

Liver Disease Prediction



INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI



INTRODUCTION

The liver is the largest internal organ of the human body and the only organ with regenerative abilities. Liver diseases account for a million deaths all around the world. There are more than 100 types of liver diseases. Our commitment was motivated by conviction that early prediction is not only just medical strategy but also a light of hope for that person and we want to accomplish that goal by using Machine Learning.

Traditional liver diagnostic methods for detecting various liver diseases are very costly and prediction is necessary if the prediction is done early then the damaged part of the liver can be removed and even if a person is left with only 10% of a healthy liver he can easily regenerate it.



The research papers are as follows:

1. Liver Disease Detection by Deepika Bhupathi, Christine Nya-Ling Tan and Sayan Kumar Ray of Manukau Institute of Technology

In this the authors have majorly used five research papers along with some minor ones:

- Logistic Regression which is a simple linear classifier.
- Decision Trees are used to capture non-linear patterns and interactions between features and Classification and Regression Trees (CART) were also used.
- Random Forest is an ensemble of decision trees offering robustness against overfitting.
- Support Vector Machines (SVM) are Useful for high-dimensional data and complex decision boundaries.
- Neural Networks are deep learning models that capture intricate patterns in data.
- Some of the other methods included were K-Nearest Neighbors (K-NN), Linear Discriminant Analysis (LDA), Unsupervised Algorithm (Autoencoders) and Naive Bayes.

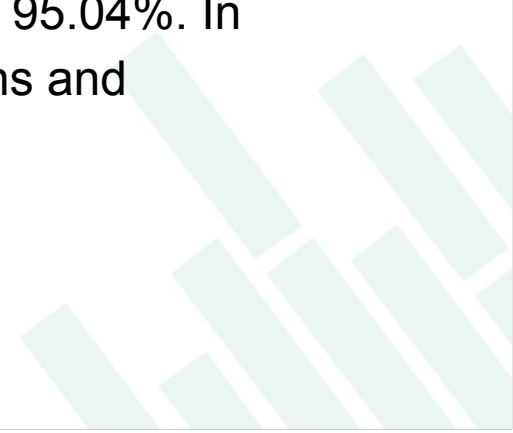
After that evaluation is done using Accuracy, Precision, Sensitivity, F1 Score and ROC curve.

Autoencoders and K-NN had the highest accuracy with 92.1% and 91.7% respectively.

2. Liver Disease Prediction using Machine Learning by Vasan Durai, Suyan Ramesh, and Dinesh Kalthireddy of SRM Institute of Science and Technology

The research paper emphasizes the significance of digital technologies in medical procedures. The research focuses on the combination of the SVM and Naive Bayes for the diagnosis. The primary literature includes the prediction of the life expectancy of cirrhosis patients, the detection of alcohol-induced diseases, and the prognosis of liver cancer using the Bayes Theorem.

Many factors, like data quality, feature selection, etc., will significantly influence the accuracy prediction. The experimental study includes many evaluations using different machine learning algorithms but showcased the J48 algorithm as the standout, giving an accuracy rate of 95.04%. In essence, machine learning presents a viable route for improving liver disease predictions and addressing the current challenges.



