Parshvanath Charitable Trust's
A. P. SHAH INSTITUTE OF TECHNOLOGY, THANE
(All Programs Accredited by NBA)
Department of Information Technology

Academic Year: 2024-25
Semester: VI
Class / Branch: TE-IT
Subject: BI Lab
Name of Instructor: Prof. Apeksha Mohite

Name of Student: Aaryan Ghawali
Student ID: 22104071
Date of Performance: 5/04/25
Date of Submission: 5/04/25

Aim: Business Intelligence Mini Project

1. Problem Definition:

The project aims to apply data mining classification techniques such as Random Forest, Naive Bayes, and J48 Decision Tree on the HCV dataset to predict patient outcomes (LIVE or DIE). By analyzing biochemical and medical attributes, the goal is to assist in early detection and decision-making in hepatitis care.

2. Data mining task to be performed:

1] Classification using Random Forest Algorithm.

2] Classification using Naive Bayes Algorithm

3] Classification using Decision Tree (J48) Algorithm

3. Dataset identified: Name of your dataset.

HCV dataset

4. Source of dataset:

https://archive.ics.uci.edu/dataset/571/hcv+data

Details of the dataset:

Name of Dataset: HCV Dataset

Brief Description:

The Hepatitis dataset contains medical records of 155 patients with 20 features related to liver function and Hepatitis C infection. It is used to predict patient survival with the target label as **LIVE** (survived) or **DIE** (not survived). The dataset includes **6 numeric attributes** (e.g., AGE, BILIRUBIN, SGOT) and **14 categorical attributes** (e.g., SEX, STEROID, FATIGUE). Categorical features are label-encoded, where values like 1 = no/female and 2 = yes/male. It serves as a benchmark for classification tasks in medical data analysis. The dataset has some missing values, making it ideal for testing imputation and model robustness.

Target Label:

The target label indicates the patient's survival outcome:

- LIVE means the patient survived Hepatitis.
- DIE means the patient did not survive.

This helps in predicting patient risk and guiding treatment. Out of 155 records, 123 are LIVE and 32 are DIE, showing a class imbalance.

Feature Type:

- **Numeric Attributes (6):**

  AGE, BILIRUBIN, ALK PHOSPHATE, SGOT, ALBUMIN, PROTIME

- **Categorical (Symbolic) Attributes (14):**

  SEX, STEROID, ANTIVIRALS, FATIGUE, MALAISE, ANOREXIA, LIVER BIG, LIVER FIRM, SPLEEN PALPABLE, SPIDERS, ASCITES, VARICES, HISTOLOGY

Encoding:

Categorical attributes are label encoded (e.g., SEX: 1=female, 2=male; STEROID: 1=no, 2=yes).

5. Algorithms to accomplish the task:

1) Classification using Random Forest Algorithm

A Random Forest classifier was applied to the Hepatitis C Virus (HCV) dataset using the Weka tool. This algorithm is based on ensemble learning, where multiple decision trees are trained and their outputs are combined (via majority voting) to make the final prediction. In this implementation, 100 trees were generated to classify patients into two categories: LIVE (survived) and DIE (not survived). Random Forest helps reduce overfitting and improves prediction accuracy, especially on imbalanced datasets like this one**.**

Implementation Details:

- **Tool Used:** Weka
- **Algorithm:** Random Forest

Performance Metrics:

- **Accuracy:**85.16%

  This indicates that the model correctly predicted the survival status of patients in approximately 85 out of 100 cases.

