Liver Cancer Detection Using SVM

M.V.B.T. Santhi,

Associate Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, India, Email: santhi_ist@kluniversity.in

Abstract

The healthcare industry has been using data mining to anticipate diseases in recent years. The process of retrieving or transforming specific information from sizable archives, warehouses, or other databases is known as data mining. For academics, predicting the illnesses from the vast medical information is a highly challenging task. To overcome this issue, the researchers employ data mining techniques such as clustering, grouping, and association rules, among others. The primary objective of this research is to predict liver disorders using classification algorithms. The techniques utilized in this paper are support vector machine (SVM), random forest, and naive bayes. These categorization methods are contrasted based on their accuracy and performance.

The performance and accuracy of Naïve Bayes and Support Vector Machine algorithms are examined and compared with the algorithm outcomes. In order to assess the accuracy and performance, we employed a dataset known as ILDP. Aspects including Name, Age, Total Bile, Sgpa, Sgpt, and so on are included in this dataset. When compared to other algorithms, the Support Vector Machine had the most accurate performance and accuracy.

Keywords: Liver, **S**upport Vector Machine (SVM), Random Forest, Naive Bayes.

1. Introduction

Predicting diseases using the massive amounts of medical records in the healthcare industry is a more challenging problem for researchers. These days, data mining is more prevalent in the healthcare industry. Medical data is subjected to data mining techniques such as classification, clustering, and association rule mining in order to identify recurring patterns that might be used to forecast disease. In data mining, classification algorithms are widely used for illness prediction and diagnosis. The classifier algorithms Naive Bayes and SVM will be employed in our study to predict liver illness. There are several liver illnesses that require the clinical care of a physician.

The primary goal of this work is to use algorithms to predict liver disorders such as cirrhosis, bile duct, hepatitis B, hepatitis C, and liver cancer from the ILDP dataset.

In human body liver is one of most powerful internal organ. To play an specific role in metabolism and to perform many essential roles, e.g. Red blood cell decomposition, etc. The weight of the organ is three kilos. The liver performs many important actions like digestion, metabolism, immunity, and nutrient conservation in the body. This actions will make the liver an important organ, with this tissues the body will be rapidly die due to lack of energy and nutrient supplements. There are different factors that lead to increase the risk in cause of hepatitis. Cancer is causes due improper growth of cells in the human body. The cells which are produced given result to cancer.

Since the Liver is a football shaped organ that present in the upper right portion of your uterus, behind your diaphragm and above your stomach. Several cancers may develop in the liver. The most common form of hepatocellular carcinoma is common. The main form of liver cancer is the hepatocellular carcinoma. Secondary hepatic cancer is the cancer that originates from the different organs and that spreads to the liver.

There are many methodologies for detecting hepatic cancer using deep learning. Of which one is image processing. First we take an image of the liver MRI and use many enhancement techniques to get a clearer picture by eliminating noise from the picture, and then segmentation of the liver cancer. We describe the liver cancer in this paper that we are taking an ILDP dataset. For data training and the data testing the dataset is split into two sections. Training requires for 75% of the dataset and the remaining 25% is used for testing, using SVM. The confusion matrix is used when the algorithm confuses.

Theseclassification algorithms are evaluatedbased on the output parameters, i.e. classification acc uracy or Precision noise from the image and then detecting liver cancer by segmenting. In this paper we are identifying the liver cancer we are taking a ILDP dataset .The dataset is divided into two parts for training and as well as testing the dataset. Training gets 75% of dataset and another 25% data is used for testing. The training is done using SVM and Naive Bayes. The confusion matrix is for accuracy check. These algorithms are compared on the analysis of the output factors i.e. precision or accuracy of classification.