Dataset Description



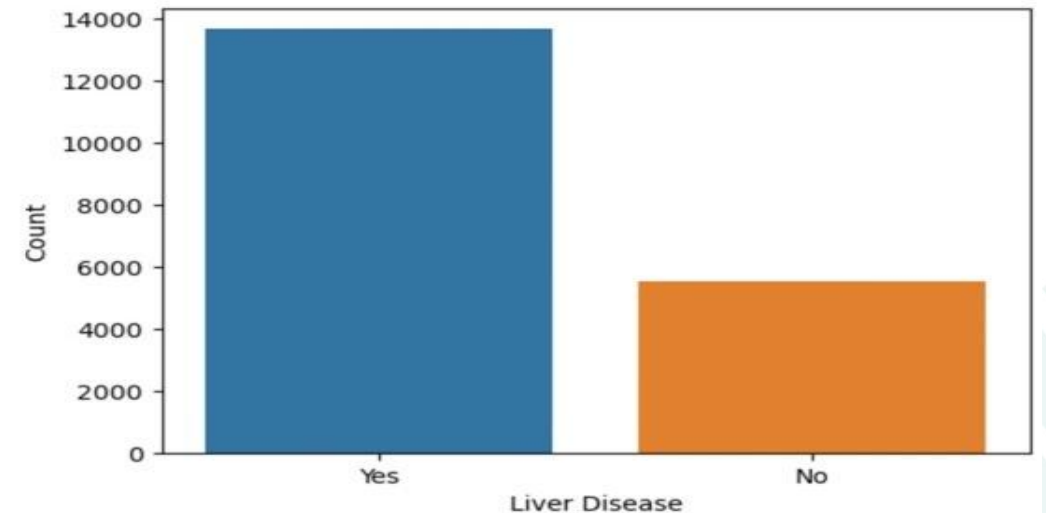
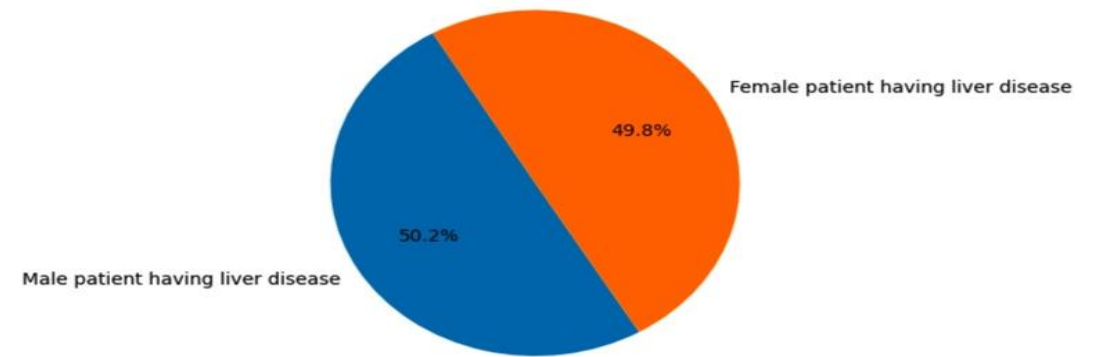
Dataset Details

- The Liver Disease Patient dataset on Kaggle is used to classify liver disease.
- It contains records of liver patients from across the world. It is a structured dataset and has a total of 30691 rows and 11 columns.
- The columns contain features like age, gender of the patient, total bilirubin, direct bilirubin, Alkphos Alkaline Phosphotase, Sgpt Alamine Aminotransferase, Sgot Aspartate Aminotransferase, Total proteins, ALB albumin and A/G Ratio Albumin and Globulin Ratio and the target variable, Result.
- The Result column consists of class 1 for those with liver disease and class 2 for those without liver disease.

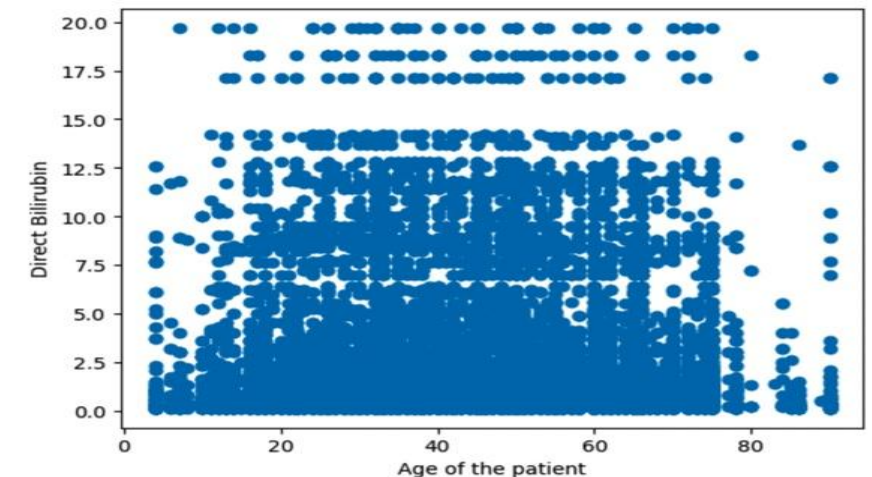
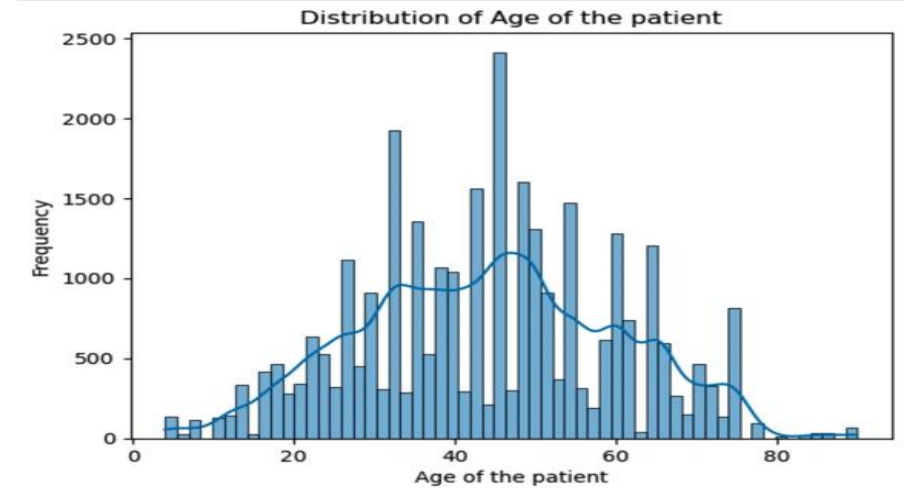
Exploratory Data Analysis




