CONTENTS

Liver Disease Prediction Analysis

1.1.Introduction:

Problems with liver patients are not easily discovered in an early stage as it will be functioning normally even when it is partially damaged. An early diagnosis of liver problems will increase patient's survival rate. Liver failures are at high rate of risk among Indians. It is expected that by 2025 India may become the World Capital for Liver Diseases. The widespread occurrence of liver infection in India is contributed due to deskbound lifestyle, increased alcohol consumption and smoking.

1.2 Objectives of Research:

Using data mining techniques predicting patients liver disease is a time consuming task which degrades patients survival rate, By Appling different machine learning techniques An early diagnosis of liver problems will increase patients' survival rate based on accuracy and F score to find the best suitable algorithm for diagnosis of liver disease which gives best performance. It added a greater advantage to medical field. Some of the classification algorithms used are: 1.Decision trees 2. Support Vector Machine 3.logistic regression 4.K-NN

1.3 Problem Statement:

Problems with liver patients are not easily discovered in an early stage as it will be functioning normally even when it is partially damaged. proposed a system to diagnose medical diseases considering 6 benchmarks which are liver disorder, heart diseases, diabetes, breast cancer, hepatitis and lymph.

2. Review of literature:

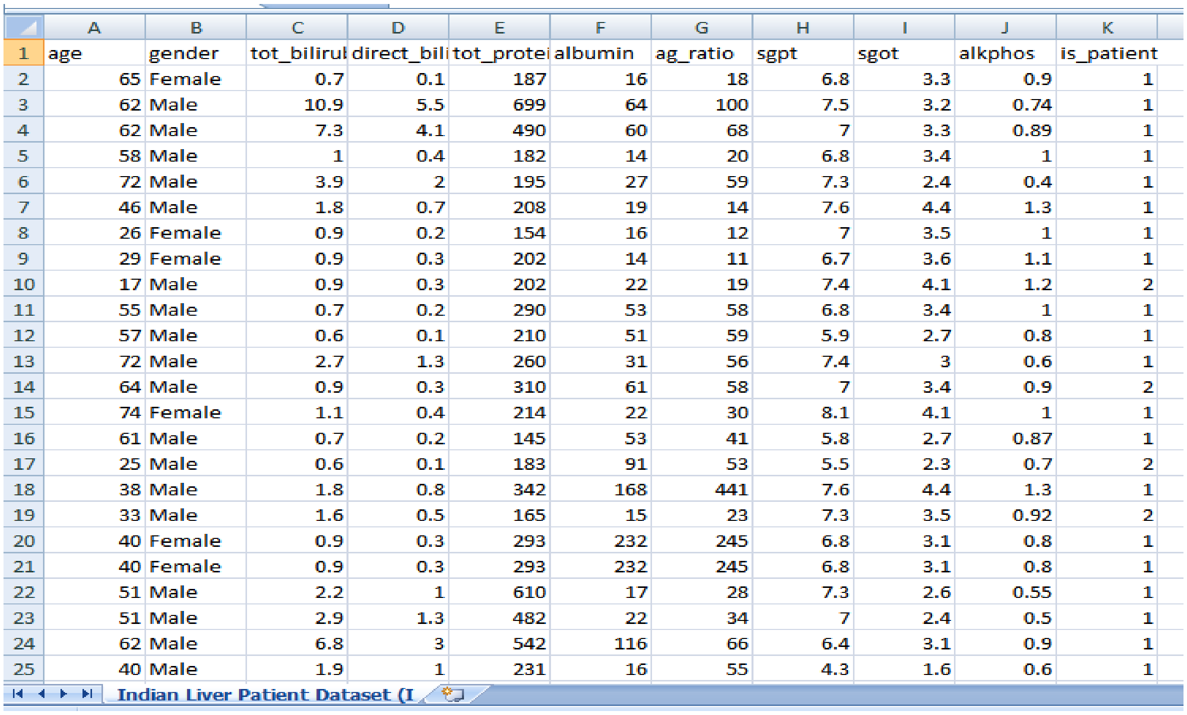
Survival Analysis is an extensively used procedure in the field of medical science. The idea of being able to predict the life expectancy of the subject is of immense value and utility to both, the doctors and the patients. There are three preliminary steps that serve as the elementary foundation of any medical treatment paradigm. The diagnosis stage, the classification stage, the assessment stage, the conclusion stage and finally the treatment stage. All these stages are expected to be accurate to the parameters and effective in their measure to distinctly reflect the quantified magnitude and the intensity of the study of the disease in the context. One of the most widely used classification methodologies that have been used for an extensive assessment of liver diseases, particularly cirrhosis is the Child-Pugh classification method

