MSc Data Science

Component-3

Modelling, Regression & Machine-Learning

By: ANGAD PARTAP SINGH (M00912257)

Title - Early Detection of Liver Disease for Diagnosis

1. Introduction

According to the WHO, 2.9% of all fatalities are caused by liver disorders. India is currently ranked 63rd among them. Globally, liver disease is a serious health concern.

The objective of this work is to develop a machine learning-based technique for early liver disease detection using a dataset made up of a variety of clinical and laboratory indicators.

Abstract

Included in the procedure are feature engineering, exploratory data analysis, and the application of machine learning algorithms for categorization. Feature engineering involves the extraction of pertinent information from the raw data. Several classification techniques are put to the test, including logistic regression, decision trees, random forests, and support vector machines, to determine which one is the most effective at predicting liver disease.

The major findings of this study suggest that a few clinical and laboratory variables, including total bilirubin, direct bilirubin, alkaline phosphatase, and albumin, have a substantial role in the early diagnosis of liver illness. The created machine learning model exhibits encouraging performance and accuracy in categorising people with liver illness, offering a useful tool for early diagnosis.

Objective

This study seeks to address the problem of early liver disease diagnosis. For bettering patient outcomes and determining the most appropriate therapy approaches, early diagnosis of liver disease is essential. For prediction, a variety of machine learning techniques are utilised, including SVM and tree-based techniques like Decision Trees. The goal of this research study is to provide a comparative review of the five machine learning methods used in the medical field to diagnose and forecast liver disease.

Methodology

Data Description

The UCI ML Repository has provided this dataset for download. Information was acquired from the University of California, Irvine's Machine Learning Repository. Individuals' total bilirubin, alkaline phosphatase, total proteins, and demographic information like age and sex are all included in the information's lab reports. The data collection includes 11 attributes. Among these 11, the Sex and Result columns are categorical. In the Age column, continuous numbers are displayed. Data from 583 patients, including 441 men and 142 women, were included in the current investigation.

Models and algorithms deployed

By applying various machine learning models to a dataset, you may take advantage of their individual strengths, increase accuracy, learn about the peculiarities of the data, evaluate resilience, and select the model that will work the best for your particular purpose. Your machine learning system's overall performance and interpretability could be improved, resulting in more dependable and well-informed decision-making.

2. Data Preprocessing & EDA

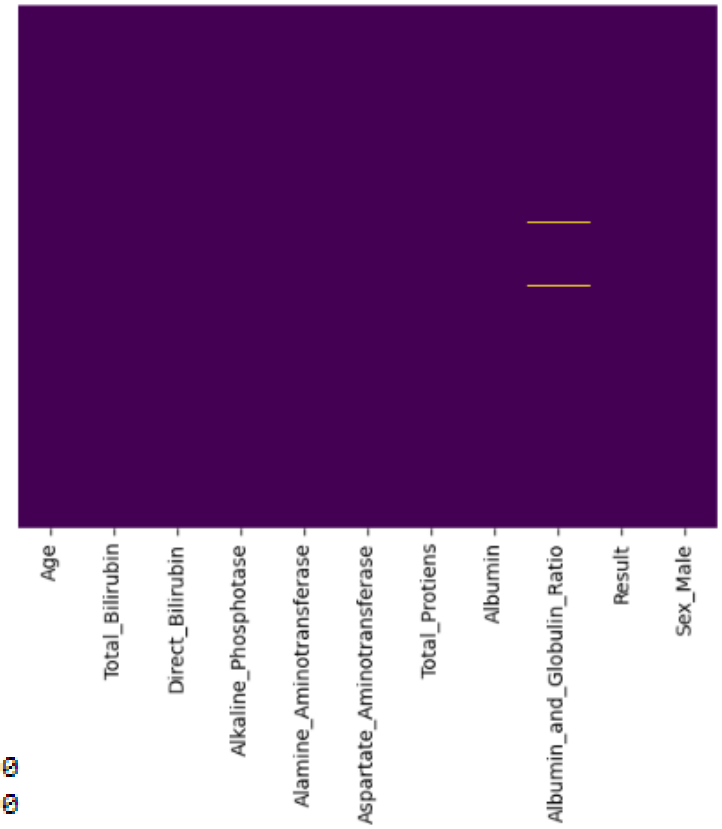
2.1 Null Value-Detection

An essential part of data preprocessing and analysis is finding null values in a dataset. There are 4 null values in our dataset. Null elements have a very low value, yet missing data are troublesome because, depending on their nature, they can occasionally lead to sampling bias. As a result of our data's unrepresentative sample size, it is possible that our findings cannot be applied to situations outside of our study.

IMPUTATION - Statistical methods to handle missing data.

One technique for replacing the missing values in a dataset is imputation. To fill in the gaps left by the missing data and account for their inherent variability an d uncertainty, multiple imputation is performed.