- The pie chart suggests the percentage of male patients having liver disease(50.2%) is nearly equivalent to that of female patients(49.8%).
- From the following frequency distribution, it can be inferred that 13677 patients have liver disease and 5529 do not have liver disease.



- It can be observed from the following histogram that most patients coming for liver disease diagnosis lie between the age group of nearly 30 to 60 years.
- It can be observed from the scatter plot that the data is positively skewed and prone to outliers. The median of the direct bilirubin is 0.3mg/dl. The value of direct bilirubin should be less than 0.3mg/dl; higher levels of it indicate liver damage.



Preprocessing Steps- The following steps were carried out to make the data suitable for analysis:

- We applied Median imputation for missing values in the dataset.
 - The numerical values were normalised using MinMaxScaler, and the outliers were removed using the Interquartile range (IQR) method.
 - The gender of the patient is categorical data and was encoded using BinaryEncoder.
 - Duplicate rows were present in the data, so they were dropped. After data preprocessing, the dataset had 19206 rows and 13 columns.
 - The class distribution for 'Result' depicts an imbalance in the data. The Synthetic minority oversampling technique (SMOTE) method was used to balance the data.
- 
- A decorative geometric pattern is located in the bottom right corner of the slide. It consists of several light blue, semi-transparent rectangular bars of varying lengths and orientations, creating a modern, abstract design.

Methodology



Feature Selection

1. ANOVA F-test:

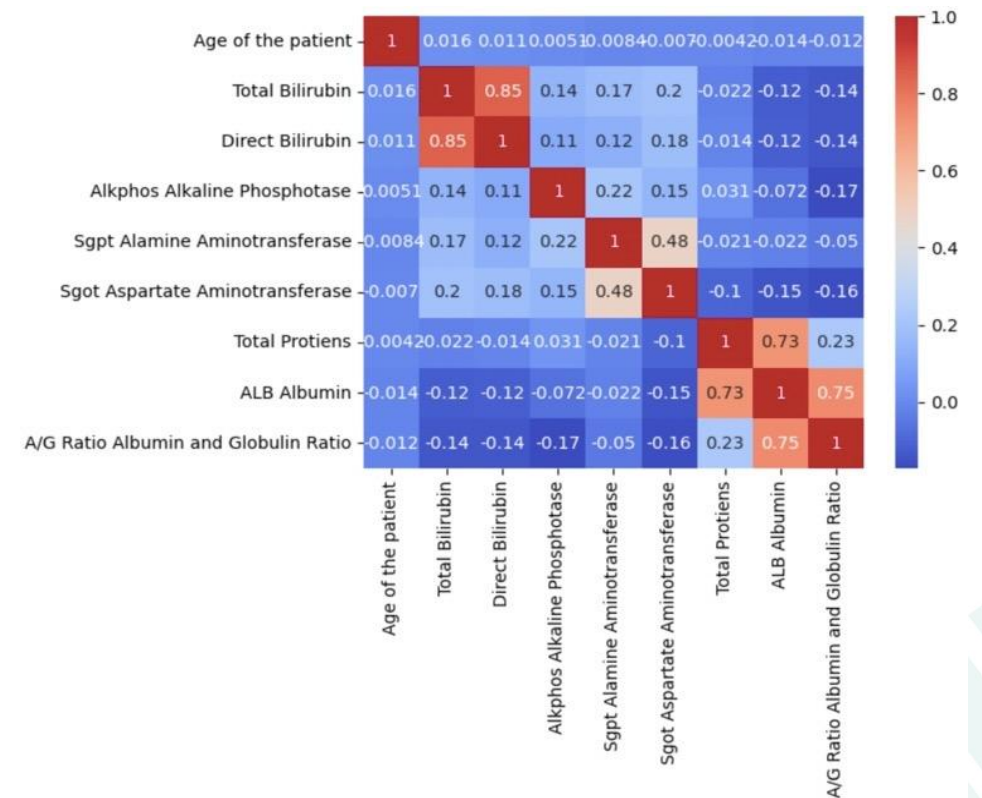
Used to assess the link between numerical features and 'Result'; 'Age of the patient' was dropped due to non-significant results.

