**SMARTBRIDGE INTERNSHIP**

**LIVER-PATIENT ANALYSIS**

**Submitted by**

Abirami.S

Nithyashee.S

1. **INTRODUCTION**

1.1 OVERVIEW

There are many disorders of the liver that require clinical care by a physician or other healthcare professional. Diagnosis of liver disease at a preliminary stage is important for better treatment. It is a very challenging task for medical researchers to predict the disease in the early stages owing to subtle symptoms. Often, the symptoms become apparent when it is too late. To overcome this issue, this project aims to improve liver disease diagnosis using machine learning approaches.

AI Machine Learning is the scientific study of algorithms and statistical models that computer systems use in order to perform a specific task effectively without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as “training data”, in order to make predictions or decisions without being explicitly programmed to perform the task.

Machine learning tasks are classified into several broad categories. In supervised learning, the algorithms build a mathematical model from a set of data that contains both the inputs and desired outputs. Classification algorithms and regression algorithms are types of supervised learning. Classification algorithms are used when the outputs, are restricted to a limited set of values. Regression algorithms are named for their continuous outputs, meaning they may have any value within a range. In unsupervised learning, the algorithm builds a mathematical model from a set of data which contains only inputs and no desired output labels. Unsupervised learning can discover patterns in the data, and can group the inputs into categories.

1.2 PURPOSE

In India, delayed diagnosis of disease is a fundamental problem due to a shortage of medical professionals. An early diagnosis of liver problems will increase patient’s survival rate.

Therefore, developing a machine that will enhance in the diagnosis of the disease will be of a great advantage in the medical field. These systems will help the physicians in making accurate decisions on patients and also with the help of Automatic classification tools for liver diseases, one can somewhat reduce the time delay.

1. **LITERATURE SURVEY**

2.1 EXISTING PROBLEM

The Indian Liver Patient Dataset comprised of 10 different attributes of 583 patients. The patients were described as either 1 or 2 on the basis of liver disease. The detailed description of the dataset is shown in Table.

|  |  |  |
| --- | --- | --- |
| **No.** | **ATTRIBUTES** | **ATTRIBUTE TYPE** |
| 1. | Age | Numeric |
| 2. | Gender | Textual |
| 3. | Total Bilirubin | Numeric |
| 4. | Direct Bilirubin | Numeric |
| 5. | Alkaline Phosphatase | Numeric |
| 6. | Alamine Phosphatase | Numeric |
| 7. | Total Proteins | Numeric |
| 8. | Albumin | Numeric |
| 9. | Albumin and Globulin Ratio | Numeric |
| 10. | Result | Numeric (1,2) |

This dataset seems to be a classic example of supervised learning. We have been provided with a fixed number of features for each data point, and our aim will be to train a variety of Supervised algorithms on this data, so that, when a new data point arises, our best performing classifier can be used to categorize the data point as a positive example or negative.

2.2 PROPOSED SOLUTION

The main objective of this research is to use classification algorithms to identify the liver patients from healthy individuals. This project also aims to compare the classification algorithms based on their performance factors.

To serve the medicinal community for the diagnosis of liver disease among patients, a graphical user interface is developed. The GUI can be readily utilized by doctors and medical practitioners as a screening tool for the liver disease.

The following algorithms are trained for this purpose, to choose the best performing classifier.

* Logistic Regression
* Naïve Bayes
* K-Nearest Neighbors
* SVM
* Decision Tree Classifier
* Random Forest

