International journal of basic and applied research

www.pragatipublication.com

ISSN 2249-3352 (P) 2278-0505 (E)

Cosmos Impact Factor-5.86

LIVER DISEASE PREDICTION USING SVM ANDNAÏVE BAYES ALGORITHMS

[1]Mr. K.V. Rajesh,[2]M. Sana,[3]M. Sushma,[4]M. Divya Sri

[1]Assistant Professor,[234]Students

Department Of CSE

Malla Reddy Engineering College for Women

ABSTRACT

When it comes to automated illness detection and prediction, data mining is a crucial component. The process uses data mining algorithms and methods to examine health records. One of the leading causes of death in a number of nations is liver disease, which has been on the rise in recent years. Predicting liver disease using classification models built from datasets of liver patients is the goal of this thesis. This thesis improved the prediction accuracy of Indian liver patients in three stages by constructing feature models and comparing them. The first step is applying the min max normalization technique to the original datasets of liver patients obtained from the UCI repository. During the second step of liver dataset prediction, a subset of the normalized liver patient dataset is produced using PSO feature selection. This subset contains just the most important features. In the third step, the data set is subjected to categorization methods. In the last stage, we'll figure out how accurate it is by calculating the root mean square and root mean error values. After using PSO feature selection, the J48 method is thought to be the superior algorithm in terms of performance. The assessment is concluded by looking at the accuracy values.

I. INTRODUCTION

Sitting on the right side of the abdomen is the huge and meaty organ known as the liver. Liver weighs around 3 pounds, has a rubbery texture, and is reddishbrown in color. The right lobe and the left lobe are the two main portions of the liver. The gallbladder, a portion of the pancreas, and the intestines all rest underneath the liver. All of these organs work in tandem with the liver to break down and absorb nutrients from meals.Filtering the blood that comes from the digestive system and sending it on to other parts of the body is the primary function of the liver. The liver is responsible for metabolizing medications and substances. In doing so, the liver reroutes bile to the intestines. Proteins necessary for several bodily processes, including blood clotting, are also produced by the liver.

Any illness-causing disruption of liver function is referred to as liver disease. When the liver is unhealthy or injured, it stops performing several vital processes that keep the body safe. When this happens, the body might suffer serious harm. Hepatic disease and liver disease are interchangeable terms. The inability of the liver to carry out its normal tasks is the result of a wide variety of conditions that together make up liver disease. For a reduction in liver function to occur, it is often necessary to impact more than 75%, or three quarters, of the liver's tissue.

II. II. ALGORITHMS

LOGISTIC REGRESSION

When the dependent variable is binary, logistic regression is the right regression strategy to use. Logistic regression, like other regression studies, is a predictive analysis. When one dependent binary variable has relationships with one or more independent variables at the nominal, ordinal, interval, or ratio levels, logistic regression is used for data description and explanation.

Logistic regressions might be difficult to understand, but the Intellectus Statistics tool makes it easy to run the analysis and provides an explanation of the results in simple English. Binary logistic regression main assumptions: .A presence/absence metric would work well for the dependent variable.Converting the continuous predictors to standardized scores and deleting values below -3.29 or more than 3.29 is one way to ensure that the data is free of outliers.All of the predictors should not be highly correlated with one another (multicollinearity). A correlation matrix among the predictors may be used to evaluate this. As long as the correlation coefficients among the independent variables are less than 0.90, according to Tabachnick and Fidell (2013), the assumption is satisfied. Logistic chances estimation is the meat and potatoes of logistic regression analysis. Logistic regression is a mathematical method for estimating a multivariate linear regression function, which is:

logit(p)

for i = 1…n .

Overfitting. Model fit is an additional critical factor to consider when choosing a model for logistic regression analysis. The amount of variation explained in the log odds (usually stated as R^2) is always increased when independent variables are added to a logistic regression model. On the other hand, overfitting happens when there are too many variables in the model, which limits its applicability outside of the original data set.

Reporting the R2. In binary logistic regression, several pseudo-R2 values have been proposed. These should be read with great care since they are inflated or deflated due to several computational errors. It is more appropriate to provide one of the existing goodness-of-fit tests; one such test is Hosmer-Lemeshow, which is based on the Chi-square test and is widely utilized.

SUPPORT VETOR REGRESSION

Support Vector Machine, or SVM, is a word that data scientists and machine learning professionals often hear. However, SVR differs little from SVM. Instead of using SVM for classification, which deals with discrete values, we may use SVR, a method for regression, to work with continuous values.

