AI-Statistical Machine Learning Approaches to Liver Disease Prediction

Team ID: PNT2022TMID48272

Faculty Mentor:

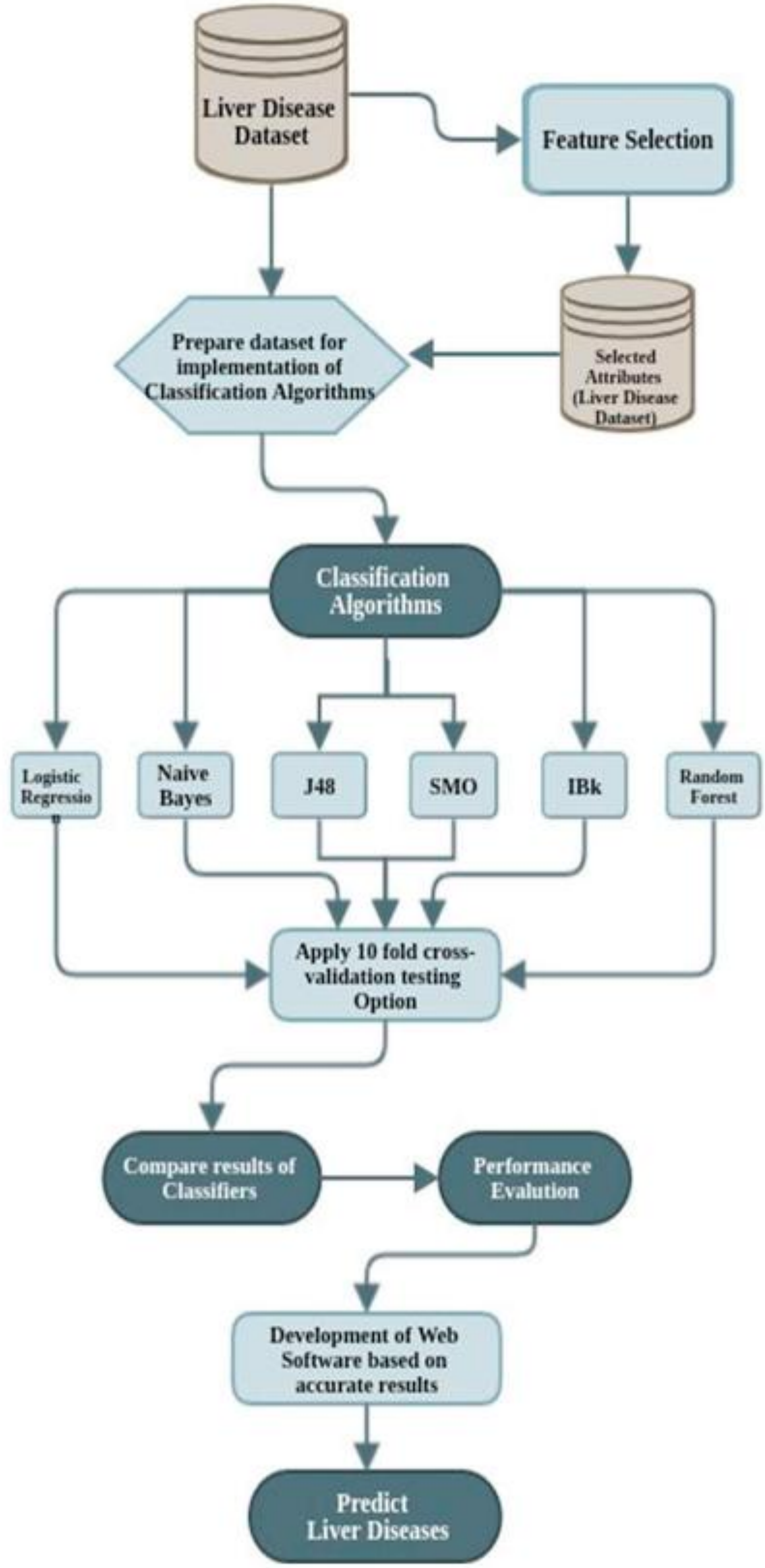
D.**Pradhiba**

Team Leader: G.lydia
Team Member: R.priya
Team Member: U.lavanya sri
Team Member: G.nagalakshmi