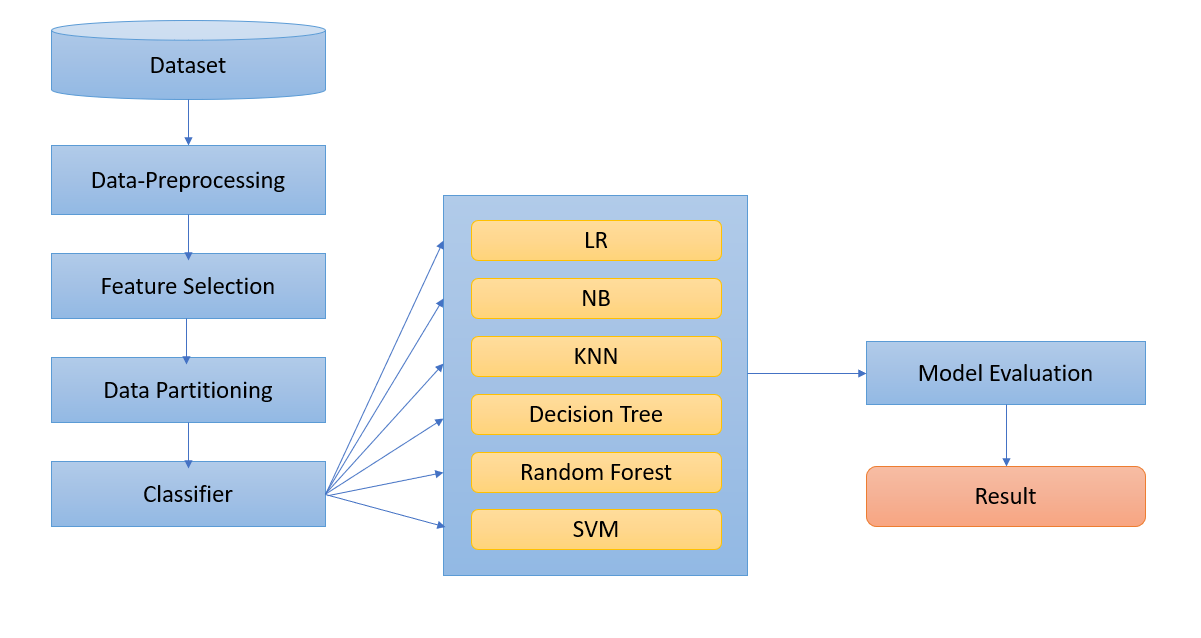
The results of each of the classification algorithm is summarized in the table shown below.

|  |  |
| --- | --- |
| **Model** | **Accuracy Score** |
| Logistic Regression | 0.754902 |
| SVM | 0.725490 |
| Decision Tree Classifier | 0.715686 |
| Random Forest | 0.686275 |
| KNN | 0.66667 |
| Naïve Bayes | 0.529412 |

As clearly summarized in the table, Logistic Regression gave the best result.

1. **THEORITICAL ANALYSIS**

3.1 BLOCK DIAGRAM



3.2 HARDWARE/SOFTWARE DESIGNING

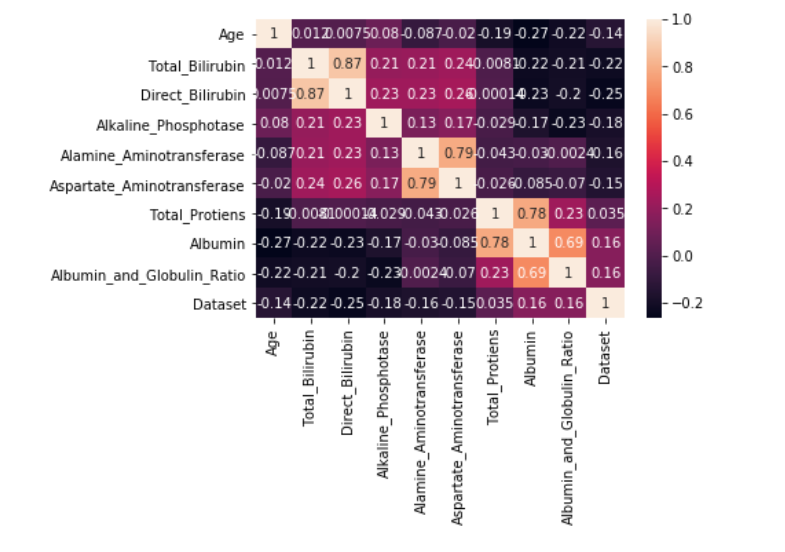
The steps followed in developing the model:

* Data Collection: The dataset was downloaded from the UCI ML Repository.
* Data Analysis: Evaluating cleanliness of the dataset by looking for any irrelevant data and handling missing data.
* Search for any trends, relations and correlations.
* Developing a GUI where the patient can be identified to be having liver disease or not.