2. Chi-Squared Test:

Checks 'Gender' dependency on 'Result'; with a p-value of 0.47 (≥ 0.05), 'Gender' is dropped.


3. Correlation Heat Map:

Strong correlations led to dropping 'ALB Albumin,' 'Total Bilirubin' for reduced feature dependency.




Model Details

We have used 6 models:-

- Support Vector Machines (SVM) were used due to their effectiveness in high-dimensional spaces.
 - Naive Bayes: Chosen for its simplicity and fast computational speed.
 - Decision Trees: To capture complex relationships in the data.
 - Random Forest: An ensemble of decision trees to improve generalisation..
 - KNN(K- Nearest Neighbors): For classification and regression tasks when data has local patterns or similarity matters.
 - Neural Networks: For capturing intricate patterns and interactions between features.
- 

Model Evaluation: Model performance will be tabulated and compared using the predefined evaluation metrics of :

- Accuracy: The proportion of true results among the total number of cases examined.
 - F1-Score: It is the harmonic mean of precision and recall.
 - ROC-AUC Curve: Quantifies a classifier's overall performance, showing its ability to distinguish between classes.
 - Confusion matrix: Shows a classifier's performance, summarising true positives, true negatives, false positives, and false negatives in a tabular format.
 - Precision: Measures the proportion of true positive predictions among all positive predictions.
- 
- In the bottom right corner, there is a decorative graphic consisting of several light teal diagonal bars of varying lengths and orientations, creating a modern, abstract pattern.

Hyperparameter tuning: The following table depicts the parameters used to optimize the performance of the models:-

Models	Hyperparameters	Optimal Values
Decision Trees	max_depth: [None, 10,20, 30] min_samples_split: [2,5,50,100] min_samples_leaf: [1,2,4]	None 2 1
Random Forest	n_estimators: [50,100] max_depth: [None, 10, 30] min_samples_split: [2,3,5] min_samples_leaf: [1,2,4]	100 30 2 1
KNN	n_neighbors:[3,5,7,9] Weights:['uniform', 'distance'] metric:['euclidean', 'manhattan']	7 distance manhattan