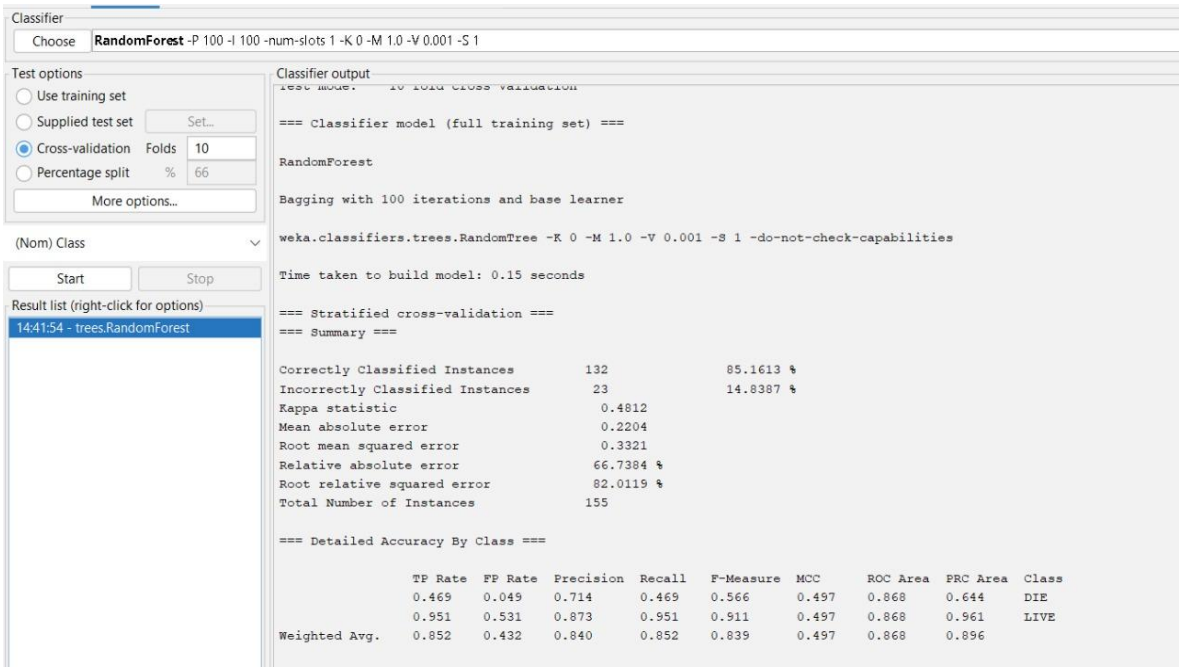

Fig 6.3 Accuracy of Random Forest classification Algorithm

2) Classification using Naive Bayes Algorithm:

Naive Bayes is a probabilistic classifier based on Bayes' Theorem, assuming feature independence. It is efficient for medical datasets and performs well with both categorical and numerical data, making it suitable for the HCV dataset. Its ability to handle large datasets with multiple features without requiring extensive computational resources makes it an ideal choice for real-time prediction tasks in healthcare. Additionally, the simplicity and interpretability of Naive Bayes provide valuable insights, especially when evaluating the likelihood of a disease based on various medical attributes.

Implementation:

- **Tool Used:** Weka
- **Algorithm:** Naive Bayes (Bayes.NaiveBayes in Weka)

Performance Metrics :

