Liver Cancer Prediction using Classification Approach

Sukanya MN[1], Anjan Babu G[2]

Dept of Computer Science, SV University, Tirupati

Abstract**—** *In therapeutic, Liver Cancer is a champion among the most unavoidable and deadly harmful developments in people. Liver harm is difficult to be investigated at a starting period in light of the peril factors. In this paper, Support Vector Machine (SVM), is applied on Indian Liver Patient dataset. SVM, an incredible machine strategy created from factual learning and has made critical accomplishment in some field. In our investigation, the help vectors, which are basic for characterization, are acquired by gaining from the preparation tests. In this paper we have shown the near outcomes utilizing two SVM kernels, polynomial and RBF portions. The polynomial part has accomplished most elevated exactness*.

I. INTRODUCTION

Liver disease is an immense term that covers every one of the potential issues that reason the liver to disregard to play out its allotted limits. Regularly, more than 75% or 75% of liver tissue ought to be impacted before a decrease in limit happens [5]. Liver harmful development is the most perilous and subverting diseases in the whole world [7]. Liver harmful development is unbending to recognize at the outset time frame in view of the absence of appearances.

The liver's rule work is to strain the blood starting from the stomach related plot, prior to passing it to whatever is left of the body. The liver moreover detoxifies artificial materials and cycles drugs. As it does accordingly, the liver hides bile that breezes up back in the absorption parcels. The liver similarly makes proteins basic for blood thickening and various limits [5]. Liver ailment is any bother of liver limit that causes contamination. The liver is accountable for various unsafe limits inside the body and should it end up contaminated or hurt, the deficiency of those limits can make basic harm the body. Liver sickness is moreover insinuated as hepatic affliction.

One of the uses of Information Mining is helpful discovering which is generally used in research zone. Clinical Analysis is the place where various Scientists are concentrating. To reduce the investigation time and upgrade the discovering exactness, it has transformed into a basic concern. In Medical, Lever Cancer is a champion among the greater part unavoidable and savage harmful developments in individuals.

II. CLASSIFICATION

Technique is the way toward finding a model or a cutoff that depicts and sees data classes and contemplations, to use the model to anticipate the classes of things whose class mark isn't known. Data sales can be viewed as a two-stage measure: learning step in which a classifier is made depicting a predetermined outline of classes or thoughts by segregating the status set contained instructive rundown tuples and their related names [2]. In the ensuing advancement model is used for request by first evaluating the sensible exactness of classifier worked during the secret turn of events. It is done using the test data. The precision of classifier on a given test set tuples is level of tuples that are unquestionably referenced by the classifier. In case the accuracy is over some acceptable level, the classifier can be used to expect future tuples whose class mark isn't known.

Portrayal is a kind of data evaluation that can be used to make models portraying epic data classes. System is a data mining reasoning used to expect pack income for data models. It is one of the fundamental systems in data mining and is used in various applications, for instance, plan demand, torment certification, customer relationship the pioneers, and given out appearing. The goal of the portrayal appraisals is to gather a model from a huge load of planning data whose target class names are known and appropriately this model is used to pack covered cases [4].

Plan is the most conventional and most popular data mining techniques. Outline maps data into predefined get-togethers or classes. It is ordinary proposed as directed getting considering how the classes are settled going prior to taking a gander at the data [6]. Methodology is the way toward finding a model that sees data classes, to use the model to predict the class of things whose class name is dull. The picked model relies on the assessment of a huge load of planning data. Instructive varieties are rich with shrouded information that can be used for cautious dynamic.

III. METHODOLOGY

SVM, an incredible machine technique created from factual learning and has made huge accomplishment in some field. Presented in the mid 90's, they prompted a blast of revenue in AI. The establishments of SVM have been created by Vapnik and are acquiring ubiquity in field of AI because of numerous alluring highlights and promising observational execution.

3.1 Support Vector Machine

Support Vector Machines (SVM) is an AI computation that is all around used for request issues. SVM computation is potentially the most noteworthy portrayal strategies that were successfully applied to various genuine issues [1][9]. SVM rely upon arranging data centers to a high dimensional segment space where a segregating hyper-plane can be found. The rule reasoning used by SVM for data request is to drawn ideal hyper-plane which goes probably as a separator between the two classes. The vectors near the hyper-plane are called support vectors. This arranging can be carried on by applying the bit stunt which unquestionably changes the data space into another high dimensional component space. The hyper-plane is handled by intensifying the distance of the closest plans, i.e., edge support, avoiding the issue of overfitting [10].