3.Data Collection:

We have collected the Data from kaggle. This data should contain 11 columns and contain 70000 records as rows.All of the rows are for predicting the Liver Disease Prediction. Based on Some factors, we are predicting this disease. The factors to predict liver disease

- Age
- Gender
- Total_bilirubin
- Direct_ bilirubin
- Total_protiens
- Albumin

- A/G ratio
- SGPT
- SGOT
- Alkphos



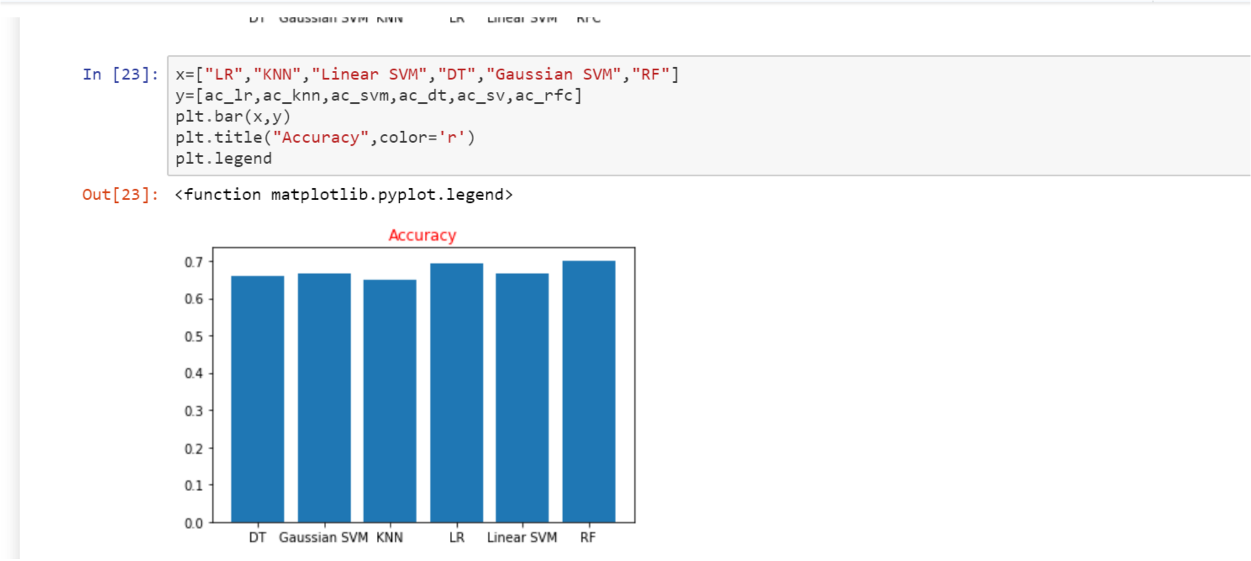
	A	B	C	D	E	F	G	H	I	J	K
1	age	gender	tot_bilirul	direct_bili	tot_protei	albumin	ag_ratio	sgpt	sgot	alkphos	is_patient
2	65	Female	0.7	0.1	187	16	18	6.8	3.3	0.9	1
3	62	Male	10.9	5.5	699	64	100	7.5	3.2	0.74	1
4	62	Male	7.3	4.1	490	60	68	7	3.3	0.89	1
5	58	Male	1	0.4	182	14	20	6.8	3.4	1	1
6	72	Male	3.9	2	195	27	59	7.3	2.4	0.4	1
7	46	Male	1.8	0.7	208	19	14	7.6	4.4	1.3	1
8	26	Female	0.9	0.2	154	16	12	7	3.5	1	1
9	29	Female	0.9	0.3	202	14	11	6.7	3.6	1.1	1
10	17	Male	0.9	0.3	202	22	19	7.4	4.1	1.2	2
11	55	Male	0.7	0.2	290	53	58	6.8	3.4	1	1
12	57	Male	0.6	0.1	210	51	59	5.9	2.7	0.8	1
13	72	Male	2.7	1.3	260	31	56	7.4	3	0.6	1
14	64	Male	0.9	0.3	310	61	58	7	3.4	0.9	2
15	74	Female	1.1	0.4	214	22	30	8.1	4.1	1	1
16	61	Male	0.7	0.2	145	53	41	5.8	2.7	0.87	1
17	25	Male	0.6	0.1	183	91	53	5.5	2.3	0.7	2
18	38	Male	1.8	0.8	342	168	441	7.6	4.4	1.3	1
19	33	Male	1.6	0.5	165	15	23	7.3	3.5	0.92	2
20	40	Female	0.9	0.3	293	232	245	6.8	3.1	0.8	1
21	40	Female	0.9	0.3	293	232	245	6.8	3.1	0.8	1
22	51	Male	2.2	1	610	17	28	7.3	2.6	0.55	1
23	51	Male	2.9	1.3	482	22	34	7	2.4	0.5	1
24	62	Male	6.8	3	542	116	66	6.4	3.1	0.9	1
25	40	Male	1.9	1	231	16	55	4.3	1.6	0.6	1