2. Related Work

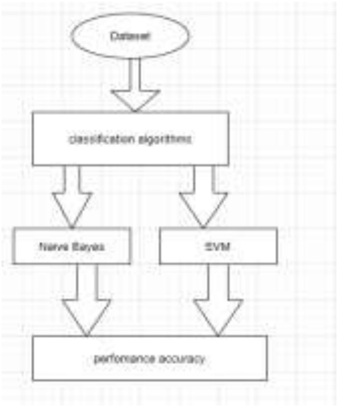
The liver is the first organ in the human abdomen, and it has a triangle form. The liver is made up of both the left and right hemi livers. Is a single heart. Our bodies once required the liver to function. It is the main organ responsible for controlling molecules like glucose and regulating different types of fat, carbs, vitamins, cholesterol, and hormones. The patient's survival rate needs to increase during the initial stages of treatment for liver disorders. The blood diagnosis of liver illnesses can benefit from an analysis of those enzyme levels. In the Information Discovery Process, the data mining approaches are separated into two categories. These are composed and predicted in an expressive manner. Each of these types will have a unique mining style.

Data mining techniques have the ability to extract information from even the most basic data. Knowledge mining is one step in the KDD process. It is the operation with the most checks. Information mining is a database search that looks for relevant examples or connections to help gather information. The analysis employs sophisticated observable procedures, such bunch research, and occasionally makes use of different ways involving the neurological system or a fabricated conscience. Finding the optimal relationships between the data beforehand is a noteworthy mining task, especially when the statistics come from various databases. Classification algorithms are used in chronic knowledge-based tasks.One branch of computer science is machine learning, which uses computer algorithms to find patterns in data.

The logistic regression framework is the model that has been suggested for prediction. Based on blood tests, this model will calculate the likelihood of getting liver disease. The following are the main contributions we have made. In order to forecast the chance of liver disease occurrence, a predictive model must first be created. Its effectiveness will be assessed, along with the significance of the disease prediction tests that are included.

3. **Methodology**

FLOW CHART



3.1 SUPPORT VECTOR MACHINE (SVM)

Support Vector Machine is a linear model for problems with the classification and regression. It can solve both linear and non-linear problems and function well for other practical issues. SVM's idea is simple: The algorithm generates a line or hyper plane that divides the data into groups.

SVM is an algorithm that takes the data as an input and outputs a line that divides the .SVM algorithm classes from both classes to find the points nearest to the line. Those points are called vectors of support. Now, we are measuring the distance from the line to the support vectors. The gap is known as the margin. Our goal is to optimize the margin. The hyper-plane the optimal margin for is optimal hyper plane.

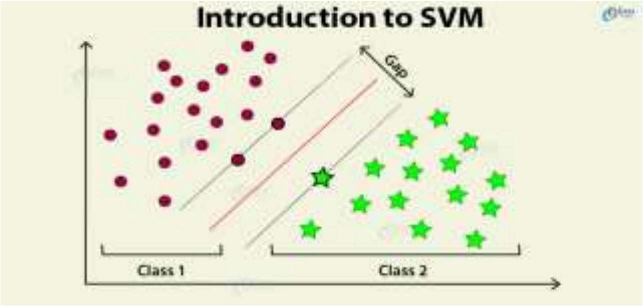


Figure 1 Classification of data into classes

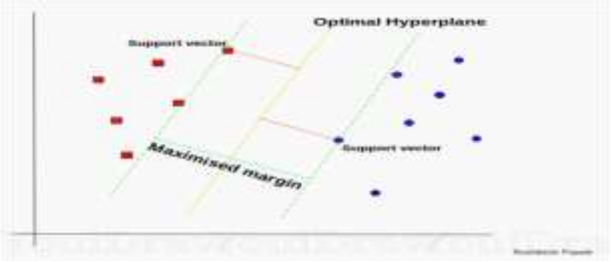


Figure 2 Choosing of Optimal Hyper plane

3.2 CONFUSION MARIX

The number of predictions that are right and incorrect is summarized with count values and broken down by gender. That's the secret to matrix uncertainty. The uncertainty matrix reveals how confused the model of classification is as it makes predictions.

The confusion matrix gives us insight not only into a classifier's errors but, more significantly, the types of errors that are made.