Solution Architecture

Today's health care is very important aspect for every human, so there is a need to provide medical services that are easily available to everyone. In this paper, the main focus is to predict the liver disease based on a software engineering approach using classification and feature selection technique. The implementation of proposed work is done on Indian Liver Patient Dataset (ILPD) from the University of California, Irvine database. The different attributes like age, direct bilirubin, gender, total bilirubin, Alkphos, sgpt, albumin, globulin ratio and sgot etc, of the liver patient dataset, are used to predict the liver diseases risk level. The various classification algorithms such as Logistic Regression, SMO, Random Forest algorithm, Naive Bayes, J48 and k-nearest neighbor (IBk) are implemented on the Liver Patient dataset to find the accuracy. The comparison different classifier results are done of feature selection and without using feature selection technique. The development of intelligent liver disease prediction software (ILDPS) is done by using feature selection and classification prediction techniques based on software engineering model. This disease contains many conditions like inflammation (hepatitis B, C) from infectious, non-infectious causes (chemical or autoimmune hepatitis), Tumors, malignant and scarring of

the liver (Cirrhosis) and Metabolic Disorders. In this paper, six classification algorithms J-48, Random Forest, Logistic Regression, SMO (Support Vector Machine), IBk (k nearest Neighbor), Logistic Regression, Naive Bayes have been considered for implementation and comparing their results based on the ILPD (Indian Liver Patient Dataset) Software's developed for the healthcare plays a crucial role in delivering efficient services to the physicians and ultimately better treatment can be offered to the patients. An efficient healthcare software can assist in several activities like forecasting of the diseases based on the historical data of some another patient, image processing of medical images, a data warehouse for management of the whole institution etc. Proposed work focuses on the development of the software that will help in the prediction of the level diseases based upon the various symptoms. The development stage of the given software includes continuous interaction with the physicians so that more accurate results can be generated. The data may contain redundant and irrelevant attributes, there is a need to remove these attributes withoutdecreasing the accuracy using a feature selection technique. In the present work, the WEKA tool is used for the implementation of a feature selection technique. There is a number of feature selection methods available in the WEKA environment. Feature selection can provide us with several benefits like reducing lifting, reducing training time and improving accuracy etc. In Weka tool, the feature evaluator and search method were used to perform feature selection on any dataset. In our proposed work the Correlation-based Feature Selection Subset Evaluator was used as Feature evaluator and Greedy Stepwise used as search method Based on above techniques these 5 features/attributes are select


Liver Disease Dataset
Feature Selection
Selected Attributes (Liver Disease Dataset)
Prepare dataset for implementation of Classification Algorithms
Classification Algorithms
Logistic Regressio n
Naive Bayes
J48
SMO
IBk
Random Forest
Apply 10 fold cross-validation testing Option
Compare results of Classifiers
Performance Evalution
Development of Web Software based on accurate results
Predict Liver Diseases