The graph shows that our missing val ues lie in the column ‘Albumin_and_ Globulin_Ratio’.




```
Age                          0
Total_Bilirubin              0
Direct_Bilirubin             0
Alkaline_Phosphotase         0
Alamine_Aminotransferase     0
Aspartate_Aminotransferase   0
Total_Protiens               0
Albumin                      0
Albumin_and_Globulin_Ratio   4
Result                       0
Sex_Male                     0
dtype: int64
```

It is also evident from the image that there are null values ion our dataset.

The Null Values needs to be taken care of before proceeding with the models.

Within regressive classes (i.e., categories like Albumin_and_Globulin_Ratio’ here), mean imputation can be done, and it can be stated as:

$$\hat{y}_{mi} = b_{r0} + \sum_{j} b_{rj} z_{mij} + \hat{e}_{mi}$$

Missing values are imputed, just like single imputation. The imputed values, however, are not selected from a distribution once, but rather m times. There should be m finished datasets at the conclusion of this stage.

2.2 Categorical Data

Different categories of data that can be categorised are represented by categorical variables. Examples of categorical variables are race, sex, age, and educational attainment. It is frequently more informative to group such variables into a relatively small number of groups, even though the age and highest grade earned for the latter two elements may also be analysed numerically by using exact values.

The column in our dataset 'Gender' depicts the categorical variables. The method used here to handle categorical variable is:

ONE-HOT ENCODING

The original data is supplemented by additional columns that indicate whether each potential value is present (or not) in the one-hot encoding. Utilising nomological categorical properties is this strategy. In one Hot Encoding technique, each category value is converted into a new column and given a value of 1 or 0. Before eliminating the first column, we will do this using the pandas get dummies function to avoid the dummy variable trap.

We have created a new column 'Male', wherein,

Male = 1

Female = 0

Gender	Male
Male	1
Female	0
Male	1
Male	1
Female	0
Male	1
Female	0
Female	0

2.3 Feature Selection

Dropping Features using PEARSON CORRELATION

This is a key method for choosing a specific feature for machine learning models. It is a way to express how linearly connected two sets of data are.

It effectively measures covariance in a normalised manner, with the result always falling between 1 and 1.

It is the ratio of the standard deviations of two variables' standard deviations to their covariance.

Pearson Correlation Coefficient can be determined by:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

where

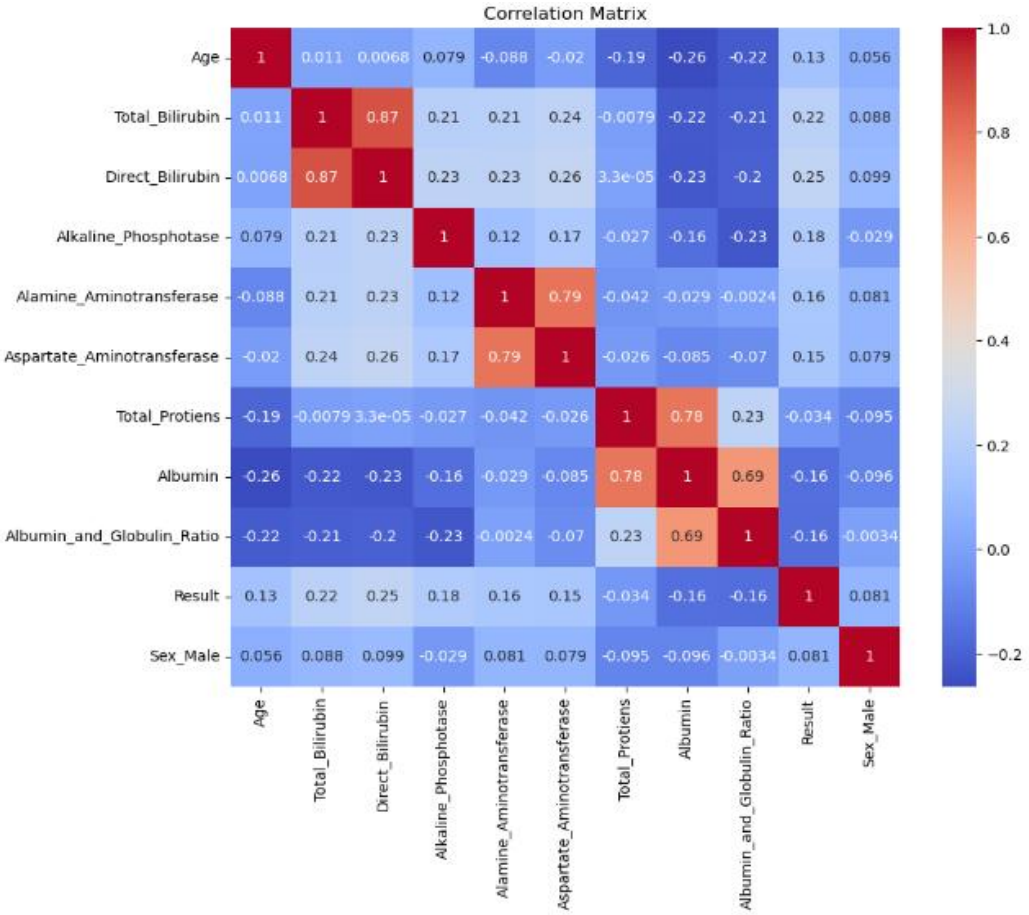
- cov is the covariance
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y.