III. DATA CLASSIFICATION

Dataset

We used the Indian Liver Patient Dataset (ILPD) that was stored at UCI. There are 576 occurrences and 10 characteristics in this collection. The attributes are age, gender, tuberculosis, database, antigen, sgpt, sgot, tumor, tumor-to-graft ratio, and antigen-to-graft ratio. You may find information on liver function tests (LFTs) in this dataset.

Naïve Bayes

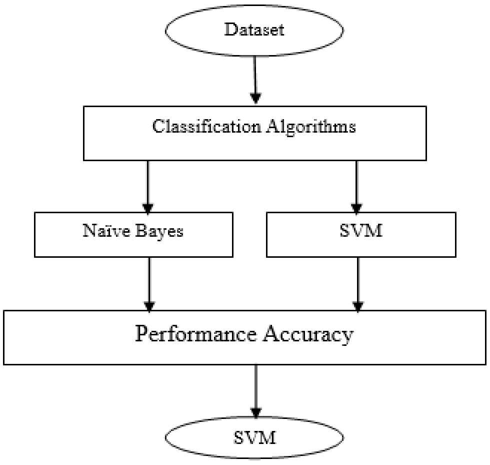
Using Bayes' theorem and strong independent assumptions, a Naive Bayes classifier is a straightforward probabilistic classifier. The underlying probability model is better described by the self-determining feature model. To put it simply, a Naive Bayes classifier takes it as read that there is no correlation between the existence of one class characteristic and any other feature [11]. In spite of a false underlying assumption, the Naive Bayes classifier nonetheless manages to do decent work.

$$\hat{P}(y = j|x_0) = \frac{\hat{\pi}_j \hat{f}_j(x_0)}{\sum_{k=1}^{k} \hat{\pi}_k \hat{f}_k(x_0)}$$

One benefit of using the Naive Bayes classifier is that it can estimate the means and variances of the classification variables with very little training data. Since there is no information about the independent variables, we simply need to find their variances for each label rather than the whole covariance matrix. When applied to numerical characteristics, the Naive Bayes (Kernel) operator differs from the Naive Bayes operator.

Using Bayes' theorem and kernel density estimation, this may be done in a straightforward manner:

IV. METHODOLOGY



Where

$\hat{\pi}_j$ is an estimate of the prior probability of class j; usually, $\hat{\pi}_j$ is the sample proportion falling into the j_{th} category

$\hat{f}_j$ is the predictable density at x0 based on a kernel density fit involving only observations from the j_{th} class

This is essentially the same idea as discriminant analysis, only instead of assuming normality, were estimating the probability density of the classes using a nonparametric method Patrick

Support Vector Machine

Support Vector Machine was first found in 1979 by Vapnik [5]. For classification and regression, Vapnik reiterated his recommendation in 1995 [4]. With its multilayer perceptron and radial-basis function networks, support vectors may be used for pattern categorization [8]. Statistical learning theory is at the heart of SVM, a modern technology that incorporates maximum classification techniques [9]. Both linear and non-linear data may be classified using SVM algorithms. By using non-linear mapping, it raises the dimensionality of the initial training data. Finding the linearly optimum separating hyperplane is its goal in this additional dimension. Hyperplanes with suitable nonlinear mappings to sufficiently high dimensions may split data into two groups. These hyperplanes are found by the SVM using margins and support vectors [6]. SVM carries out the classification job by reducing classification errors and maximizing margin of classification for both classes. While support vector machines (SVMs) have many potential optimization applications, including regression, data categorization remains the traditional challenge. In figure 2, we can see the main concept. Each data point is labeled as positive or negative; the objective is to locate a hyper-plane that maximally separates them.

1. Collecting needs: During this stage, we collect all of the client's needs, such as the inputs and outputs that the customer expects.

2. Analysis: During this phase, we create a document known as the "High Level Design Document" based on the client's needs. There are several sections in it, including Abstract, Functional Requirements, Non Functional Requirements, Current System, Future System, SRS, and more.

3. Design: Because the High Level Design Document is not readily understandable by all members, we utilize the "Low Level Design Document" to make it easier to grasp. The Unified Modeling Language (UML) was used in the creation of this document. Use case, sequence, collaboration, etc. are all part of this.Oh no...

Phase 4: Coding entails building the code in discrete modules. The modules are integrated when they have been developed.

5. Testing: Once development is complete, we need to verify whether the client's requirements have been met. If it doesn't work, we'll go back to developing.

6. Implementation: After the testing phase, we go on to implementation if the client's needs have been met. i.e., the program must be deployed to a server.