1. **EXPERIMENTAL INVESTIGATIONS**

CORRELATION GRAPH

Correlation is an indication about the changes between two variables. We can plot correlation matrix to show which variable is having a high or low correlation in respect to another variable.

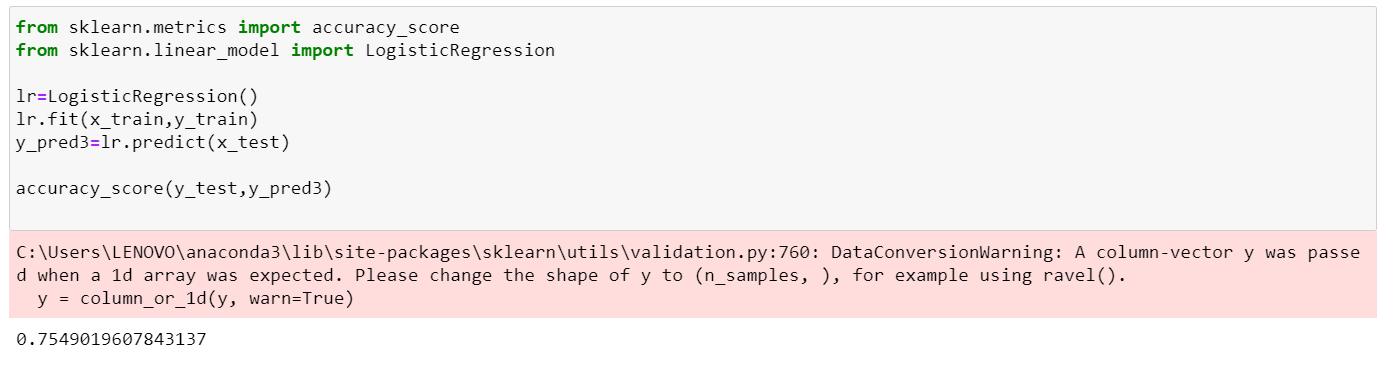


From the above output of correlation matrix, we can see that it is symmetrical i.e. the bottom left is same as the top right. It is also observed that each variable is positively correlated with each other.

1. Logistic Regression

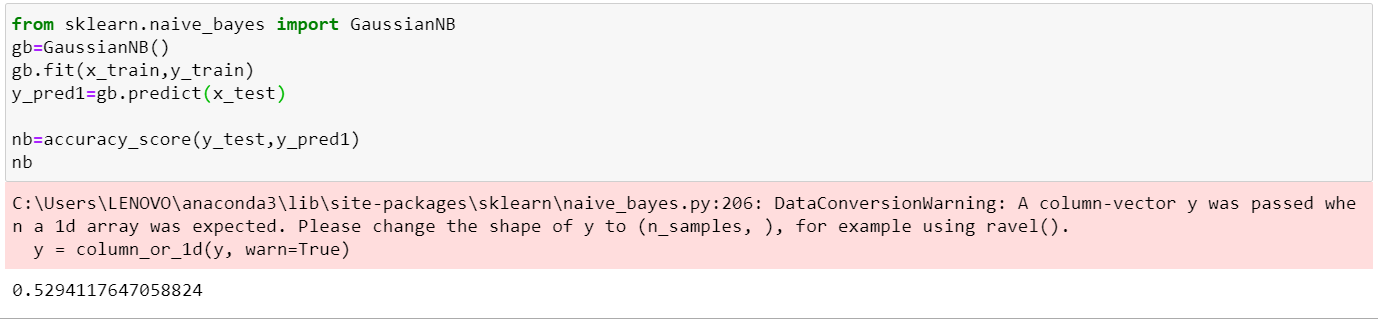
Since the outcome is binary and we have a reasonable number of examples at our disposal compared to number of features, this approach seems suitable. Since for this data, it already knows the output beforehand, it continuously adjusts the weights such that when these weights summed up with their features are introduced in the logistic function, the results are as near as possible to the actual ones.

Once presented with a test value, it again inserts the value into our logistic function and returns the output as a number between 0 and 1, which represents the probability of that test value being in a particular class.