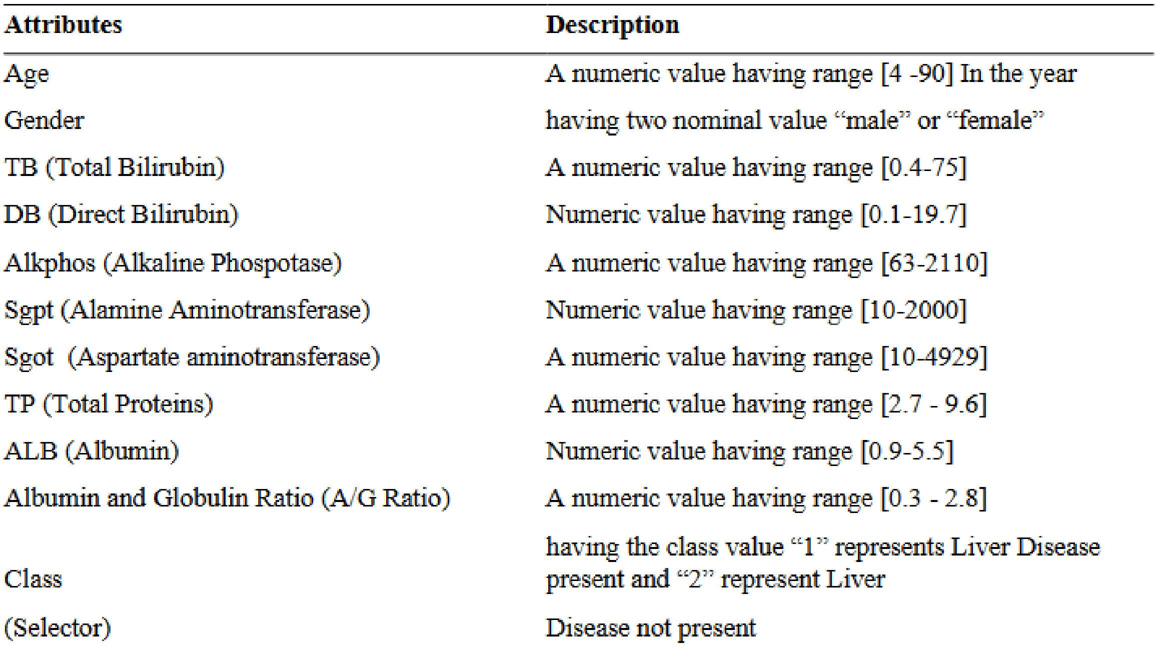
In their research work authors have worked on doing an analysis of the data related to Liver Disorder with the help of Naive Bayes, Decision Table, and J48. However, attributes like case history of the patient, diabetes, smoking, obesity, alcohol intake, smoking etc were used. Based upon the given database it has concluded that male people are having more liver disorder than the females. Age group of 35-65 is mostly affected and out of these 26% people are having the disorder because of alcohol, smoking contributed to 22% of people, obesity, and diabetic of 4 & 5 percent respectively. A. Gulia et al. in their proposed work researchers have done classification of the liver patient data using the algorithms like Bayesian Network, Support Vector Machine, J48, Multi-Layer Perceptron and Random Forest. The data from the UCI repository which is afforded by Center of Machine Learning and Intelligent Systems has used. After completion of their three-phase analysis, the Random Forest Algorithm is the bestone with an accuracy of 71.87% has been concluded. Y. Kumar et al. [12] in their proposed work researchers have used Rule-Based Classification Model (RBCM) for the prediction of liver diseases. Without the rule-based classification the efficiency of all the common algorithm decreases was analyzed. In their proposed work 20 rules were used for the classification of liver diseases. The decision tree-based algorithm gives the best performance using rule-based classification and accordinglyits accuracy decreases when rule-based is not used. M. Pasha et al. work on the dataset from the UCI repository is used which is having 583 instances, the meta-learning algorithms like Grading, logit boost, Adaboost, and Bagging were used. The comparisons of the algorithms based upon the amount of correct and incorrect classifications and time of execution have done. After doing detailed analysis the grading is the best algorithm in terms of accuracy and execution time have been concluded. M. Abdar et al. [14] in their research work focuses on the early prediction ofliver diseaseusing Multilayer Perceptron Algorithm (MLPNN) which uses (CART) (classification and regression tree, (CHAID) Chi-square

