



# Concepts and Technologies of AI

## 5CS037

Student Name: Aaditya Acharya

Student ID: 2510333

## Abstract

The goal of this report is to identify the categorical target variable using the classification techniques. 30691 patient data of Liver Patient Disease Dataset with 11 features that supports the UN sustainable development Goal 3 (Good Health and Well-Being) is used. The project analyzed the 19083 patients after the successful cleaning of the data. The methodology includes the EDA, Neural Network (MLP), Classical ML (Logistic Regression, Decision Tree Classifier), Hyperparameter Optimization (Randomized Search CV), Feature Selection (Recursive Feature Elimination) and final model comparison. For the evaluation of the model Accuracy, Precision, Recall, F1-Score was calculated. The Decision Tree model achieved the accuracy of 98% showing the strong predictive capability for the liver disease. Feature selection using RFE identified 6 features which is important for the final model.

## Table of Contents

<b>1</b>	<b>Introduction</b>	1
1.1	<b>Problem Statement</b>	1
1.2	<b>Dataset</b>	1
1.3	<b>Objective</b>	2
1.4	<b>Research Questions</b>	2
<b>2</b>	<b>Methodology</b>	3
2.1	<b>Data Pre Processing</b>	3
2.2	<b>Exploratory Data Analysis (EDA)</b>	3
2.4	<b>Model Evaluation</b>	6
2.5	<b>Hyperparameter Optimization</b>	7
2.6	<b>Feature Selection</b>	8
<b>3</b>	<b>Results And Conclusion</b>	9
3.1	<b>Key Findings</b>	9
3.2	<b>Final Model</b>	9
3.3	<b>Challenges</b>	10
3.4	<b>Future Work</b>	10
<b>4</b>	<b>Discussion</b>	11
4.1	<b>Model Performance</b>	11
4.2	<b>Impact of Hyperparameter Tuning and Feature Selection</b>	11
4.3	<b>Interpretation of results</b>	11
4.4	<b>Limitations</b>	12
4.5	<b>Suggestions for Future Research</b>	12
<b>5</b>	<b>References</b>	13

## Table Of Figures

Figure 1 class distribution of target variable .....	3
Figure 2 correlation with target variable .....	4
Figure 3 gender based analysis .....	4
Figure 4 Final Model .....	9

## 1 Introduction

### 1.1 Problem Statement

Liver disease shows the important health challenge so Early and effective diagnosis are crucial for the treatment of the patient. The goal of this project is to develop the model that predicts the patients at the risk of liver disease based on the clinical parameters. The challenge it to build a model that can help in identifying the risk of liver disease

### 1.2 Dataset

**Dataset Name:** Liver Disease Patient Dataset 30K train data

**Source :** Kaggle

**Total records:** 30,691

**Number of features:** 10 features with 1 target variable.

Features	Description
Age of the patient	Patients age
Gender of the patient	Patients gender
Total Bilirubin	Yellow pigments produced
Direct Bilirubin	Processed from bilirubin
Alkphos	Enzyme found in liver
Sgpt Alamine Aminotransferase	enzyme
Sgot Aspartate Aminotransferase	enzyme
Total Proteins	Total proteins blood
ALB	Main protein
A/G Ratio	Ration of albumin to proteins
Result	Liver disease or not

The Dataset aligns with the **SDG3: Good Health and Well Being**. The project directly supports the Sustainable Development Goal 3 by early detection of liver disease of the patients. By developing the model accurately the model aims to improve the health outcomes as early detection of disease helps to promote the well-being across the populations.

### 1.3 Objective

The primary objectives of this model are:

- Develop Multiple models for the liver disease classification.
- Perform the comprehensive data preprocessing.
- Compare the model using the standard evaluation metrics.
- Optimize the models through the hyperparameter tuning.
- Identification of important features for disease prediction.

### 1.4 Research Questions

The project seeks to answer the following questions:

- Can machine learning accurately predict the liver disease?
- What is the impact of hyperparameter optimization on model performance?
- How do classification algorithms differ from each other?