1. Naïve Bayes

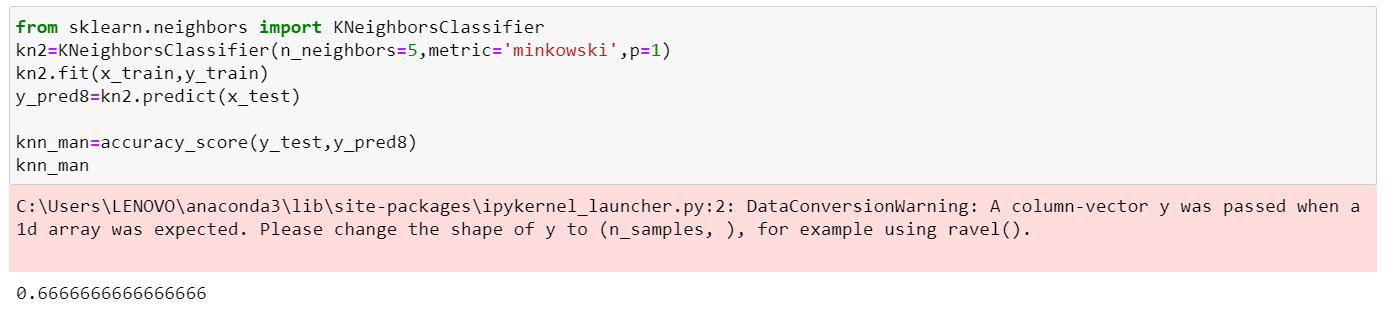
In [statistics](https://en.wikipedia.org/wiki/Statistics), **Naïve Bayes classifiers** are a family of simple "[probabilistic classifiers](https://en.wikipedia.org/wiki/Probabilistic_classification)" based on applying [Bayes' theorem](https://en.wikipedia.org/wiki/Bayes%27_theorem) with strong (naïve) [independence](https://en.wikipedia.org/wiki/Statistical_independence) assumptions between the features. They are among the simplest [Bayesian network](https://en.wikipedia.org/wiki/Bayesian_network) models. Naïve Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables in a learning problem.



1. K-Nearest Neighbors

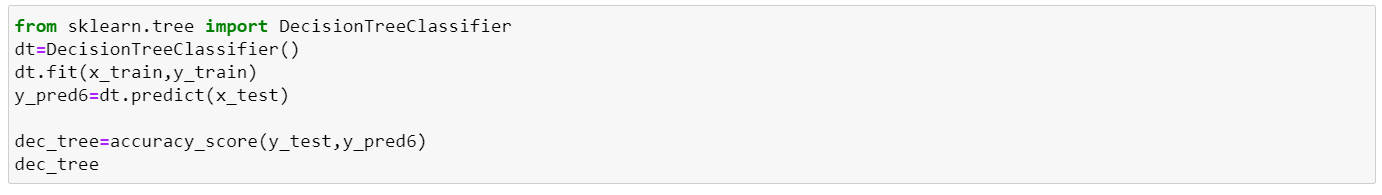
In [pattern recognition](https://en.wikipedia.org/wiki/Pattern_recognition), the **k-nearest neighbors algorithm** (***k*-NN**) is a [non-parametric](https://en.wikipedia.org/wiki/Non-parametric_statistics) method proposed by Thomas Cover used for [classification](https://en.wikipedia.org/wiki/Statistical_classification) and [regression](https://en.wikipedia.org/wiki/Regression_analysis). In both cases, the input consists of the *k* closest training examples in the [feature space](https://en.wikipedia.org/wiki/Feature_space).

In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its *k* nearest neighbors (*k* is a positive [integer](https://en.wikipedia.org/wiki/Integer), typically small). If *k* = 1, then the object is simply assigned to the class of that single nearest neighbor.