	Class 1 ***Predicted***	**Class 2** ***Predicted***
Class 1	**TP**	**FN**

Actual		
Class 2 **Actual**	**FP**	**TN**

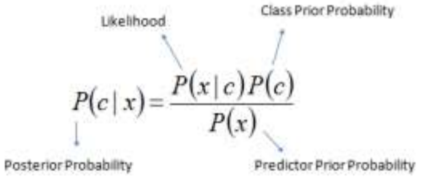
Table 1Confusion matrix

3.3 NAIVE BAYES ALGORITHM

It is a classification method predicated on the independence of predictors and the Bayes theorem. Put simply, a Naive Bayes classifier holds that the presence of one feature in a class has no bearing on the existence of any other feature. The Naive Bayes model can be easily constructed and is very helpful for very big data sets. In addition to being simple, Naive Bayes is thought to perform better than even the most complex classification algorithms.

It is easy and quick to predict the test data set's class. When independence is assumed, a Naive Bayes classifier outperforms other models, such as pragmatic regression, which requires less training data, in multi-class prediction as well. When dealing with categorical input variables, it works well in comparison to a numerical variable or variables. For number variables, the standard distribution (bell curve), which is a strong assumption, is made.

Bayes Theorem provides a way for P(c), P(x) and P(x) to measure posterior likelihood. Look at the equation underneath:



$$P(c\,|\,X) = P(x_1\,|\,c)\times P(x_2\,|\,c)\times\cdots\times P(x_n\,|\,c)\times P(c)$$

4. **Results and Discussions**

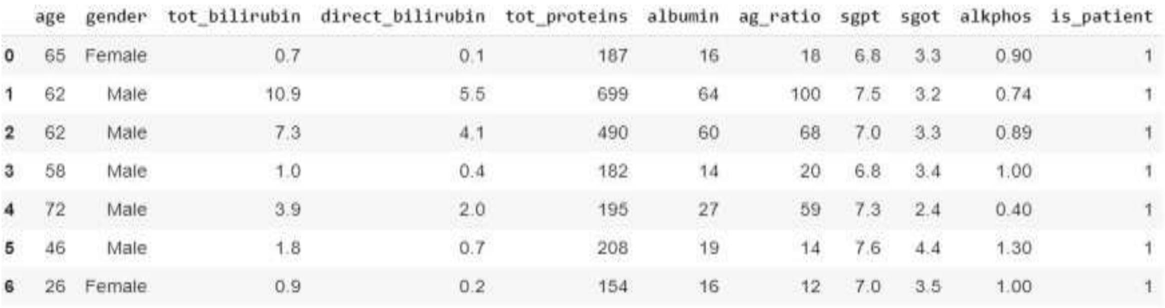
Loading the ILPD which is a dataset that we used to detect liver cancer. It consists of age, gender, TB, DB, TP and so on.

Data Analysis: This typically looks at the data to find out what is going on.

Check the data: Search for missing data, irrelevant data and do a cleanup.

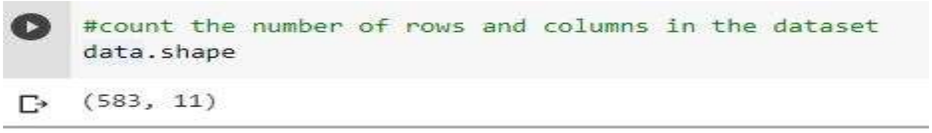
Data visualization: Data visualization is the description of data and information in graphic form. Visual components are used, such as charts, diagrams, and maps.

a) Reading the data



	age	gender	tot_bilirubin	direct_bilirubin	tot_proteins	albumin	ag_ratio	sgpt	sgot	alkphos	is_patient
0	65	Female	0.7	0.1	187	16	18	6.8	3.3	0.90	1
1	62	Male	10.9	5.5	699	64	100	7.5	3.2	0.74	1
2	62	Male	7.3	4.1	490	60	68	7.0	3.3	0.89	1
3	58	Male	1.0	0.4	182	14	20	6.8	3.4	1.00	1
4	72	Male	3.9	2.0	195	27	59	7.3	2.4	0.40	1
5	46	Male	1.8	0.7	208	19	14	7.6	4.4	1.30	1
6	26	Female	0.9	0.2	154	16	12	7.0	3.5	1.00	1

