Data Analysis and Data Mining
Assignment 2

FULL NAME: *Brahim Anis Belferkous* **Neptun Code:** *NI9WQN*

Topic: Hepatitis C Data Analysis

Introduction:

Hepatitis C is a viral infection that causes liver swelling, called inflammation. Hepatitis C can lead to serious liver damage. The hepatitis C virus (HCV) spreads through contact with blood that has the virus in it.

Data Description and Source:

The Hepatitis C dataset was obtained from the UCI Machine Learning Repository. It contains data on 159 patients with hepatitis C, including their age, sex, blood test results, and disease category (normal, fibrosis, or cirrhosis).

Repository: https://archive.ics.uci.edu/ml/datasets/HCV+data

Content:

All attributes except Category and Sex are numerical.

Attributes 1 to 4 refer to the data of the patient:

1) X (Patient ID/No.)

2) Category (diagnosis) (values: '0=Blood Donor', '0s=suspect Blood Donor', '1=Hepatitis', '2=Fibrosis', '3=Cirrhosis')

3) Age (in years)

4) Sex (f,m)

Attributes 5 to 14 refer to laboratory data:

5) ALB, ALP, ALT, AST, BIL, CHE, CHOL, CREA, GGT, PROT

The target attribute for is Category (2): blood donors vs. Hepatitis C patients (including its progress ('just' Hepatitis C, Fibrosis, Cirrhosis).

Data Analysis Procedures:

The data was first cleaned and preprocessed to handle missing values and outliers. Then, the data was explored using descriptive statistics and correlation analysis. Finally, K-means clustering and k-nearest neighbors' classification were used to analyze the data and make predictions about the disease category.

Results:

Descriptive Statistics

- The average age of the patients is 46.18 years.
- The majority of the patients are male (57.47%).
- The laboratory values exhibit varying ranges and distributions.

Optimal Number of Clusters

- Based on the silhouette analysis, the optimal number of clusters in the dataset is two.

Silhouette Score

- The silhouette score of 0.3873 indicates a moderate separation between the identified clusters.

The correlation analysis Showed that there were moderate to strong correlations between several of the blood test results such AST-ALT, ALB-PROT, CREA-SEX

Accuracy of KNN Model

- The k-nearest neighbors classification model achieved an accuracy of 97.14% on the test set, suggesting high predictive performance for distinguishing between blood donors and Hepatitis C patients.

Evaluation of Results

The results of the analysis suggest that the blood test results can be used to predict the disease category of patients with hepatitis C. The k-nearest neighbors classification model achieved a reasonable level of accuracy, but further work is needed to improve the data to get more accurate model.