7. Maintenance: We provide maintenance for the application after deployment in case any issues arise on the client side.

The words and phrases that will appear often throughout this piece

1. The kernel is the main function that converts data from lower dimensions to higher ones.

SVM's second plane is the hyperplane, which serves as a demarcation line for the various data classes. However, in SVR, it will be defined as a line that aids in predicting the goal value or continuous value.

3. Boundary line: Support vector machines (SVMs) have two lines that generate a margin, apart from the Hyper Plane. Anywhere outside or on the boundary lines are acceptable for the support vectors. This dividing line distinguishes between the two categories. The idea is the same in SVR.

IV. Vectors of support: These information sites are located in close proximity to the border. The points are at their farthest points from one another.

One supervised machine learning approach that works for both regression and classification problems is "Support Vector Machine" (SVM). Nevertheless, categorization difficulties are its primary use. This technique takes a set of coordinates as input and uses them to create a point cloud representing each data item in an n-dimensional space (where n is the number of features). After that, we classify the data by locating the hyper-plane that effectively divides the two groups (see the picture below for an example). In real life, a kernel is used to implement the SVM algorithm. This introduction to support vector machines does not include how to train the hyperplane in linear SVM; doing so involves converting the issue using some linear algebra. Rephrasing the linear SVM using the inner product of any two provided data instead of the observations themselves is a major idea. Each set of input values is multiplied by itself twice to get the inner product of two vectors. Take [2, 3] and [5, 6] as an example; their inner product is 28, which is 2*5 + 3*6. The following is the formula for predicting a new input by joining all of the support vectors (xi) with the input (x):

$$f(x) = B0 + \mathrm{sum}(ai * (x,xi))$$

DECISION TREE

There are several real-world parallels to trees, and it turns out that trees have impacted many areas of machine learning, including regression and classification. One way to graphically and unambiguously depict choices and decision making in decision analysis is via a decision tree. The decision-making process is structured like a tree, as the name suggests. Despite its prevalence in machine learning, it is most often used in data mining to determine the best way to accomplish a certain objective.

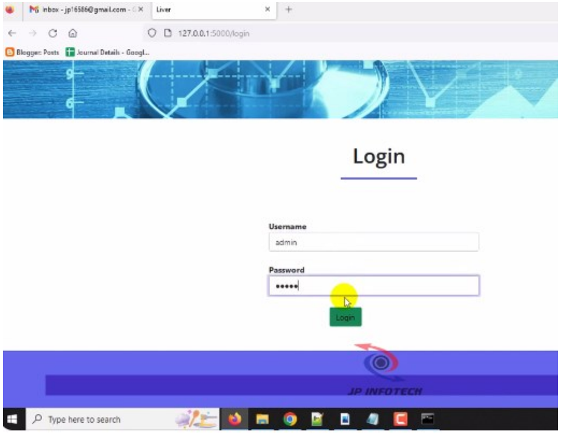
V. FUTURE SCOPE

Methods for selecting the Indian Liver Patient Database. In this thesis, the liver illness was examined by means of J48, MLP, SVM, Random Forest, and Bayesnet Classification, among other algorithms. The PSO feature selection model is the basis for the many outputs produced by these algorithms.When compared to other classification algorithms, bayes net and J48 Classification have shown to be more effective. This thesis employed characteristics such as Total bilirubin, Direct bilirubin, Total proteins, Albumin, A/G ratio, SGPT, SGOT, and Alkphos to assess the efficacy of various classification algorithms. There are several criteria for assessing feature subsets. Our next step is to try to categorize feature selection algorithms into four distinct types: full search, heuristic search, meta-heuristic approaches, and artificial neural network methods. A lot of people have been using PSO to choose features that would make liver categorization better. In addition, there is a lot of effort going into selecting features for liver categorization using multi-objective PSO, with the goal of both improving performance and decreasing the amount of features used. When choosing between uniform and non-uniform mutation, the majority of current PSO-based multi-objective feature selection methods employ binary tournament selection. Using improved selection and mutation processes, we may further decrease the search space for greater liver categorization accuracy. Separating the liver area into its component parts (liver, for example) is a goal of future research methods. In particular, the technique needs work in the area of feature selection for the liver's several anatomical regions, which include the renal cortex, renal column, renal medulla, and renal pelvis. Aside from that, we want to increase the size of the database used to test the system. In the future, using the heart dataset and illness categorization, the approach developed in this thesis may also be used to identify cardiac disorders.

VI. CONCLUSION

When it comes to medical diagnosis and illness prediction, classification is the data mining approach most often utilized in the healthcare industry. For the purpose of predicting liver illness, this study used two classification methods, namely Naïve Bayes and Support Vector Machine (SVM). Both the classification accuracy and the execution time are used as performance metrics to compare these algorithms. This study finds that the SVM classifier has the greatest classification accuracy out of all the algorithms tested. In contrast, the Naïve Bayes classifier requires the shortest possible execution time when comparing execution times.