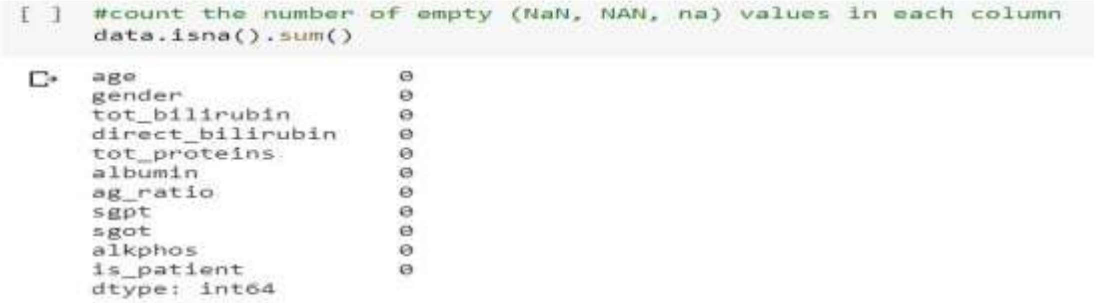
b) count the no. of rows and columns in the dataset



```
#count the number of rows and columns in the dataset
data.shape
```

```
(583, 11)
```

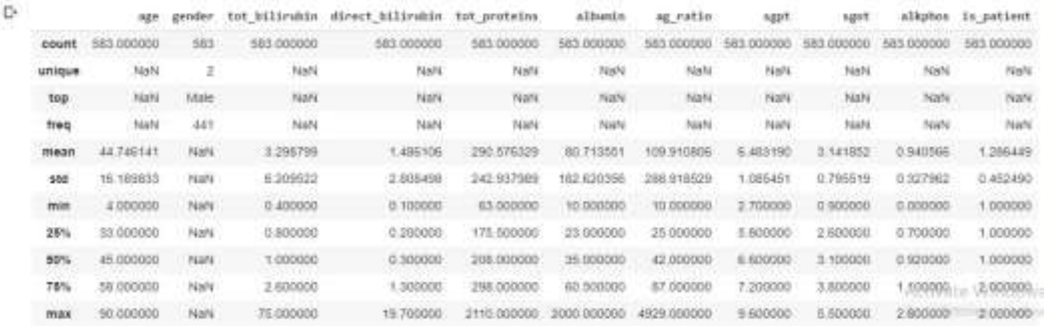
c) count the no. of empty (NaN, NAN, na) value in the each column



```
[ ]  #count the number of empty (NaN, NAN, na) values in each column
     data.isna().sum()
```

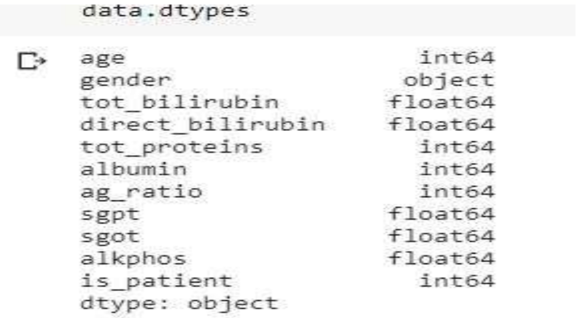
```
age                  0
gender               0
tot_bilirubin        0
direct_bilirubin     0
tot_proteins         0
albumin              0
ag_ratio             0
sgpt                 0
sgot                 0
alkphos              0
is_patient           0
dtype: int64
```