## 2 Methodology

### 2.1 Data Pre Processing

For the preprocessing of the data different preprocessing steps was performed. Initially data contains 30,691 data with different missing values, duplicate values, etc. for the cleaning the data columns names was standardized by removing the extra spaces with underscores. For the missing values Numerical features was filled with median values and categorical features (Gender) was filled with the mode. Duplicates patients was removed to prevent the bias in the data training. Outliers was detected using the IQR method which was clipped rather than removing it. After all this cleaning process final datasets contains 19.083 datasets without missing values, duplicate values, etc.

### 2.2 Exploratory Data Analysis (EDA)

For the exploratory data analysis different EDA techniques was performed.

#### Class Distribution

The dataset shows more liver disease cases(1) than healthy cases (2).

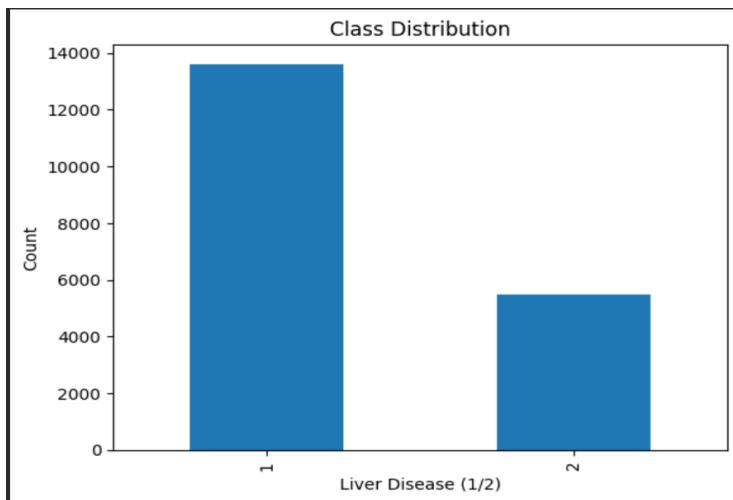


Figure 1 class distribution of target variable

## Correlation Analysis

Correlation analysis shows the Relation with the target variable.

Age_of_the_patient	-0.004597
Total_Bilirubin	-0.313699
Direct_Bilirubin	-0.320679
Alkphos_Alkaline_Phosphotase	-0.230792
Sgpt_Alamine_Aminotransferase	-0.286602
Sgot_Aspartate_Aminotransferase	-0.298538
Total_Protiens	0.036707
ALB_Albumin	0.165456
A/G_Ratio_Albumin_and_Globulin_Ratio	0.179971
Result	1.000000
Name: Result, dtype: float64	

Figure 2 correlation with target variable

Direct Bilirubin shows the strongest negative correlation as age of the patient shows the minimal correlation which shows that age is not only the factor for the disease. Total proteins shows the positive correlation which shows that increase in proteins decreases the chance of disease.

## Gender Based Analysis

To know the rate of disease across male and female gender based analysis was performed which shows that more number of patients is male compared to the female. This also helps to know the demographic factors in liver disease.

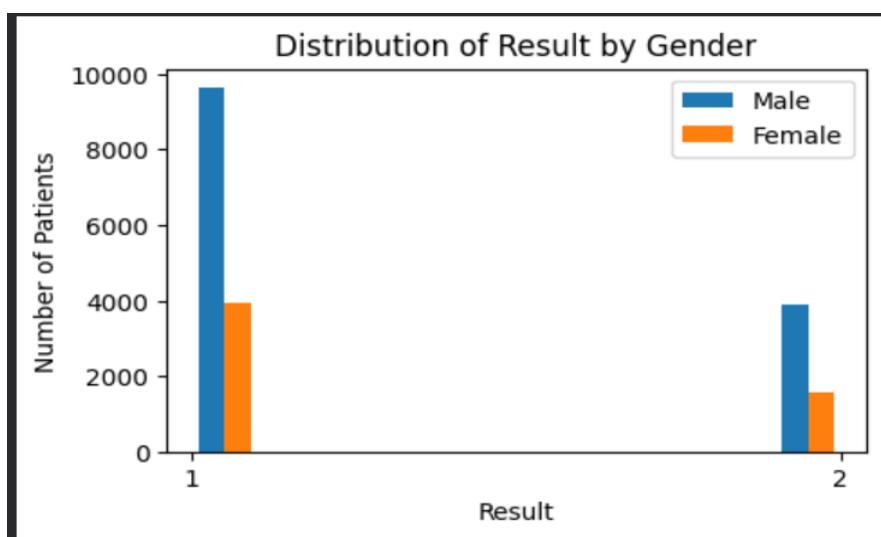


Figure 3 gender based analysis

## 2.3 Model Building

### Feature Selection

- The gender was removed before training the model as it is categorical data.
- For the training and testing the dataset was split into training(80%) and testing(20%).
- Standard scaler was applied as standardization is crucial for the algorithms like MLP and Logistic Regression.