Automatic interaction detector, See5(C5.0). Their dataset is from UCI repository of the University of California, Irvine relevant to Indian Liver Patient Dataset (ILPD). From their results, it can be concluded that MLPNNB-CHAID is the best algorithm with an innovative accuracy of 14.57%. The 70% of the data as a training data and rest of the 30% for the testing stage were used. Author EI-Shafeiy at al. [15] in their research work focuses on electronic health records, metabolomics analyses are some of the digital information related to the health and with the passage of time, the shape of Big Data has been taken. In the present work dataset having 23 attributes of 7000 patients with 5295 as male and rest as female is used. Their proposed work make use of Boosted C5.0, Support Vector Machine, Naïve Bayes (NB) along with the feature selection. Vijayarani et al. [16] in their research paper classification algorithms are used for the prediction of liver diseases. Famous algorithms like Naïve Bayes and Support Vector Machine (SVM) are used in the proposed work. The dataset from the UCI repository and it is having fields like Gender, Sgot, ALB, ALP, DB etchave taken. Based upon their present work SVM is best in terms of accuracy and Naïve Bayes is good in terms of execution time. Soegijoko et al. in their proposed work have developed e-Health system for the improvement of community health care system. The use of various Open source software elements like Gammu as SMS engine, MySQL for database handling and Apache as a web server are used. The authors proposed system classifies the mother and child care activities as automatic message generation for the first-trimester pregnancy, second and third -trimester pregnancy, post-partum (after delivery), 40 days after delivery etc. Implementation issues like increasing electricity power, infrastructure preparations, health providers, education and training, technical hardware/software problems etc has highlighted. Their initiative for the paperless prescription system helps in reducing the paperusage is very effective. Skevoulis et al. in their proposed work engineered a software which calculates the Coronary Heart disease (CHD) risk factor. It is a joint effort between the Pfizer Inc's and Seidenberg School of Computer Scienceat Pace University (SCSIS). Their proposed work provides the choice of selecting three CHD risk algorithms, a customized health report and a progress chart. Their proposed work makes use of client-server-based model and the content is dynamically generated using ASP.NET programing language and C#, Microsoft SQL server 2000 and Crystal reports for generating the health reports, ChartFX for representing progress chart. Their proposed work also makes use of JAVA and JavaScript for displaying CHD results. Weber-Jahnke in their proposed work discusses various challenges related to Software engineeringin Health Care (SEHC). Their proposed work requires the community to adopt knowledge translation (KT) for better resultgeneration by the software systems. Their recommendation of using the KT is based upon the survey which highlights that number of software engineering tools, methods, processes that works perfectly for other domains cannot be easily transferred to the healthcare applications. Patent Dataset The total number of 583 instances/rows with eleven distinct attributes were composed from the (ILPD) Indian Liver Patient Dataset from the UCI Repository [17] to solve the purpose of this paper. The attribute "class" described as the disease with value "1" mean a person having liver disease present and "2" represent liver disease not present. Table 1 shows the description of attributes, values of liver disease database. The database having 416 with a patient having liver diseases and 167 are non-liver diseases instances .

Result and Discussion:

The results of the proposed work have been obtained by implementing feature selection techniquesand different classifiers. The judgment of different results was done based on the following categories. First, the results of different classifiers are compared the on basis of correctly classified instances with feature selection techniques shown in table 3 and without using feature selection techniques shown in table Secondly, the para meterslike Kappa statistic value, mean absolute error are compared using on 10-fold cross-validation testing option.

Finally, the execution time of different classifiers also compares for both the above techniques shown in table4. The table 2 show the results of different classifiers without using feature selection technique. Logistic Regression shows the higher accuracy value 72.50 and IBk (k nearest Neighbor) least value 64.15 percentage. The table 3 shows the results of different classifiers using feature selection technique. Using this technique Random Forest shows the higher accuracy value 74.36 and IBk(k nearest Neighbor) least value 67.41 percentage.