Consider the two-class issue where the classes are straightly detachable. Let the dataset D be given as (x1, y1), (x2, y2) …. (xn, yn) ε Rn, where xi is the arrangement of preparing tuples with related class marks, yi. Every yi can take one of the two qualities, either +1 or - 1. The information is directly detachable in light of the fact that many numbers of straight lines can isolate the information focuses into two particular classes where, in class 1, y = +1 and in class 2, y = - 1. The best isolating hyperplanes will be the one which have the maximal edge between them. The most extreme edge hyperplane will be more precise in ordering the future information tuples than the more modest edge [3][10].

3.2 Kernel selection of SVM

Kernel selection plays an important role in SVM training and classification. A properly designed kernel function can minimize generalization error, accelerate convergence speed, and increase prediction accuracy. There are two common optimization methods, adding parameters and kernel alignment. Adding parameters is a method for putting additional parameters in the kernel and optimizing those parameters so as to improve the performance. There are four kernel methods are available:

- Linear kernel: $K(xi, xj) = xi^T xj$.
- Polynomial kernel: $K(xi, xj) = (\gamma\, xi^T xj + r)^d$, $\gamma > 0$
- RBF kernel : $K(xi, xj) = \exp(-\gamma \|xi - xj\|^2)$, $\gamma > 0$
- Sigmoid kernel: $K(xi, xj) = \tanh(\gamma\, xi^T xj + r)$ Here, γ, r and d are kernel parameters.

One of the best-known methods is the support vector machines (SVM), a kernel-based method which has found applications in many pattern recognition problems [2], [3]. Only polynomial and RBF kernels were analyzed in this work.

IV. EXPERIMENTAL RESULTS

We have utilized the Python Language to test our proposed calculations. The Python Scikit-learn is a bundle for information arrangement, relapse, grouping and representation. The proposed SVM kernel-based component determination strategies have been tested for Indian Liver Patient dataset has been taken from the UCI Machine Learning Repository [8].In this dataset, there are 576 instances and 11 traits and two class labels are Liver cancer present class contains 167 instances and Absent class has 416 instances. The information is partitioned in two sets. The preparation set is 70% (408 records) and the staying 30% (175 records) is utilized for testing. The detailed analysis of the dataset is shown in the figure-1 and figure-2 through density and histogram plots.