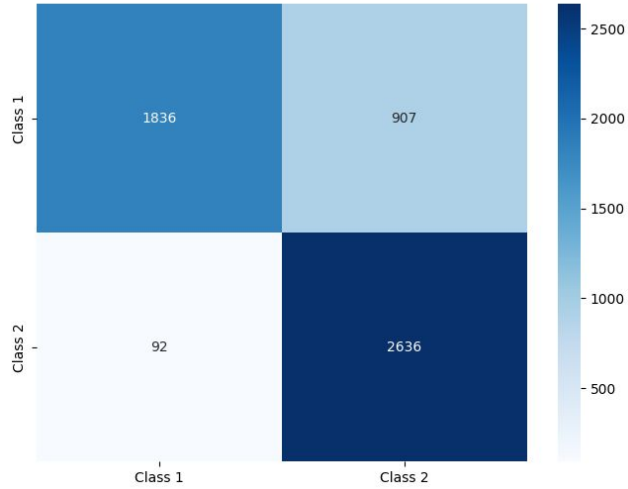
Result and Analysis



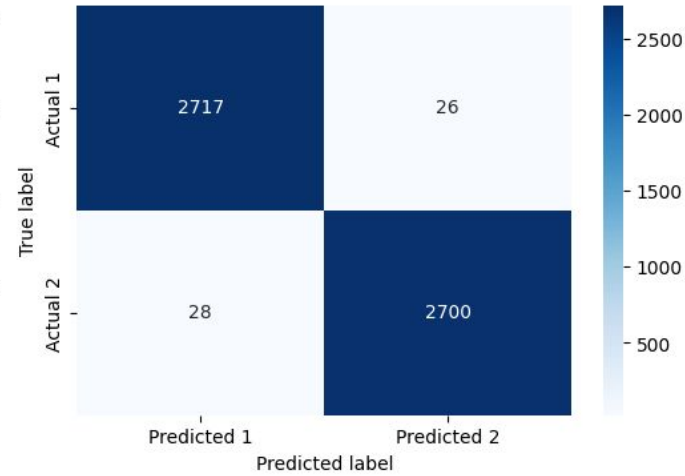
The model yielding the highest performance across evaluation metrics among the models implemented is Random Forest, with an accuracy of 99.43%.

Models	Accuracy (in %)	F1-score (in %)	Precision (in %)
Naive Bayes	65.47	58.18	74.06
SVM	81.74	78.6	95.22
Decision Trees	99.01	99.01	98.97
Random Forest	99.43	99.43	99.09
KNN	99.39	99.39	99.59
Neural Network	88.52	88.92	85.71