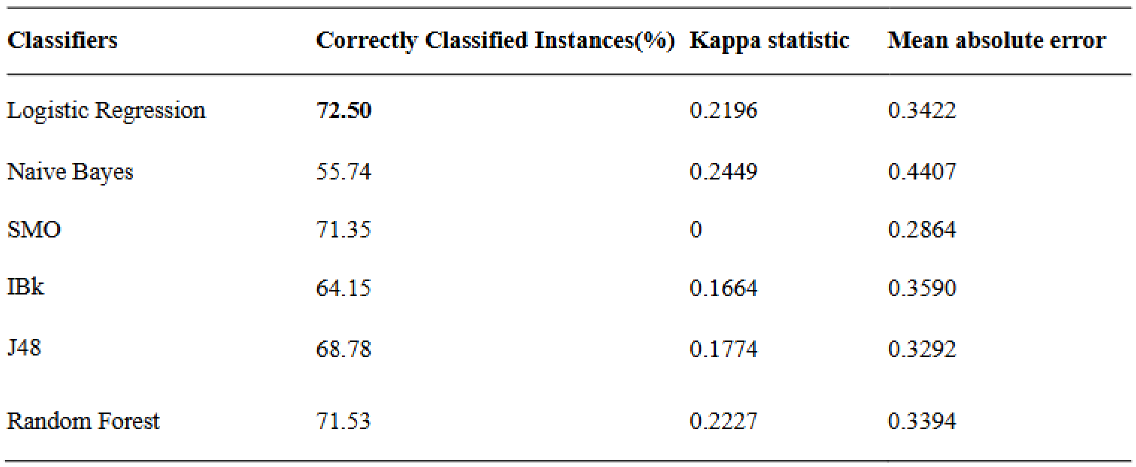


Attributes	Description
Age	A numeric value having range [4 -90] In the year
Gender	having two nominal value "male" or "female"
TB (Total Bilirubin)	A numeric value having range [0.4-75]
DB (Direct Bilirubin)	Numeric value having range [0.1-19.7]
Alkphos (Alkaline Phospotase)	A numeric value having range [63-2110]
Sgpt (Alamine Aminotransferase)	Numeric value having range [10-2000]
Sgot (Aspartate aminotransferase)	A numeric value having range [10-4929]
TP (Total Proteins)	A numeric value having range [2.7 - 9.6]
ALB (Albumin)	Numeric value having range [0.9-5.5]
Albumin and Globulin Ratio (A/G Ratio)	A numeric value having range [0.3 - 2.8]
Class	having the class value "1" represents Liver Disease present and "2" represent Liver
(Selector)	Disease not present

Database Selection:

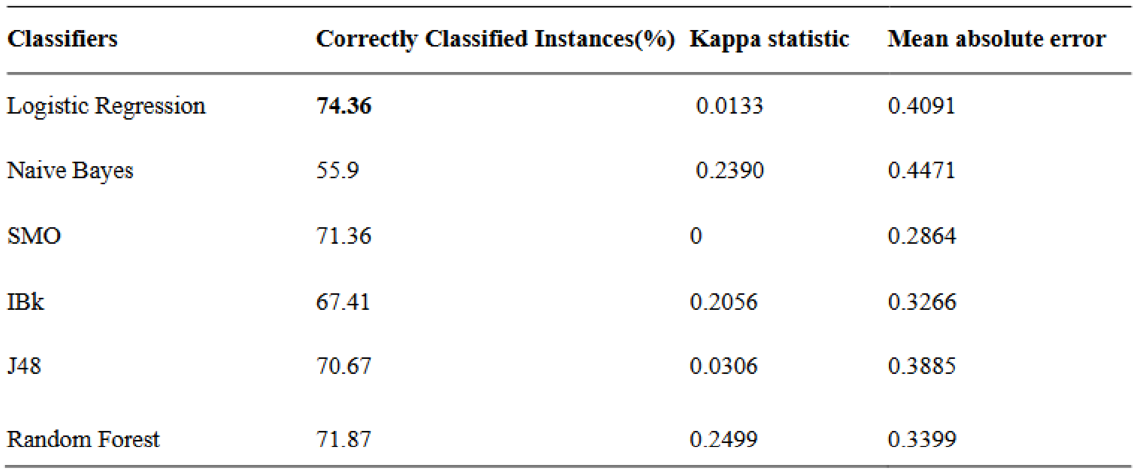
The liver patient dataset was prepared in ARFF (Attribute Relation File Format) file format. The data was processed into the correct format for implementation of different classifiers. Also required to remove an unnecessary field, missing records and duplicate record sifany to prepare data setina standard format. B. Feature Selection: The liver patient dataset was prepared in ARFF (Attribute Relation File Format) file format. The data was processed into the correct format for implementation of different classifiers. Also required to remove an unnecessary field, missing records and duplicate records if any to prepare data setina standard format.