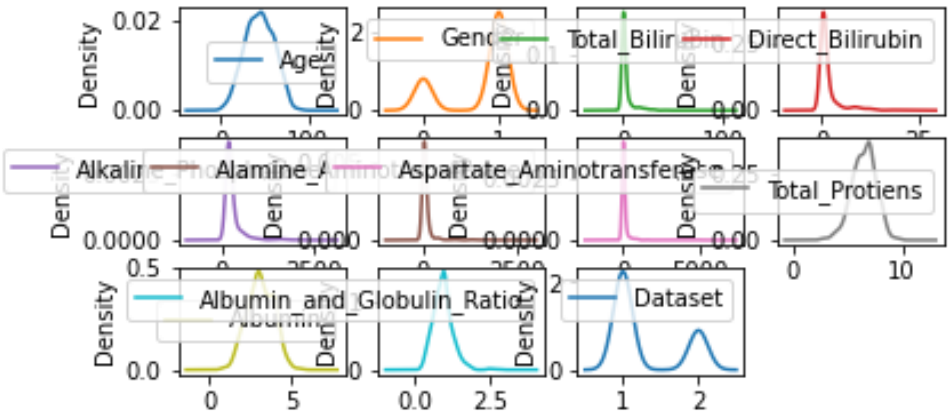


FIGURE 1: Density plot of dataset

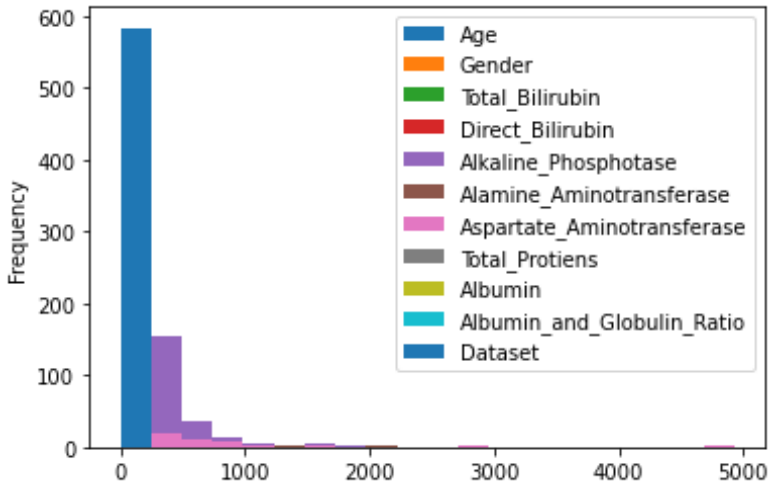
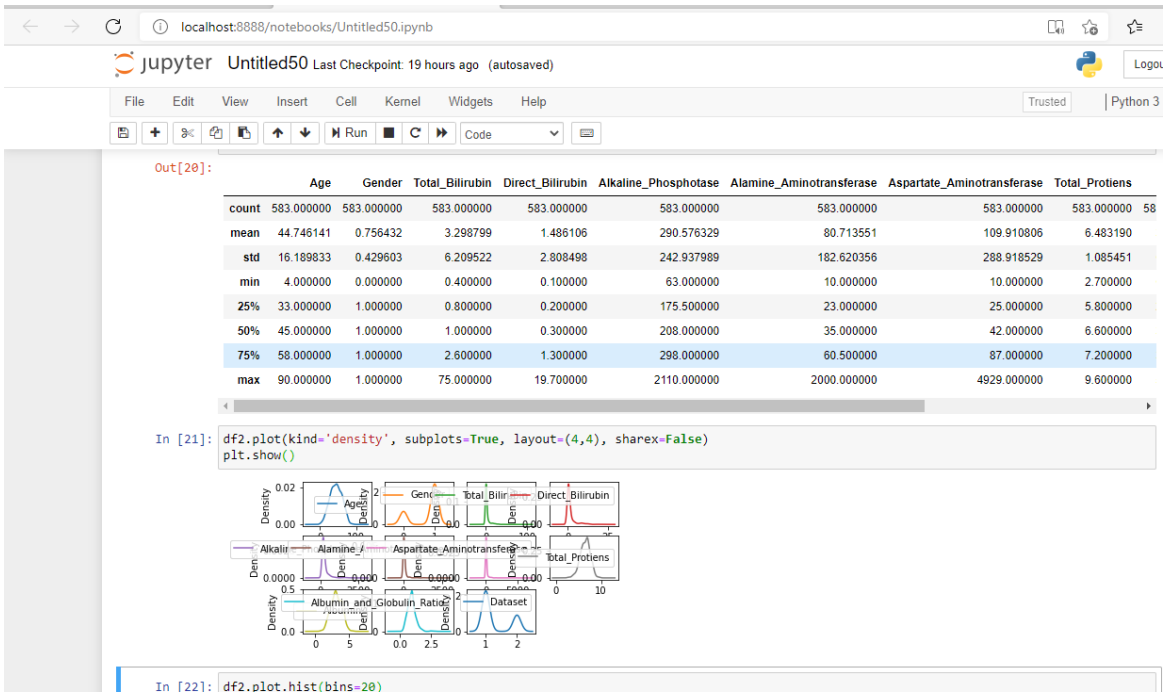


FIGURE 2: Histogram plot of entire dataset

The statistical summary information of the dataset is shown in the figure-3.



	Age	Gender	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Protiens
count	583.000000	583.000000	583.000000	583.000000	583.000000	583.000000	583.000000	583.000000
mean	44.746141	0.756432	3.298799	1.486106	290.576329	80.713551	109.910806	6.483190
std	16.189833	0.429603	6.209522	2.808498	242.937989	182.620356	288.918529	1.085451
min	4.000000	0.000000	0.400000	0.100000	63.000000	10.000000	10.000000	2.700000
25%	33.000000	1.000000	0.800000	0.200000	175.500000	23.000000	25.000000	5.800000
50%	45.000000	1.000000	1.000000	0.300000	208.000000	35.000000	42.000000	6.600000
75%	58.000000	1.000000	2.600000	1.300000	298.000000	60.500000	87.000000	7.200000
max	90.000000	1.000000	75.000000	19.700000	2110.000000	2000.000000	4929.000000	9.600000

FIGURE-3: Statistical description of the dataset

4.1 Results

In our experiment, the performance metrics of two kernels are compared to find an optimal and efficient kernel and it is carried out using Python software. A comprehensive performance study has been conducted to evaluate kernel selection using real-life Indian Liver Patient dataset obtained from the UCI Machine Learning Repository [6] to test its performance against two kernels with different parameters. We implemented SVM classification for two types of kernels: polynomial and Radial basis function (RBF).

It has been recommended that exactness acquired by the SVM relies generally upon the part chose and the boundaries. The investigation zeroed in on polynomial and outspread premise piece (RBF). The polynomial and RBF portions have boundary "C" that relates to the punishment for misclassification. The higher the worth of "C" is, the more the punishment, driving the arrangement model to be over-fitting. Alternately, more modest worth of "C" prompts a more summed up model that will be unable to order the obscure information precisely. In this paper, "C" was fluctuated from 1 to 50 for polynomial and 1 to 50 for the RBF portion and hence the ideal worth of "still up in the air alongside computing the expectation rate. In light of the normal expectation rate acquired by shifting the referenced boundaries they proposed ideal qualities for "C" as 1 for polynomial portion and 10 for RBF. SVM calculation was assessed utilizing two portions for our dataset. For SVM utilizing polynomial portion, the worth of p was changed from 1 to 3. Just 1, 2 and 3 were picked in this assessment. For SVM utilizing RBF part, the worth of γ was changed from 1 to 3. Just 1, 2 and 3 were picked in this assessment. In any case, in our dataset the ideal boundaries of p=2 and γ=2 shows the most elevated exactness are gotten our outcome as displayed in figure-4 and same displayed in the table-1.