1. Decision Tree Classifier

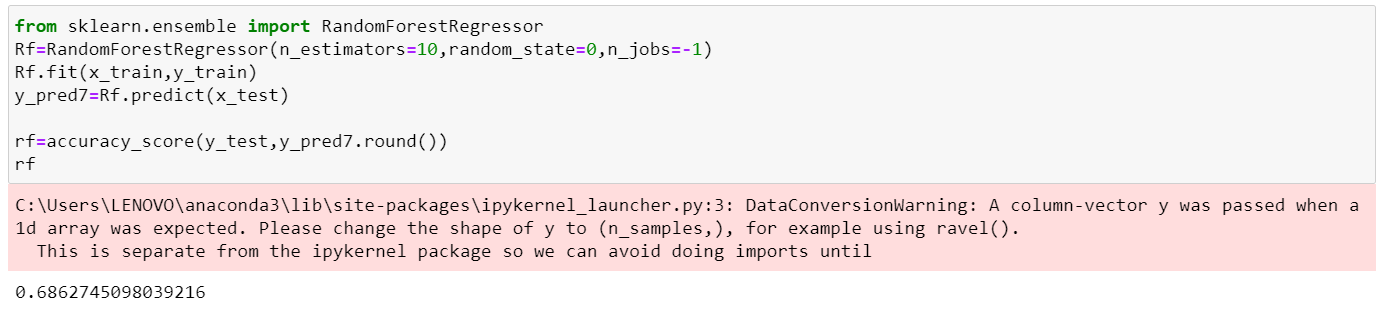
**Decision Trees (DTs)** are a non-parametric supervised learning method used for [classification](https://scikit-learn.org/stable/modules/tree.html#tree-classification) and [regression](https://scikit-learn.org/stable/modules/tree.html#tree-regression). The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.



0.715686

1. Random Forest

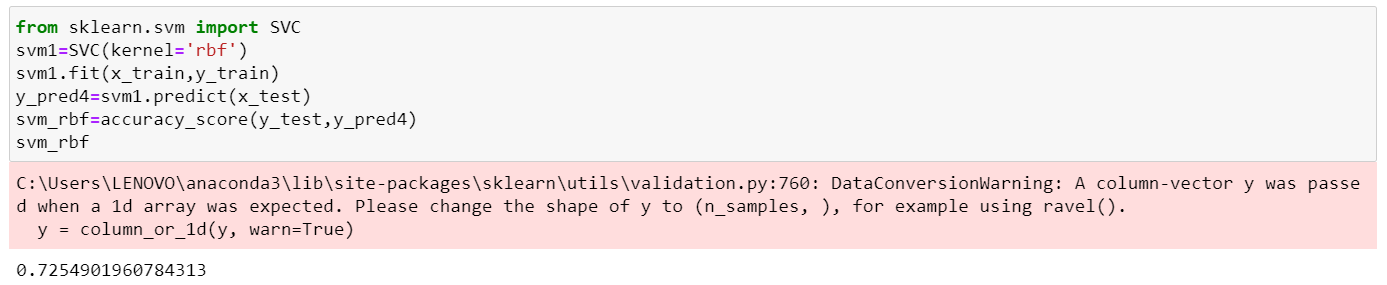
**Random forests** or **random decision forests** are an [ensemble learning](https://en.wikipedia.org/wiki/Ensemble_learning) method for [classification](https://en.wikipedia.org/wiki/Statistical_classification), [regression](https://en.wikipedia.org/wiki/Regression_analysis) and other tasks that operate by constructing a multitude of [decision trees](https://en.wikipedia.org/wiki/Decision_tree_learning) at training time and outputting the class that is the [mode](https://en.wikipedia.org/wiki/Mode_(statistics)) of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of [overfitting](https://en.wikipedia.org/wiki/Overfitting) to their [training set](https://en.wikipedia.org/wiki/Test_set).



1. SVM

In [machine learning](https://en.wikipedia.org/wiki/Machine_learning), **support-vector machines** (**SVMs**, also **support-vector networks**) are [supervised learning](https://en.wikipedia.org/wiki/Supervised_learning) models with associated learning [algorithms](https://en.wikipedia.org/wiki/Algorithm) that analyze data used for [classification](https://en.wikipedia.org/wiki/Statistical_classification) and [regression analysis](https://en.wikipedia.org/wiki/Regression_analysis).

