the lower data range.

A graph with blue lines

Description automatically generated

Figure 3: Relevant between ALB and Category

ALB has a higher correlation with 0s in the range of 14.9 to 24.9, and a higher correlation with cirrhosis in the range of 29 to 36.

## Verify Data Quality

**Missing value:**

|  |  |
| --- | --- |
| Attribute | Missing |
| ALB | 1 |
| ALP | 18 |
| ALT | 1 |
| CHOL | 10 |
| PROT | 1 |

Table 5: Missing values

**Duplicate value:**

There are no duplicate values

Outlier Detect:

**Outlier:**

A screenshot of a computer

Description automatically generated

Figure 4: Detect Outlier (Distance)

Using Detect Outlier (Distances) to detect outlier, there are 10 instances be categorized as outlier.

# Data Preparation

## Select Data

* Choose the HCV dataset, including all the instances.
* Select relevant attribute: Age, Sex, ALB, ALP, ALT, AST, BIL, CHE, CHOL, CREA, GGT, PROT
* Remove irrelevant attribute: Patient ID

## Clean Data

* Replace missing values with the mean value.
* Normalize data. I use Z-transformation
* Binning Age: I choose by frequency with 6 bins.

|  |  |  |  |
| --- | --- | --- | --- |
| By Size | By Frequency | By Specification | By Binning |
|  |  |  |  |

Table 6: Binning Age Comparison

## Construct Data

* This dataset is well-constructed, this step is omitted.

## Integrate Data

* This dataset is the unique data source, this step is omitted.

## Format Data

* Change the attribute **“Sex”** from Nominal to Numeric using one hot encoding.

# Modeling

## Decision Tree

**Cross-Validation**

**A computer screen shot of a diagram

Description automatically generated**

Figure 5: Decision tree cross-validation

**If PROT > -4.195, further split based on CHE.**

**if CHE > -2.005, further split based on ALB**

**if ALB > -2.990, further split based on ALT.**

**if ALT > 6.207, the classification is 0s=suspect Blood Donor**

**if ALT ≤ 6.207, further split based on ALT.**

**if ALT > -1.026, further split based on AST.**

**if AST > 0.552, further split based on GGT.**

**if GGT > -0.266, further split based on ALP.**

**if ALP > -1.737, further split based on BIL.**

**if BIL > -0.279, the classification is 2=Fibrosis**

**if BIL ≤ -0.279, the classification is 0=Blood Donor**

**if ALP ≤ -1.737, the classification is 1=Hepatitis**

**GGT ≤ -0.266, the classification is 0=Blood Donor**

**if AST ≤ 0.552, further split based on ALT.**

**if ALT > -0.835, further split based on ALP.**

**if ALP > -1.563, the classification is 0=Blood Donor**

**if ALP ≤ -1.563, the classification is 1=Hepatitis**

**if ALT ≤ -0.835, further split based on ALB.**

**ALB > 0.516, the classification is 0=Blood Donor**

**ALB ≤ 0.516, the classification is 2=Fibrosis**

**if ALT ≤ -1.026, the classification is 3=Cirrhosis**

**if ALB ≤ -2.990, the classification is 0s=suspect Blood Donor**

**if CHE ≤ -2.005, the classification is 3=Cirrhosis**

**If PROT ≤ -4.195, the classification is 0s=suspect Blood Donor**

**A table with numbers and percentages

Description automatically generated**

Figure 6: Decision tree cross validation performance

**Accuracy: 89.76**

**Precision:**

* **Blood Donor: 96.12%**
* **Suspect Blood Donor: 50.00%**
* **Hepatitis: 0.00%**
* **Fibrosis: 29.03%**
* **Cirrhosis: 73.08%**

**Recall:**

* **Blood Donor: 97.56%**
* **Suspect Blood Donor: 57.14%**
* **Hepatitis: 0.00%**
* **Fibrosis: 42.86%**
* **Cirrhosis: 63.33%**

**High Accuracy for Blood Donor (96.12%):** The model performs well in identifying Blood Donors.

**Poor Performance for Hepatitis (0.00%):** The model fails to identify Hepatitis cases.

**Moderate Performance for Cirrhosis (73.08%):** The model moderately identifies Cirrhosis cases.

**Improvement Needed for Fibrosis (29.03%) and Suspect Blood Donor (50.00%):** These classes need better precision and recall.

**Split Data:**

**A diagram of a dna

Description automatically generated with medium confidence**

Figure 7: Decision tree split data

**If ALB > -3.267, further split based on CHE.**

**if CHE > -1.993, further split based on ALT.**

**if ALT > -1.026, further split based on ASF.**

**if AST > 1.330, further split based on ALB.**

**if ALB > 0.845, the classification is 1=Hepatitis**

**if ALB ≤ 0.845, the classification is 2=Fibrosis**

**if AST ≤ 1.330, further split based on AST.**

**if GGT > 2.128, the classification is 1=Hepatitis**

**if GGT ≤ 2.128, the classification is 0=Blood Donor**

**if ALT ≤ -1.026, the classification is 3=Cirrhosis**

**if CHE ≤ -1.993, the classification is 3=Cirrhosis**

**If ALB ≤ -3.267, the classification is 0s=suspect Blood Donor**

A screenshot of a computer

Description automatically generated

Table 7: Decision tree split data performance

**Accuracy: 89.67**

**Precision:**

* **Blood Donor: 94.01%**
* **Suspect Blood Donor: 0.00%**
* **Hepatitis: 0.00%**
* **Fibrosis: 0.00%**
* **Cirrhosis: 88.89%**