### Multi-Layer Perception (MLP)

MIP is a neural network classifier with different hidden layers, activation functions, etc. which is capable of learning complex non-linear patterns in the data.

The model was configured using the following parameters:

- Hidden Layers: (50,25) neurons.
- Activation Function: ReLU
- Solver: Adam
- Regularization: alpha(0.001)
- Learning rate: adaptive
- Maximum iterations: 300
- Early stopping: True to prevent the overfitting

**Loss Function:** Log loss for the binary classification.

**Optimizer:** Adam optimizer with Adaptive Learning Rate.

### Logistic Regression

Logistic Regression is a linear classification model finds the probability of class which performs well on the binary classification data. To build the model standard scaler data was used in LogisticRegression to train and predict the model in which 80% was used for training whereas 20% was used for the testing of the data.

### Decision Tree Classifier

Random Forest is a non linear model which is capable of building the complex models . To built the model DecisionTreeClassifier was used to fit the training dataset and predictions for the training and testing was done in which 80% was used for training whereas 20% was used for the testing of the data.

## 2.4 Model Evaluation

The model performance was evaluated using the following metrices:

### Multi-Layer Perception (MLP)

Metrices	Training performance	Testing Performance
Accuracy	0.95	0.94
Precision	0.95	0.94
Recall	0.98	0.97
F1-Score	0.96	0.96

### Logistic Regression

Metrices	Training performance	Testing Performance
Accuracy	0.72	0.73
Precision	0.75	0.76
Recall	0.90	0.90
F1-Score	0.82	0.83

### Decision Tree

Metrices	Training performance	Testing Performance
Accuracy	1.0	0.98
Precision	1.0	0.99
Recall	1.0	0.98
F1-Score	1.0	0.99

These Classification metrices was chosen as this is the most commonly used metrices.

### Initial Model Comparison

Decision Tree have 99 % test accuracy and logistic regression have 73%. However decision tree shows perfect 100% accuracy showing potential overfitting.

## 2.5 Hyperparameter Optimization

RandomizedSearchCV was used in both logistic and Random Forest models. This approach explores the parameters by randomly sampling the combinations, maintaining the computational cost.

### Logistic Regression

To improve the performance of the model, Parameters includes the regularization strength (C), penalty type (L1,L2), fit intercept, maximum iteration and solver. The search identified the best hyperparameters for the model.

- **solver:** liblinear
- **penalty:** l1
- **max\_iter:** 500
- **fit\_intercept:** True
- **C:** 100

### Decision Tree

To improve the performance of the model, Parameters includes the splitter , maximum depth, minimum samples split and leaf , maximum features and criteration. The search identified the best hyperparameters for the model.

- **splitter:** 'best'
- **max\_depth:** 50
- **min\_samples\_split:** 20
- **min\_samples\_leaf:** 2
- **max\_features:** 'sqrt'
- **criterion:** log\_loss

### Cross-Validation Scores

5-Fold Cross-Validation was done during the hyperparameter search. The best cv achieved is 0.98 for decision tree and 0.72 for logistic regression.

## 2.6 Feature Selection

Feature selection was done using the Recursive Feature Elimination which is wrapper method. It works by recursively removing the least important features. The analysis helps to identify the important features that contribute to the liver disease. The top features that contribute the most include.

- Total\_Bilirubin
- Alkphos\_Alkaline\_Phosphotase
- Sgpt\_Alamine\_Aminotransferase
- Sgot\_Aspartate\_Aminotransferase
- Total\_Protiens
- ALB\_Albumin

RFE was selected as it balances between the model performance and feature selection and it also works well with various classifiers.