The features ‘Total Bilirubin’ and ‘Direct Bilirubin’ show POSITIVE COVARIANCE. This shows the POSITIVE direction of relationship between two features.

It is quite evident from the heatmap below that features ‘Total Bilirubin’ and ‘Direct Bilirubin’ features share a positive correlation, the value of which is 0.87. Thus, we can drop one of them for a better predictive modelling.

Heat map to show correlation between all the features.

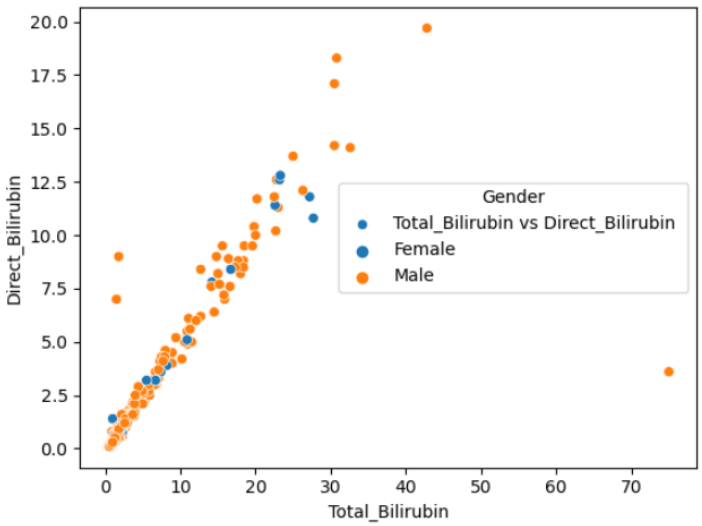
This makes it evident about the correlation of Direct Bilirubin and Total Bilirubin.



For a better model, we take the correlation among Independent Variables not more than 0.80.

Please note, the scatter plot depicts more precisely about the positive correlation between both the features.

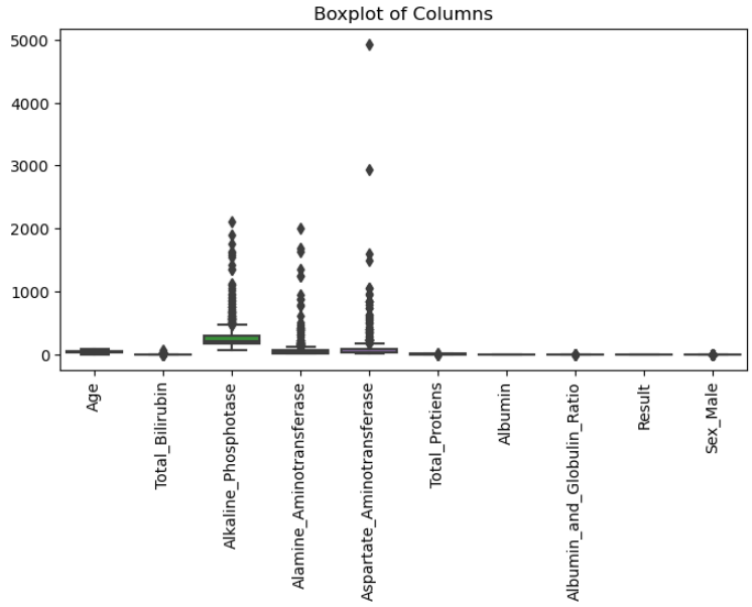
Showing positive relation between both features. Thus, we can drop one of them.



2.4 Outlier Detection:

Outliers are specific exceptional numbers that go outside of the predicted range and contrast with the rest of the data. Machine learning modelling and model performance in general can frequently be improved by comprehending and even getting rid of these outlier values.

Here is the graphical representation of the outliers in our dataset.



There are certain methods to detect these outliers, one of them, which we are incorporating is SD Method.

Standard Deviation Method: The SD is based on the formula. We can identify and exclude outliers from the dataset based on the chosen limit, which can be either 2 times or 3 times standard deviation.

Upper Limit = mean + 3*stdev

Lower Limit = mean + 3*stdev

This way the outliers are detected and have been taken care of for our predictive models.