d) Statistical Parameters of the dataset



	age	gender	tot_bilirubin	direct_bilirubin	tot_proteins	albumin	ag_ratio	sgpt	sgot	alkphos	is_patient
count	583.000000	583	583.000000	583.000000	583.000000	583.000000	583.000000	583.000000	583.000000	583.000000	583.000000
unique	NaN	2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
top	NaN	Male	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
freq	NaN	441	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
mean	44.746141	NaN	3.298799	1.486106	290.576329	80.713551	109.910806	6.483190	3.141852	0.940566	1.286449
std	16.189833	NaN	6.209522	2.808498	242.937989	182.620356	288.918529	1.085451	0.795519	0.327962	0.452490
min	4.000000	NaN	0.400000	0.100000	63.000000	10.000000	10.000000	2.700000	0.900000	0.000000	1.000000
25%	33.000000	NaN	0.800000	0.200000	175.500000	23.000000	25.000000	5.800000	2.600000	0.700000	1.000000
50%	45.000000	NaN	1.000000	0.300000	208.000000	35.000000	42.000000	6.600000	3.100000	0.920000	1.000000
75%	58.000000	NaN	2.600000	1.300000	298.000000	60.500000	87.000000	7.200000	3.800000	1.100000	2.000000
max	90.000000	NaN	75.000000	19.700000	2110.000000	2000.000000	4929.000000	9.600000	5.500000	2.800000	2.000000

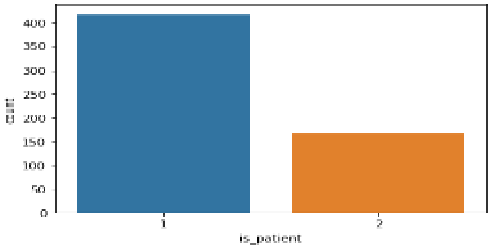
e) check at the data types to see which columns need to be encoded



```
data.dtypes
```

```
age                   int64
gender               object
tot_bilirubin       float64
direct_bilirubin    float64
tot_proteins          int64
albumin               int64
ag_ratio              int64
sgpt                float64
sgot                float64
alkphos             float64
is_patient            int64
dtype: object
```

f) Visualize the count



The above figure shows what type of cell it is either patient has cancer or non-cancer and also the gives the average count of each patient. If the patient has cancer it shows as 1 if the patient has no cancer is shows as 2

g) visualize the correlation

Correlation is a function of how strongly one variable relies on another variable. To train the model on different algorithms and check for best from them on basis of confusion matrix , accuracy etc. Support vector machine and naïve Bayes were the algorithms which we used for training and confusion matrix is used for testing.

Algorithm	Accuracy
SVM	0.78
NAIVE BAYES	0.39

Table 2. Accuracy Measure for Liver Disease Dataset

Algorithm	Training Score	Testing Score	Confusion matrix
SVM	0.74	0.80	[[0 28] [3 115]]
NAIVE BAYES	0.37	0.39	[[26 2] [86 32]]

Table 3. Training Score, TestingScore, Confusion Matrix

5. Conclusion

Early detection and treatment of liver cancer are critical for successful outcomes. Classification is one of the main data mining techniques used in the health care industry, mostly for illness prediction and medical diagnosis. In this study, SVM and naïve bayes classification algorithms were utilized to predict liver illness. On the basis of accuracy and precision, these algorithms are compared. This study's experimental results indicate that the SVM classifier is regarded as an effective algorithm.

6. References

[1] Rosalina. A.H, Noraziah.” A. Prediction of Hepatitis Prognosis Using Support Vector Machine and Wrapper Method”, IEEE, (2018), 2209- 22

[2] K.Karthik, SomasundranKrishnan.”Classification of liver patientdataset using machine learning algorithms”, IEEE,(2017).

[3] H. Blockeel and J. Struyf. “Efficient algorithms for decision tree cross-validation. Proceedings of the eighteenth International Conference on Machine Learning (C. Brodely and A. Danyluk, eds.)”, Morgan Kaufmann, 2011, pp. 11-18

[4] Haider, Nazar Ali Hussain.”A predictive model for liver disease progression based on logistic regression”,IEEE(2016).

[5] S. Rashid, A. Ahmed, I. Al Barazanchi, A. Mhana, and H. Rasheed, “Lung cancer classification using data mining and supervised learning algorithms on multi-dimensional data set,” Period. Eng. Nat. Sci., vol. 7, no. 2, pp. 438–447, 2019.