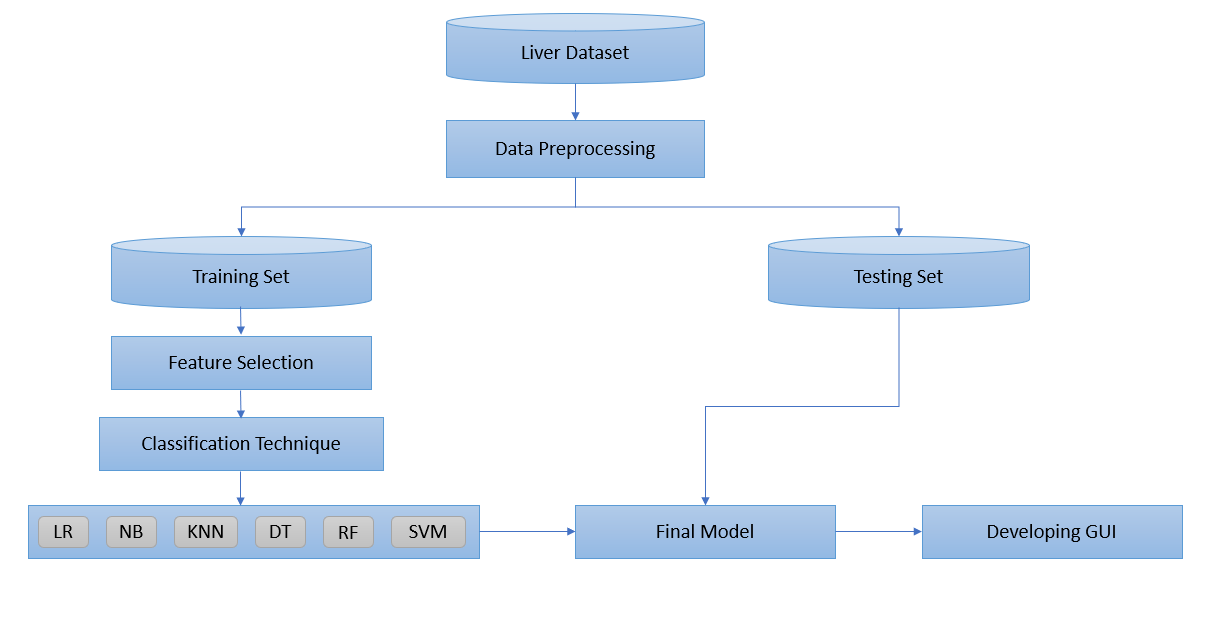
The Support Vector Machine (SVM) algorithm is a popular machine learning tool that offers solutions for both classification and regression problems.



SUMMARY

|  |  |
| --- | --- |
| **Model** | **Accuracy Score** |
| Logistic Regression | 0.754902 |
| SVM | 0.725490 |
| Decision Tree Classifier | 0.715686 |
| Random Forest | 0.686275 |
| KNN | 0.66667 |
| Naïve Bayes | 0.529412 |

1. **FLOWCHART**



1. **RESULT**

Based on the inputs entered by the user, the model predicts whether the patient has liver disease or not.

1. **ADVANTAGES AND DISADVANTAGES**

The benefits of this model are:

* Easy interface
* Straight forward results
* Accurate performance calculations

Disadvantages:

As our dataset is small, it’s training dataset is similar to test dataset. So it is difficult for the model to predict accurately for larger dataset.

1. **APPLICATIONS**

This project makes it easier to predict whether the patient has liver disease or not.

1. **CONCLUSION**

In this project, we have proposed methods for diagnosing liver disease in patients using machine learning techniques. The six machine learning techniques that were used include SVM, Logistic Regression, KNN, Naïve Bayes, Decision Tree Classifier and Random Forest. The system was implemented using all the models and their performance was evaluated. Performance evaluation was based on certain performance metrics. Logistic Regression was the model that resulted in the highest accuracy with an accuracy of 75%. A GUI, which can be used as a medical tool by hospitals and medical staff was implemented using Logistic Regression.

1. **FUTURE SCOPE**

Database should be expanded on which the system will be tested much better. Also, the model requires further improvement mostly regarding feature selection of the liver into multiple components.

1. **BIBLIOGRAPHY**

* <https://towardsdatascience.com/ways-to-detect-and-remove-the-outliers-404d16608dba>
* <https://www.youtube.com/watch?v=KFuEAGR3HS4&t=970s>