2.5 Skewness (Measure of Asymmetry)

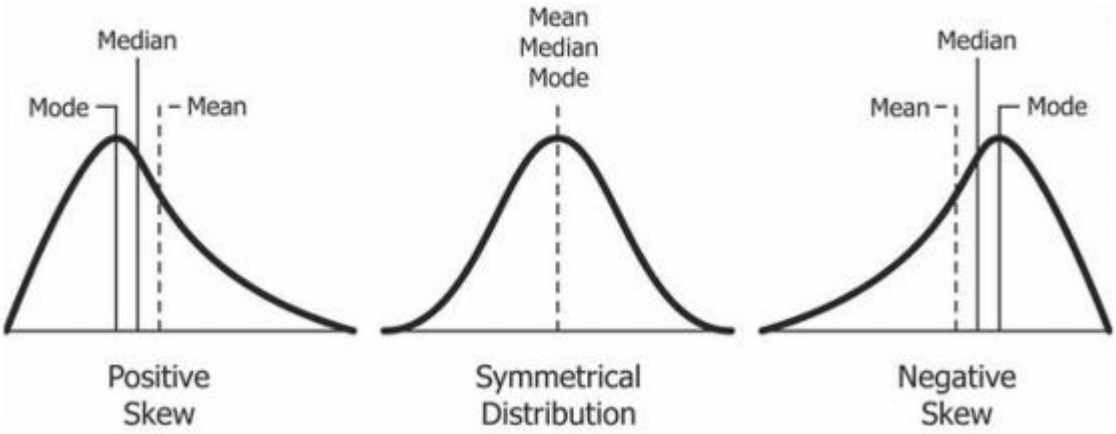
Simply put, a random variable's skewness is a measure of how much its probability distribution deviates from a normal distribution.

Based on the definition, the data can be divided into:

Positive Skewness / Right Skew

Negative Skewness / Left Skew

Our main agenda is to work on a data which is Symmetrically distributed. The image below would make it clearer.



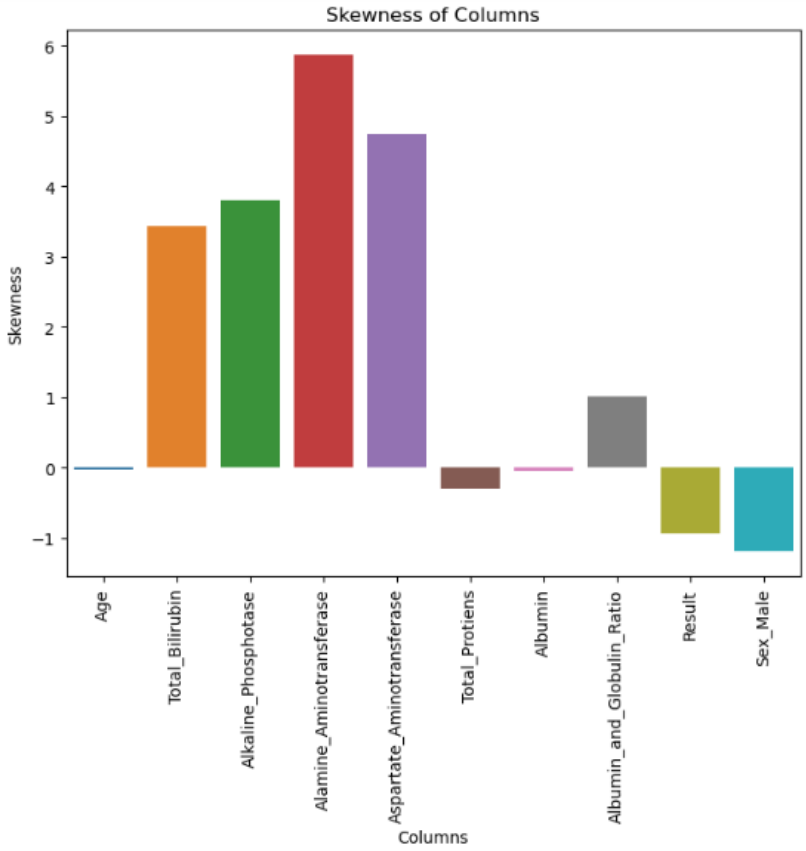
Relation between Mean, Median & Mode based on skewness:

Right Skew (Mean > Median > Mode)

Normally Distributed (Mean ~ Median ~ Mode)

Left Skew (Mean < Median < Mode)

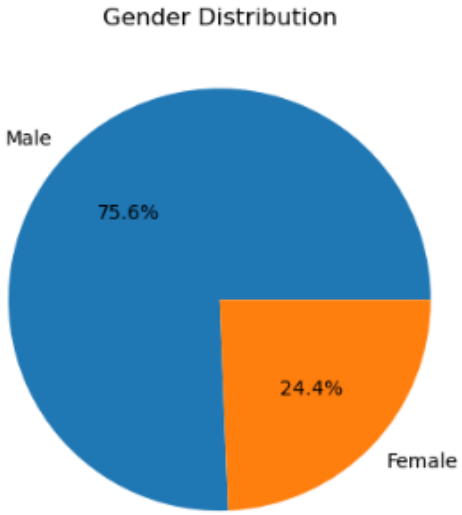
- The bar-chart below shows skewness of our features.
- The features extending above origin show positive skewness.
- Those extending below origin show negative skewness.



Please note: Only Linear and Logistic Models make assumption based on Gaussian Distribution, they assume that the variables are normally distributed.

Other models like Tree-based methods, SVM, do not make assumptions based on normal distribution, thus they are not affected by how the data is distributed.