### 3 Results And Conclusion

Model	Features	CV Score	Accuracy	Precision	Recall	F1-Score
Logistic Regression	Selected (6)	0.72	0.73	0.77	0.91	0.83
Decision Tree	Selected (6)	0.98	0.98	0.99	0.99	0.99

#### 3.1 Key Findings

The model evaluated on test data shows:

- Got 94% testing accuracy in neural Network (MLP) which is very good.
- Got 98% accuracy in Decision Tree.
- Got 73% accuracy in logistic regression, it is not good but it finds 92% recall.
- 6 features out of 9 was used in the final evaluation.

#### 3.2 Final Model

The final model was evaluated using accuracy. Based on that the best model was decision Tree as it achieved 98% accuracy on the test data with excellent precision and recall of 99%. It also have high cv score 98%. Based on the accuracy decision tree achieved 98% accuracy whereas Logistic regression achieved 73%.

	Model	Features	CV Score	Accuracy	Precision	Recall	\
0	Logistic Regression	6	0.72	0.73	0.77	0.91	
1	Decision Tree	6	0.98	0.98	0.99	0.99	
<b>F1-Score</b>							
0				0.83			
1				0.99			

Figure 4 Final Model

### 3.3 Challenges

The challenges we faced are:

- out of 30691 only 19083 was used as data was duplicated, missing, etc.
- Testing all different things took a lot of computer time.
- The data is unbalanced as there is more sick patients rather than healthy which can cause model biased.

### 3.4 Future Work

To make the model better:

- More patients data can help model learn better patterns.
- More additional information can be added.
- More advanced ml algorithms can be used to make model more better.
- Testing on external dataset can be done.

## 4 Discussion

### 4.1 Model Performance

The model performance was evaluated using accuracy, precision, recall, etc. on both training and testing the model shows the very consistent performance. On Neural Network the model shows the 95% accuracy on training and 94 % accuracy on testing which means model is performing very well. On the Logistic regression the model shows 73% accuracy on testing data which is not good but it shows 92% recall which is good. In the case of Decision Tree the model shows 98% accuracy.

### 4.2 Impact of Hyperparameter Tuning and Feature Selection

Hyperparameter tuning and feature selection helped lot in model performance. Hyperparameter was performed using RandomizedSearchCV, which finds the optimal parameter for both logistic regression and random Forest Methods. After applying the technique the model perform 73% accuracy on Logistic regression with C=100, penalty=l1, solver ‘liblinear’ as C allows model to be complex and l1 will better suit for the real world data. There in not so big changes after the hyperparameter tuning. Hyperparameter somehow prevents overfitting in Decision Tree. In both models the improvement is small but meaningful.

Feature selection was performed using the Decision tree to find the most important features affecting the liver disease patients. After applying the technique The features was reduced from 9 to 6 which make model more faster. Different features like age, total\_proteins, ALB\_Albumin was removed.

### 4.3 Interpretation of results

The chosen features and model performed in a consistent manner with expectations. The Multi-Layer Perceptron (MLP) demonstrated strong and consistent performance, achieving a testing accuracy of approximately 94%. The Logistic Regression model recorded slightly lower accuracy (72%) but high recall of 92% High recall implies that model successfully majority of patients who successfully have liver disease. The Decision Tree achieved 98% accuracy on testing.

#### 4.4 Limitations

- Initially the data contains 30,691 patients data which is good but later the size decreases to 19083 after the cleaning of the as data contains duplicated value.
- There are more patients with disease cases than healthy case.
- Missing many features like treatment history, etc.

#### 4.5 Suggestions for Future Research

- Perform testing on more independent dataset.
- Other classification algorithms can be tested.
- Use of larger dataset and increase the quality of the dataset.
- Handle the class imbalance to improve the fairness.
- Apply more better feature selection techniques.

## 5 References

- Shrivastava, A., 2021. *Liver Disease Patient Dataset 30K train data*. [Online]  
Available at: <https://www.kaggle.com/datasets/abhi8923shriv/liver-disease-patient-dataset>  
[Accessed 2026].
- United Nations, n.d. *Good Health And Well Being*. [Online]  
Available at: <https://sdgs.un.org/goals/goal3>  
[Accessed 2026].