VII. SCREENSHOTS

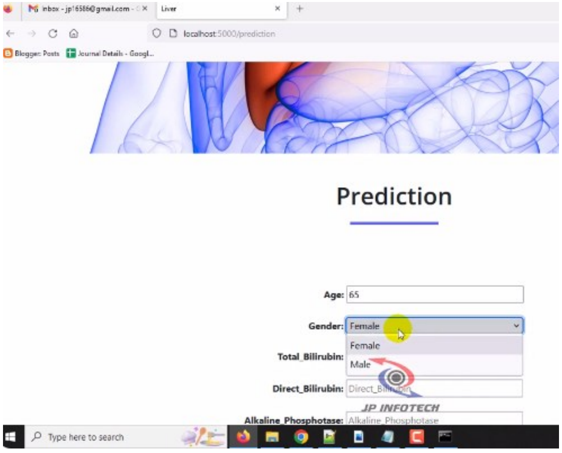


127.0.0.1:5000/login

Login

Username: admin

Password: •••••



localhost:5000/prediction

Prediction

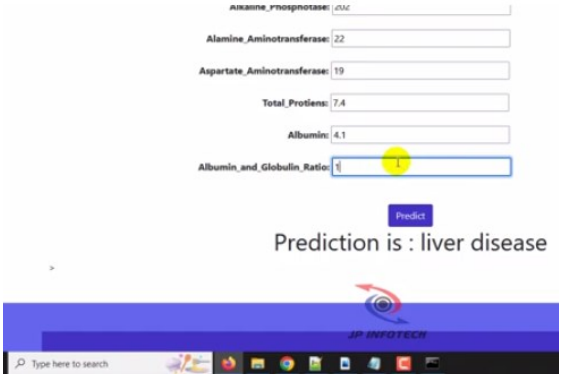
Age: 65

Gender: Female

Total_Bilirubin:

Direct_Bilirubin: Direct_Bilirubin

Alkaline_Phosphotase: Alkaline_Phosphotase



Alkaline_Phosphotase: 202

Alamine_Aminotransferase: 22

Aspartate_Aminotransferase: 19

Total_Protiens: 7.4

Albumin: 4.1

Albumin_and_Globulin_Ratio: 1

Prediction is : liver disease

REFRENCES

[1] D. Sindhuja and R. J. Priyadarsini, “A survey on classification techniques in data mining for analyzing liver disease disorder”, International Journal of Computer Science and Mobile Computing, Vol.5, no.5 (2016), pp. 483-488.

[2] B. V. Ramana, M. R. P. Babu and N.B. Venkaeswarlu, “A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis”, International Journal of Database Management Systems (IJDMS), Vol.3, no.2, (2011) , pp. 101-114.

[3] A.S.Aneeshkumar and C.J. Venkateswaran, “Estimating the Surveillance of Liver Disorder using Classification Algorithms”, International Journal of Computer Applications (0975 – 8887) , Vol. 57, no. 6, (2012), pp. 39-42.

[4] S.Dhamodharan, “Liver Disease Prediction Using Bayesian Classification”, 4th National Conference on Advanced Computing, Applications & Technologies, Special Issue, May 2014. International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.8, No.2, March 2018 9

[5] G. Selvara and S. Janakiraman, “A Study of Textural Analysis Methods for the Diagnosis of Liver Disease from Abdominal Computed Tomography”, International Journal of Computer Applications (0975-8887), Vol. 74, no.11 (2013), PP.7-13.

[6] H. Sug, “ Improving the Prediction Accuracy of Liver Disorder Disease with Oversampling”, Applied Mathematics in Electrical and Computer Engineering, American-MATH 12/CEA12 proceedings of the 6th Applications and proceedings on the 2012 American Conference on Appied Mathematics (2012), PP. 331-335.

[7] R.H.Lin, “An Intelligent model for liver disease diagnosis”, Artificial Intelligence in Medical, Vol. 47, no. 1 (2009), PP. 53-62.

[8] B. V. Ramanaland and M.S. P. Babu, “Liver Classification Using Modified Rotation Forest”, International Journal of Engineering Research and Development ISSN: 2278-067X, Vol. 1, no. 6 (2012), PP.17-24.

[9] H.R. Kiruba and G. T. arasu, “An Intelligent Agent based Framework for Liver Disorder Diagnosis Using Artificial Intelligence Techniques”, Journal of Theoretical and Applied Information Technology, Vol. 69 , no.1 (2014), pp. 91-100.