2.6 Sampling



This method is applied when we have imbalanced dataset. Out dataset is imbalanced as it consists of a smaller number of Female count than Males. This can result in biasness in our machine-learning models.

This issue can be taken care by one of the Sampling techniques, i.e OVER SAMPLING.

Random oversampling duplicates examples from the minority class in the training dataset. The library used is RandomOverSampler. This resulted in the increase in our dataset. Now our dataset looks like:

```
X_res.shape, y_res.shape
```

```
((818, 9), (818,))
```

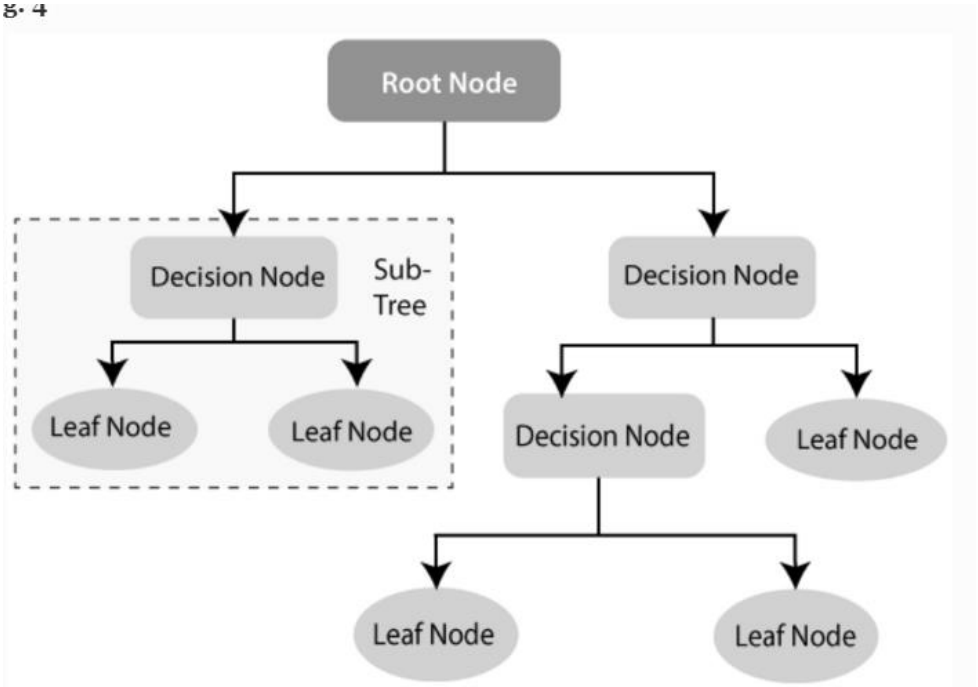
3. Multi-Class Models (Multiple ML Models)

We are now going to check various models to check with model fits best for our dataset for prediction.

3.1 Decision Tree Classifier

A decision tree model primarily consists of nodes and branches, with splitting, stopping, and pruning serving as its primary modelling functions. It uses supervised learning in a non-parametric way. By constructing the tree from the root to a few leaf nodes, the Decision Tree classifies the instances. Instances are categorised by inspecting the attribute defined by that node, starting at the root node of the tree and moving along the branch that corresponds to the

attribute value. The terms "gini" for Gini impurity and "entropy" for information gain are the two most frequently used criterion for splitting.



Mathematically,

$$\text{Entropy} : H(x) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)$$

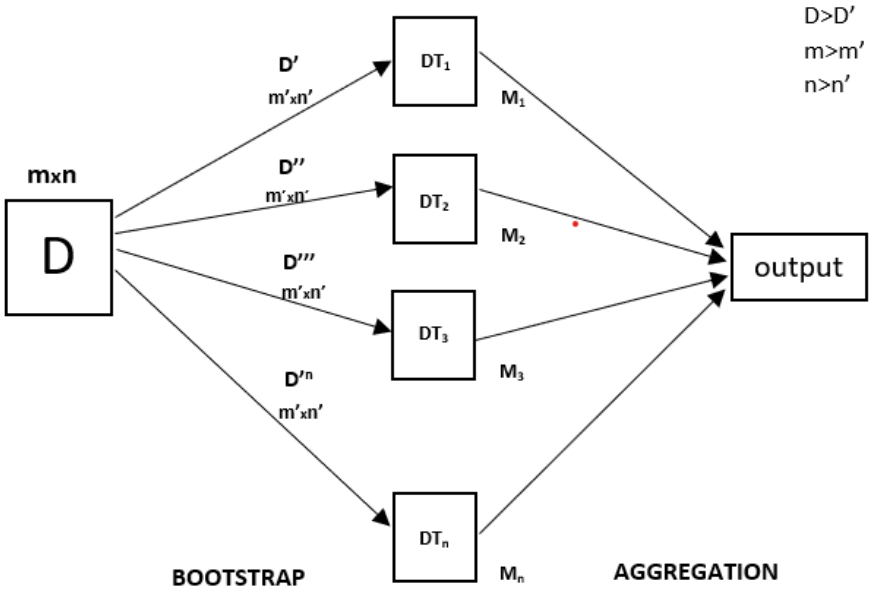
$$\text{Gini}(E) = 1 - \sum_{i=1}^{c} {p_i}^2.$$

A powerful statistical technique for our classification issue is the decision tree. It makes difficult-to-understand and -interpret relationships between the variables simpler. It is robust to outliers and manages heavily skewed data (like ours) with ease.