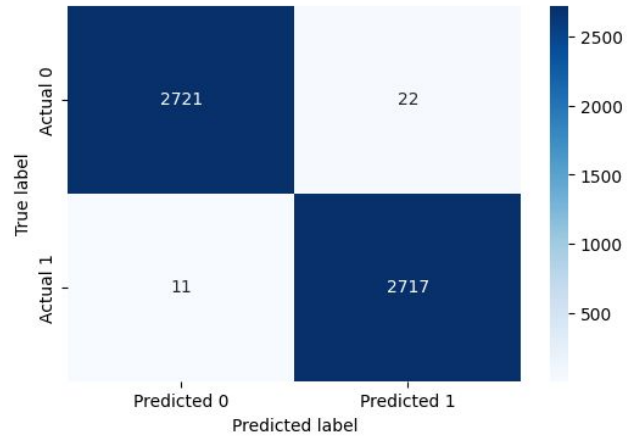
SVM Confusion Matrix



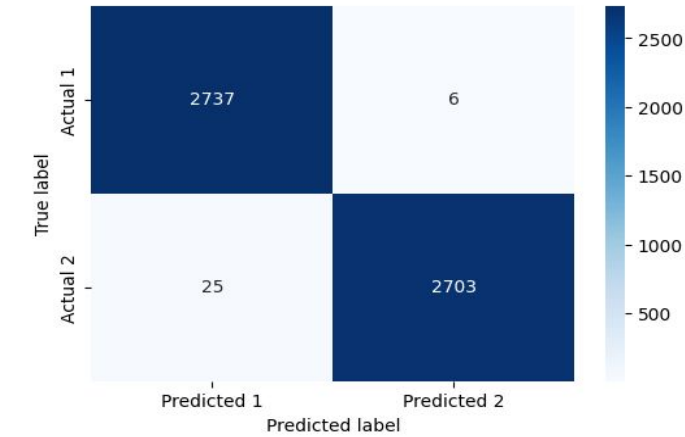
Decision Tree Confusion Matrix



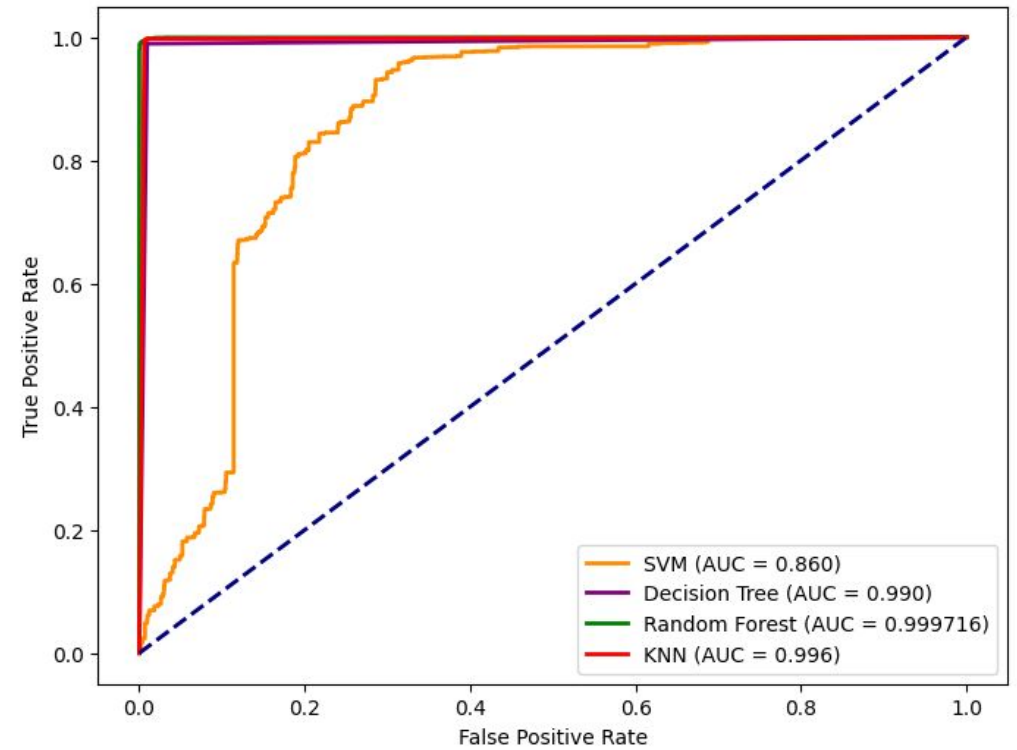
Confusion Matrix - KNN



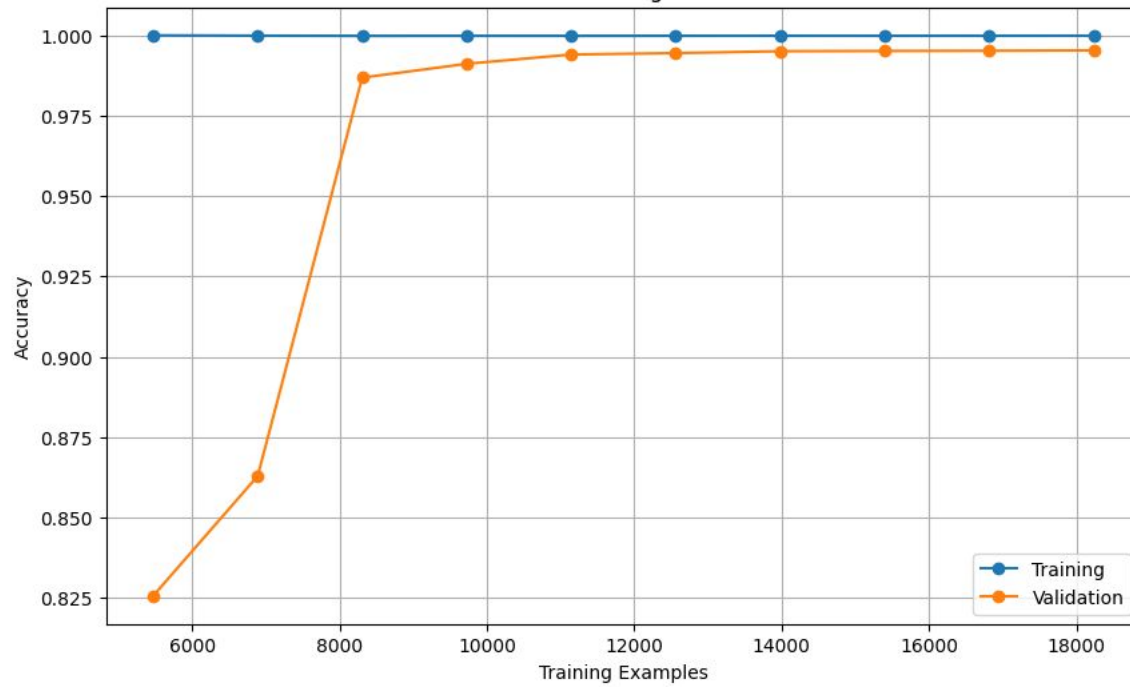
Random Forest Confusion Matrix



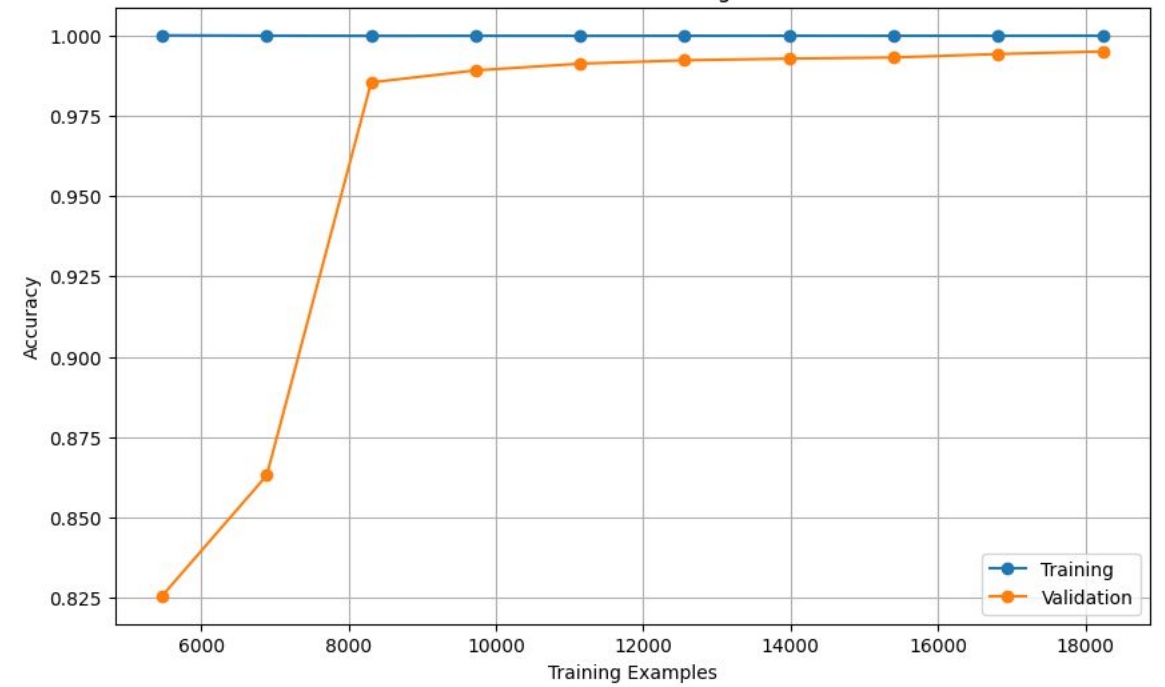
ROC-AUC Curve



KNN Learning Curve



Random Forest Learning Curve



Conclusion



The machine learning project for liver disease prediction showcased notable success, with the Random Forest model achieving the highest accuracy of 99.43%. This approach not only offers substantial savings in diagnostic costs but also marks a significant advancement in early detection. The project's findings affirm the potential of machine learning in enhancing healthcare outcomes, especially in settings with limited resources, and set a promising direction for future research in predictive medicine. For future work, further validation on different datasets is required to evaluate the model's performance on unseen data.



Contribution



- Aryan Rohilla (2021024) : Model Details (SVM, Naive Bayes, KNN(K -nearest Neighbors)), Result(Accuracy, F1-score, Precision) and hyperparameter tuning.
- Avinash Barala(2012028) : Preprocessing (describing data structure, dropping unnecessary columns, removal of Outliers by using IQR method and replacing them with Median Values, Normalizing the data and Min Max Scaling), ROC-AUC Curves and Confusion Matrix for all models (SVM, DT, RF and KNN).
- Ekansh (2021044) : Literature review in the research methodology of the Research Papers, Model Details (Neural Network), Result(Accuracy, F1-score, Precision)
- Shruti Jha (2021289) : Preprocessing(encoding and removal of duplicate rows), EDA, feature selection, models(decision trees and random forest) and hyperparameter tuning.

THANK YOU

A decorative graphic in the bottom right corner consisting of several parallel, diagonal teal bars of varying lengths, creating a sense of movement or a modern design element.