Accuracy: ~84.52%

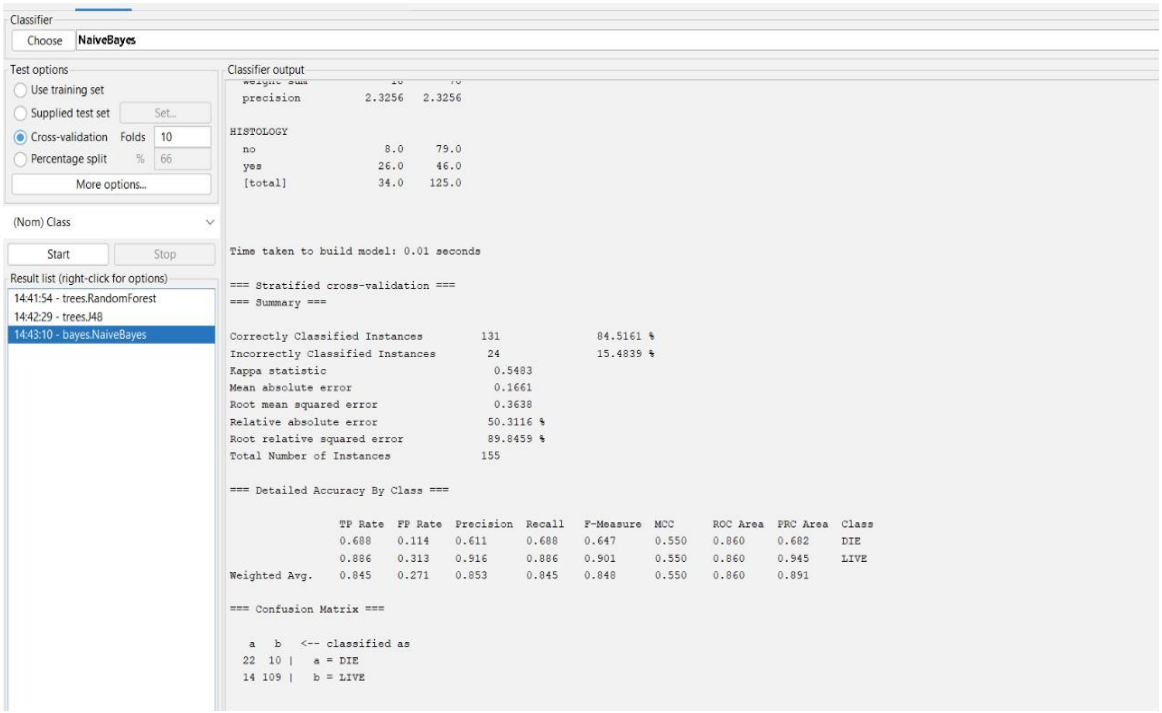

Fig 6.3 Accuracy of Naïve Baye's classification Algorithm

C. Classification using Decision Tree (J48) Algorithm

J48 is an implementation of the C4.5 algorithm, which is a popular and effective decision tree learning algorithm. It builds a tree by recursively partitioning the dataset into subsets based on information gain and entropy, making it well-suited for both categorical and numerical features. The decision tree generated by J48 helps in creating an interpretable model, where each node represents a decision rule, and each leaf node represents a class label.

Implementation:

- **Tool Used:** Weka
- **Algorithm:** J48 (C4.5 Decision Tree)

Performance Metrics: ~83.87%

- The J48 classifier produced slightly lower accuracy compared to Random Forest and Naive Bayes, but it provided an easily interpretable model through a clear decision tree structure.

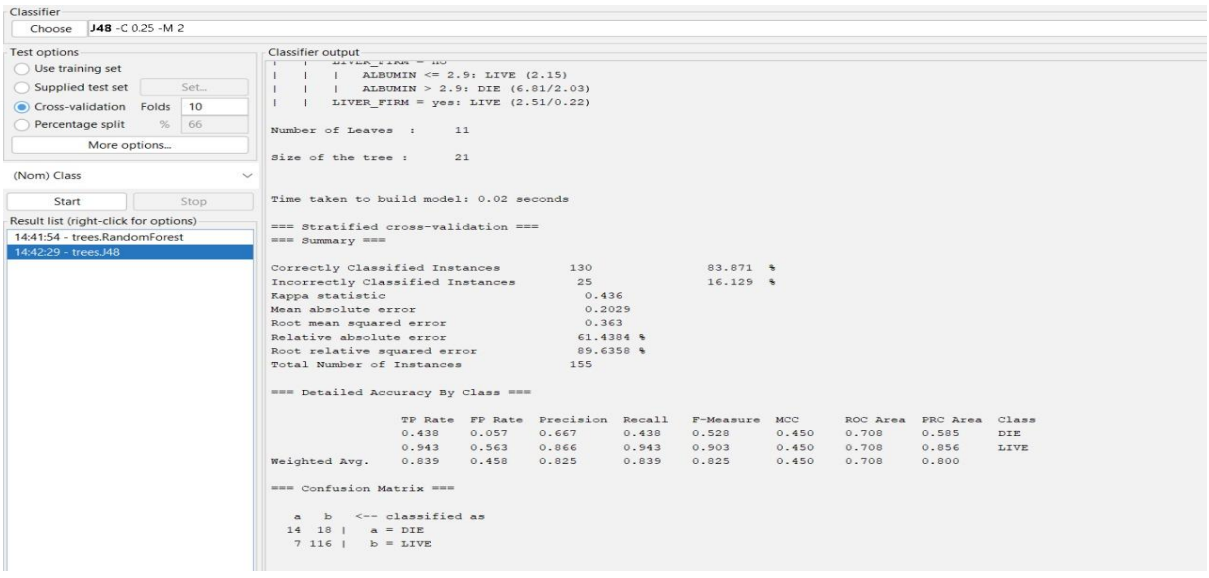

Fig 6.3 Accuracy of Decision tree of (J48) classification Algorithm

6.Conclusion:

The analysis of the **HCV (Hepatitis C Virus) dataset** using different **classification algorithms** provided meaningful insights into patient health classification and the suitability of models in medical diagnostics.

Algorithm	Accuracy
Random Forest	~85.16%
Naive Bayes	~84.52%
Decision Tree (J48)	~83.87%

- The **Random Forest classifier** achieved the highest accuracy of **85.16%**, proving to be the most effective in handling the complexity and variability within medical records.
- **Naive Bayes** also performed well, with an accuracy of **~84.52%**, though its assumption of feature independence may limit its performance in complex datasets.
- **Decision Tree (J48)** achieved **~83.87% accuracy**, offering the advantage of interpretability, although with slightly lower performance compared to ensemble methods.

Final Verdict:

Based on the evaluation, the **Random Forest Classification algorithm** can be considered the **most suitable** for the **HCV dataset**, due to its higher accuracy and robustness across diverse patient data.