3.2 Random-Forest Classifier

The Random Forest Classifier bases its operation on the number of Decision Trees that are implemented. A random vector sampled at random and with the same distribution across all of the trees in the forest determines the values of each tree in a random forest. The generalisation error converges as a limit as a forest's number of trees rises. The generalisation error of a forest of tree classifiers is determined by the strength of each individual tree

inside the forest and the correlation between them.



Random Forest Coverage

The generic mathematics behind it is

$$PE^* = PX,Y\ (mg(X,Y) < 0)$$

where the subscripts X, Y indicate that the probability is over the X, Y space.

It produces meaningful internal estimates of error, strength, correlation, and variable relevance; it is somewhat robust to outliers and noise; it is faster than bagging or boosting; it is straightforward and easily parallelized.

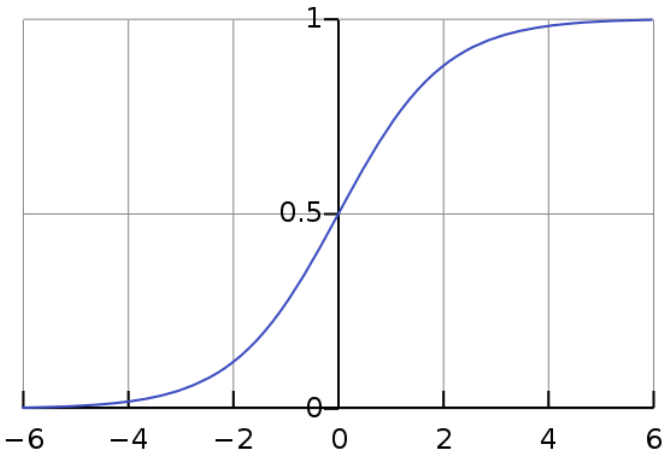
3.3 Logistic Classifier

The basic mathematical concept underpinning logistic regression is the logit, or natural logarithm of an odds ratio. A useful technique for expressing and assessing ideas about correlations between a categorical outcome variable and one or more continuous or categorical predictor variables is logistic regression.

A statistical model that is frequently used to model a binary dependent variable using the logistic function. A sigmoid function is another term for the logistic function, and it is provided by:

$$F(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{e^x + 1}$$

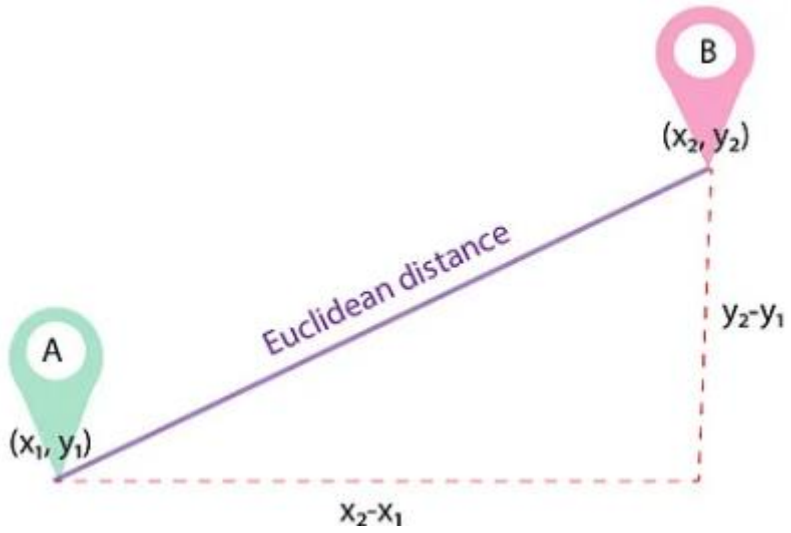
The sigmoid function or S-shaped curve is in the picture.



3.4 KNN (K-Nearest Neighbour)

Finding neighbours is the foundation of the KNN idea. Finding nearest neighbours is defined as locating the point in the given data set that is closest to the input point. Euclidean distance is the separation between the two specified points.

The Euclidean distance can be shown by:



The input x is then assigned to the class with the highest probability once the distance has been calculated:

$$P(y = j|X = x) = \frac{1}{K} \sum_{i \in \mathcal{A}} I(y^{(i)} = j)$$

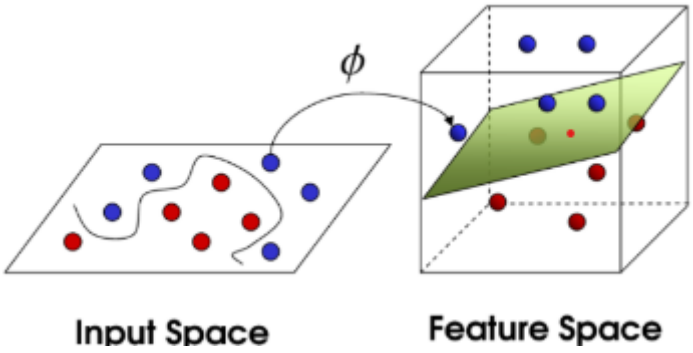
The least distance from a given point, is the time when the value of that unknown point is generalized.