TABLE 1
RESULTS OF SVM KERNELS

SVM Kernel	Accuracy	Precision	Recall
Polynomial	93.54	94	94
RBF	91.28	92	92

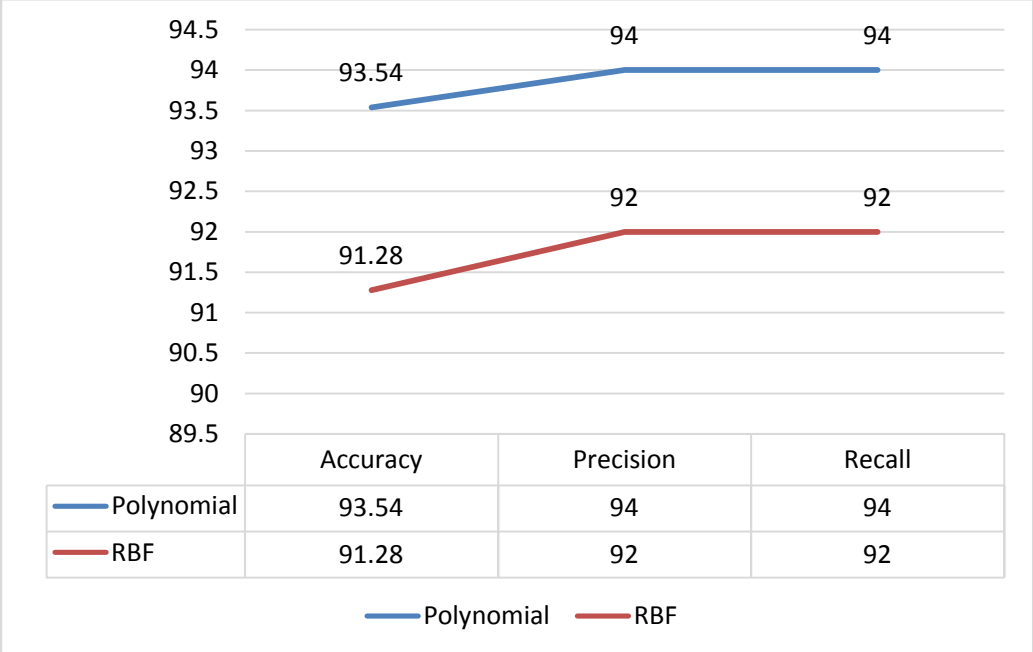


FIGURE 4: SVM Kernel results

In our experiment, the study focused on SVM algorithm was evaluated using two kernels polynomial and RBF kernels. We find in the Figure-4, the introduction of the SVM with polynomial kernel estimation has accomplished 93.54% accuracy, while SVM with RBF kernel has achieved 91.28%. It has been suggested that accuracy obtained by the SVM depends on the kernel selected and the parameters.

V. CONCLUSION

The SVM approach to machine learning is known to have both theoretical and practical advantages. Experimental results show that kernel selection greatly improves the quality of SVM classification. Our experimental results show that Polynomial kernel has achieved highest accuracy on Indian Liver Patient dataset. Experimental results show that kernel selection greatly improves the quality of classification. The selection of multiple kernel parameters is addressed to achieve accuracy, errors and time.

REFERENCES

[1] Bernhard Schölkopf and Alex Smola, Learning with kernels. MIT Press, Cambridge, MA, 2002.
[2] D. Hand, H. Mannila, P. Smyth.: Principles of Data Mining. The MIT Press. (2001)
[3] G. Bo and H. Xianwu, "SVM multi-class classification," Journal of Data Acquisition & Processing, vol. 21, pp. 334-339, 2006.
[4] Han J, Kamber M. Data Mining: Concepts and Techniques[J]. Data Mining Concepts Models Methods &Algorithms Second Edition, 2011, 5 (4).
[5] https://www.medicinenet.com/liverdisease/article.htm.
[6] Ian H. Witten and Eibe Frank. Data Mining: Practical machine learning tools and techniques.2nd ed. San Francisco: Morgan Kaufmann, 2005.
[7] Lam, Yee Hong Brian, Proteomic Classification of Liver Cancer using Artificial Neural Network", 2005.
[8] UCI machine learning repository. http://archive.ics.uci.edu/ml/.
[9] Vapnik, V.N. Statistical Learning Theory. John Wiley and Sons, New York, USA, 1998.
[10] Vapnik, V.N. The Natural of Statistical Learning theory. Springer–Verleg, NewYork, USA 1995.