**Recall:**

* **Blood Donor: 98.12%**
* **Suspect Blood Donor: 0.00%**
* **Hepatitis: 0.00%**
* **Fibrosis: 0.00%**
* **Cirrhosis: 88.89%**

**High Accuracy for Blood Donor (94.01%):** The model performs well in identifying Blood Donors.

**Poor Performance for Suspect Blood Donor, Hepatitis, Fibrosis**: The model fails to identify Hepatitis cases.

**Moderate Performance for Cirrhosis (88.89%):** The model moderately identifies Cirrhosis cases.

**Improvement Needed for Suspect Blood Donor, Hepatitis, Fibrosis:** These classes need better precision and recall.

## Cluster

A white background with black text

Description automatically generated

Figure 8: Cluster model

**Centroid Table**

A table with numbers and a few black numbers

Description automatically generated with medium confidence

Figure 9: Centroid Table

**Performance:**

A table with numbers and percentages

Description automatically generated

Figure 10: Clustering performance

**Accuracy: 86.5%**

**Precision:**

* **Cluster 0: 92.77%**
* **Cluster 1: 12.82%**
* **Cluster 2: 0.00%**
* **Cluster 3: 0.00%**
* **Cluster 4: 100%**

**Recall:**

* **Blood Donor: 98.69%**
* **Suspect Blood Donor: 71.43%**
* **Hepatitis: 0.00%**
* **Fibrosis: 0.00%**
* **Cirrhosis: 3.33%**
* The model performed well in identifying Blood Donors (Cluster 0) with a precision of 92.77% and recall of 98.69%.
* Poor performance was noted for other categories like Suspect Blood Donor, Hepatitis, and Fibrosis.
* Cirrhosis identification was moderately successful with a precision of 100% and a recall of 3.33%.

## Outlier Detect

LOF vs Distance

A screenshot of a computer

Description automatically generated

Figure 11: Outlier detection pivot table

1. Dis\_outlier = false, lof\_outlier = false, count(label) = 584

* This indicates that 584 samples were not detected as outliers by either the distance-based method or the LOF method.
* This is the case for most samples (584), suggesting that both methods agree that these samples are normal.

1. Dis\_outlier = false, lof\_outlier = true, count(label) = 21

* This indicates that 21 samples were not detected as outliers by the distance-based method but were detected as outliers by the LOF method.
* This suggests that the LOF method tends to detect some outliers that the distance-based method misses.

1. Dis\_outlier = true, lof\_outlier = true, count(label) = 9

* This indicates that 9 samples were detected as outliers by both the distance-based method and the LOF method.
* These samples are consistently identified as outliers by both methods, showing some agreement between the two.

1. Dis\_outlier = true, lof\_outlier = false, count(label) = 1

* This indicates that 1 sample was detected as an outlier by the distance-based method but was not detected as an outlier by the LOF method.
* This suggests that the distance-based method detected some outliers that the LOF method did not.

A screenshot of a computer

Description automatically generated

Figure 12: LOF vs. Distance Outlier detection comparison

These are the list of the 9 common outliers.

# Conclusion

In this analysis, we explored and analyzed the HCV dataset by performing classification using Decision Trees, clustering using k-Means, and outlier detection using LOF and distance-based methods in RapidMiner. Below are the summarized findings and conclusions:

1. Decision Tree:

* The Decision Tree model showed high accuracy in identifying Blood Donors with a precision of 96.12% and recall of 97.56%.
* The model struggled with identifying Hepatitis cases, with a precision and recall of 0%.
* Moderate performance was observed for Cirrhosis cases, with a precision of 73.08% and recall of 63.33%.
* The model requires improvement for identifying Fibrosis and Suspect Blood Donor cases.

1. Clustering using k-Means:

* The clustering model had an overall accuracy of 86.5%.
* The model performed well in identifying Blood Donors (Cluster 0) with a precision of 92.77% and recall of 98.69%.
* Poor performance was noted for other categories like Suspect Blood Donor, Hepatitis, and Fibrosis.
* Cirrhosis identification was moderately successful with a precision of 100% and a recall of 3.33%.

1. Outlier Detection:

* Two methods, LOF and distance-based, were used for outlier detection.
* Most of the samples were not detected as outliers by either method.
* The LOF method detected 21 additional outliers that the distance-based method did not detect, indicating that the LOF method may be more sensitive to certain outliers.
* 9 samples were consistently identified as outliers by both methods, showing some agreement between the two.

# Reference

[1] Lichtinghagen,Ralf, Klawonn,Frank, and Hoffmann,Georg. (2020). HCV data. UCI Machine Learning Repository. https://doi.org/10.24432/C5D612.