3.5 **SVM (Support-Vector Machine)**

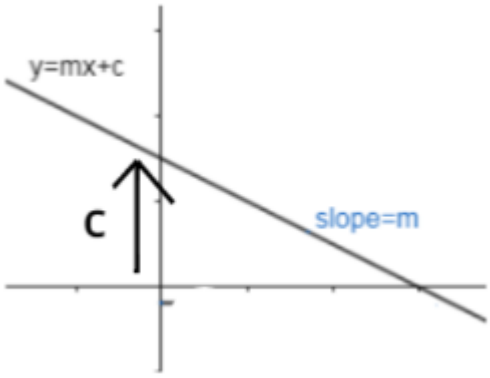
A Support Vector Machine, or SVM, is a machine learning technique that studies data and divides it into two categories.

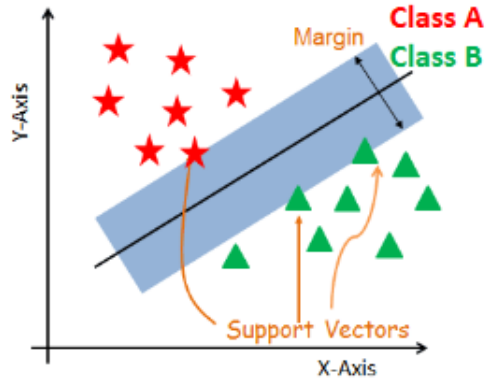
The most common application of the supervised, linear machine learning technology known as Support Vector Classification (sometimes referred to as Support Vector Machine) is to solve classification problems.

When x is specified in a real space, its domain is obvious, and range and co-domain are revealed when y is mapped to a function for f(x). In this instance, we are applying the domain, range, and mapping a function for the data points, but we are utilising a vector space for x rather than actual space.



The equation of the main separator line is called a hyperplane equation.





The goal is to maximize the Euclidean distance for better performance, Comparison and Analysis

$$y_n\left[w^T\phi(x) + b\right] = \begin{cases} \geq 0 \; if \; correct \\ < 0 \; if \; incorrect \end{cases}$$

Comparative findings for the classification methods will be discussed in this section. The purpose and restrictions of categorization algorithms are explained by this comparative examination.

4. Results and Analysis

This research has been conducted to check the performance of the 5 classification methods for our Liver Disease dataset. The Accuracy has evaluated basis (A), Precision (P), Recall (R), and F1-score (F1).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

$$F1 = \frac{2*TP}{2*TP+FP+FN}$$

The formulas for the different metrices have been shown on the left.

Analysing the Accuracies with 5 models

Accuracy With

Decision-Tree

	precision	recall	f1-score	support
0	0.82	0.90	0.86	110
1	0.87	0.77	0.82	95
accuracy			0.84	205
macro avg	0.84	0.83	0.84	205
weighted avg	0.84	0.84	0.84	205

```
[[99 11]
 [22 73]]
```

Accuracy with

Random-Forest Classifier

	precision	recall	f1-score	support
0	0.81	0.87	0.84	110
1	0.84	0.76	0.80	95
accuracy			0.82	205
macro avg	0.82	0.82	0.82	205
weighted avg	0.82	0.82	0.82	205

```
[[96 14]
 [23 72]]
```

Accuracy With

Logistic Classifier

	precision	recall	f1-score	support
0	0.71	0.84	0.77	110
1	0.76	0.60	0.67	95
accuracy			0.73	205
macro avg	0.73	0.72	0.72	205
weighted avg	0.73	0.73	0.72	205

```
[[92 18]
 [38 57]]
```

Accuracy with

KNN

	precision	recall	f1-score	support
0	0.71	0.84	0.77	110
1	0.76	0.60	0.67	95
accuracy			0.73	205
macro avg	0.73	0.72	0.72	205
weighted avg	0.73	0.73	0.72	205

```
[[92 18]
 [38 57]]
```

Accuracy with

SVC

	precision	recall	f1-score	support
0	0.67	0.84	0.74	110
1	0.73	0.52	0.60	95
accuracy			0.69	205
macro avg	0.70	0.68	0.67	205
weighted avg	0.70	0.69	0.68	205

```
[[92 18]
 [46 49]]
```