Table 2: Compare the results different classifiers on Liver Patient Diseases dataset



Classifiers	Correctly Classified Instances(%)	Kappa statistic	Mean absolute error
Logistic Regression	**72.50**	0.2196	0.3422
Naive Bayes	55.74	0.2449	0.4407
SMO	71.35	0	0.2864
IBk	64.15	0.1664	0.3590
J48	68.78	0.1774	0.3292
Random Forest	71.53	0.2227	0.3394

Table 3: Compare the results of different classifiers using feature selection technique on Liver Patient Diseases Dataset



Classifiers	Correctly Classified Instances(%)	Kappa statistic	Mean absolute error
Logistic Regression	**74.36**	0.0133	0.4091
Naive Bayes	55.9	0.2390	0.4471
SMO	71.36	0	0.2864
IBk	67.41	0.2056	0.3266
J48	70.67	0.0306	0.3885
Random Forest	71.87	0.2499	0.3399

Compare the results of different classifiers on basis of Correctly Classified Instances

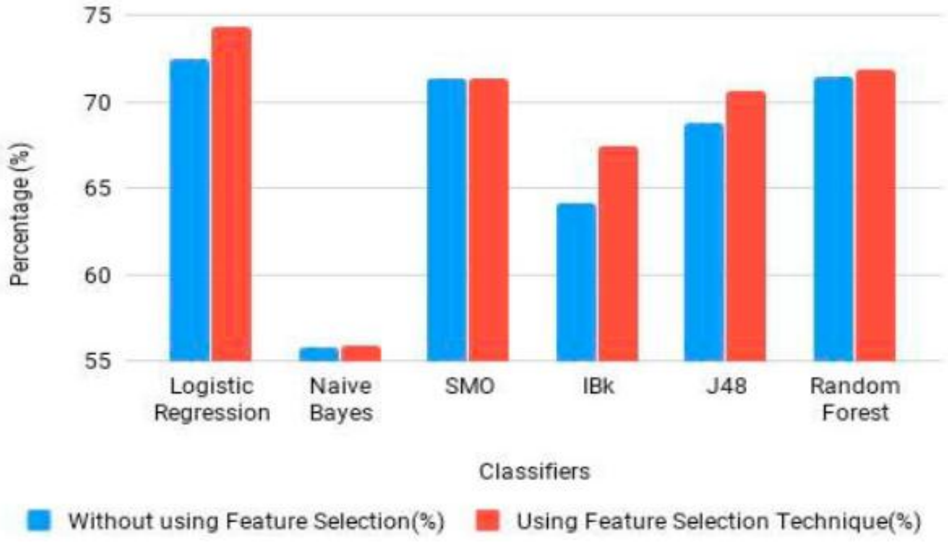


Table 4: Comparison of different classifiers based on execution time

Classifiers	**Execution time without using Feature Selection (in a sec)**	**Execution time using Feature Selection (in a sec)**
Logistic Regression	0.06	0.01
Naive Bayes	0.03	0.001
SMO	0.09	0.01
IBk	0.001	0.001
J48	0.01	0.01
Random Forest	0.14	0.12

Conclusion:

In the proposed work, different classifiers were implemented on liver patient diseases data setto predict liver diseases based on developed software. Dataset was processed and implemented on WEKA tool using feature selection techniques with 10-fold cross validation testing option. The results of the proposed work were compared using feature selection and without using feature selection techniques after the implemental on of different classifiers in terms of execution time and accuracy. During the research work the result of other parameters like kappa statistic, correctly classified instances, and mean absolute error were also compared on liver patient diseases dataset. The best result was achieved using Logistic Regression classifier with feature selection techniques and execution time of different classifiers was decreased after the implementation of feature selection technique. Finally, Intelligent liver disease prediction Software (ILDPS) is developed using concept of software engineering life cycle.