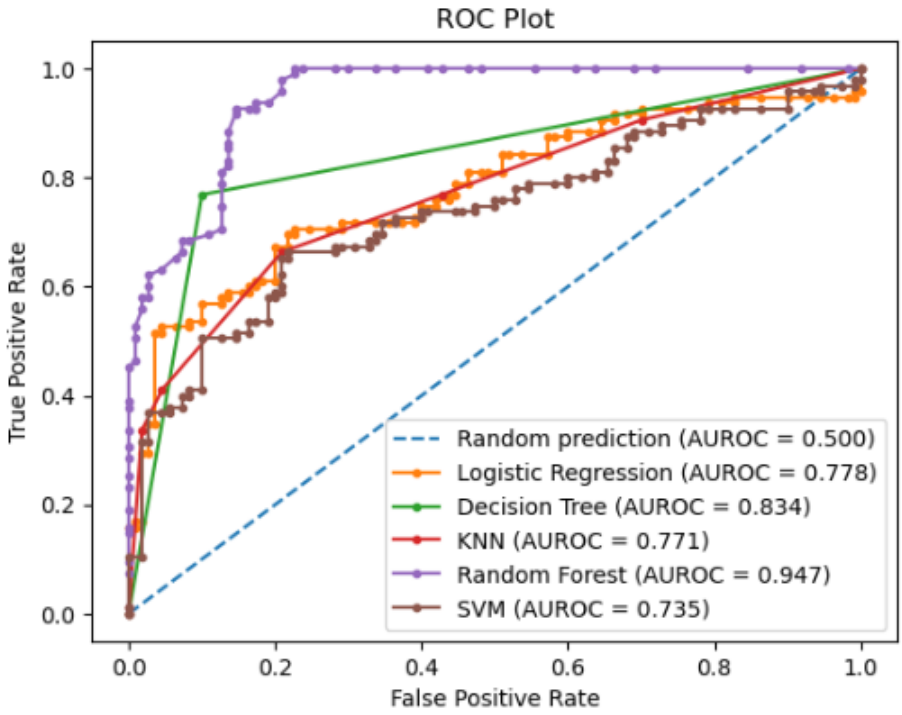
4.1 ROC/AUROC Analysis

Receiver operating characteristic (ROC) graphs are important for classifier organisation and performance visualisation. In addition to being often utilised in medical decision-making, ROC graphs are now being used more frequently in research on machine learning and data mining.

The region between the ROC curve and the diagonal shows how the risk distributions are separated from one another. The area between the ROC curve and the diagonal is larger and the AUC is higher the more space there is between the risk distributions of diseased and unaffected people.

The area under ROC (AUROC) is commonly used for assessing the ability of the prediction models. The more area under the curve means the better that model would perform for analysis.

The Different models are showing different values for AUROC. This helps us to bifurcate among the various models and find the best suited for our problem statement.



4.2 Findings

1. The value of AUROC for Random Forest is 0.947. This shows that this model provides a High Positive Rate.
2. The False Positive Rate is less in this model.
3. This also implies that the Random Forest Classifier, for out dataset, has a strong discriminatory power.
4. It is also effective when it comes down to correct classification of instances from both classes.
5. This proves high degree of confidence for analysis and accurate predictions.

5. Conclusion

The model with the greatest Area under ROC turned out to be the Random Forest Classifier. Given that the area under the ROC curve is 0.94, this model is the most effective for predicting and diagnosing Chronic Liver Diseases early on. It still depends on the circumstances surrounding the data collection. If the data were accurate, there may be situations when alternate models would function just fine. Chronic liver disease can be identified by clinicians who are adept at identifying notable findings and classifying them as normal or abnormal using background information and other context clues. Similar to how machine learning algorithms may aid medical personnel, ML algorithms can be trained to recognise the potential for liver disease.

By using the correlation of each variable with the risk of liver disease to train the model, ML methods may discriminate between blood donors with and without liver disease with high accuracy.

Many of the constraints in healthcare related to inaccurate diagnoses, missing data, expense, and time can be lessened by using these techniques. By increasing awareness of risk factors and diagnostic indicators, the use of ML techniques can contribute to a reduction in the overall burden of liver disease on public health globally.

References

Indian Liver Patient Records | Kaggle

https://pubmed.ncbi.nlm.nih.gov/35470133/

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8867112/

https://fg.bmj.com/content/13/5/367

https://www.mdpi.com/2673-4389/1/4/23

Imputation (statistics) - Wikipedia

http://www.stat.yale.edu/Courses/1997-98/101/catdat.htm

https://www.researchgate.net/publication/343487307_A_Survey_on_machine_learning_techniques_for_the_diagnosis_of_liver_disease

Handling Categorical Features - With Examples | kaggle_tutorials – Weights & Biases (wandb.ai)

How to Handle Categorical Features | by Ashutosh Sahu | Analytics Vidhya | Medium

https://machinelearningmastery.com/how-to-use-statistics-to-identify-outliers-in-data/

https://towardsdatascience.com/outlier-detection-part1-821d714524c

https://www.mdpi.com/2673-4389/1/4/23

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4466856/

https://link.springer.com/article/10.1023/a:1010933404324

https://www.analyticsvidhya.com/blog/2020/10/the-mathematics-behind-svm/

https://link.springer.com/article/10.1007/s42979-021-00592-x