Figure

4. Methodology

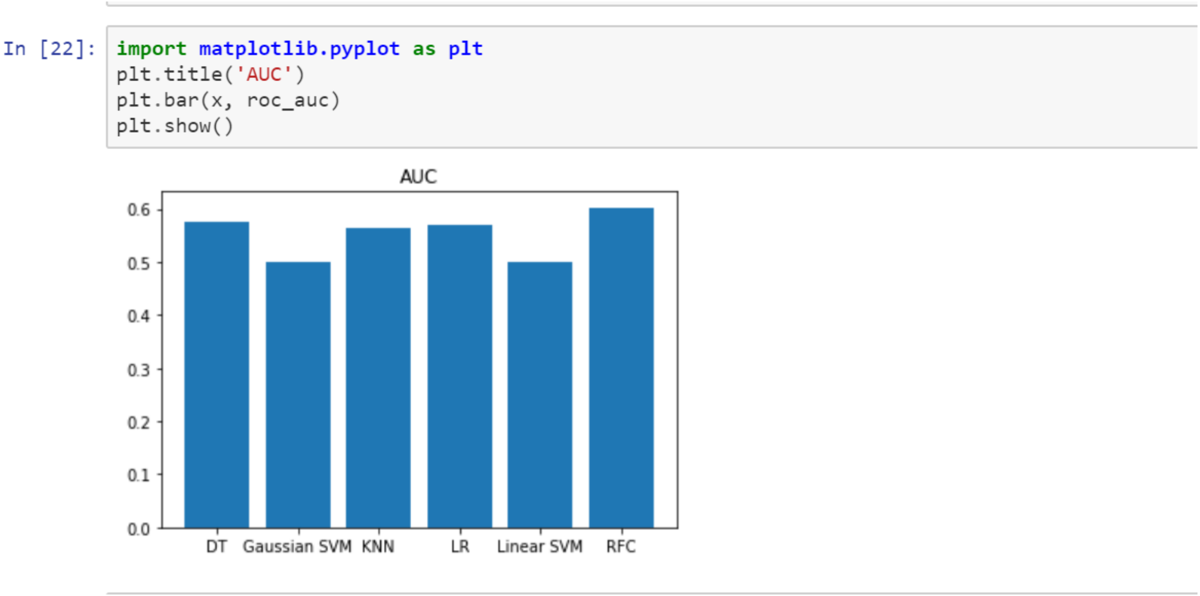
4.1 Exploratory Data Analysis

4.1.1 Figures and Tables :

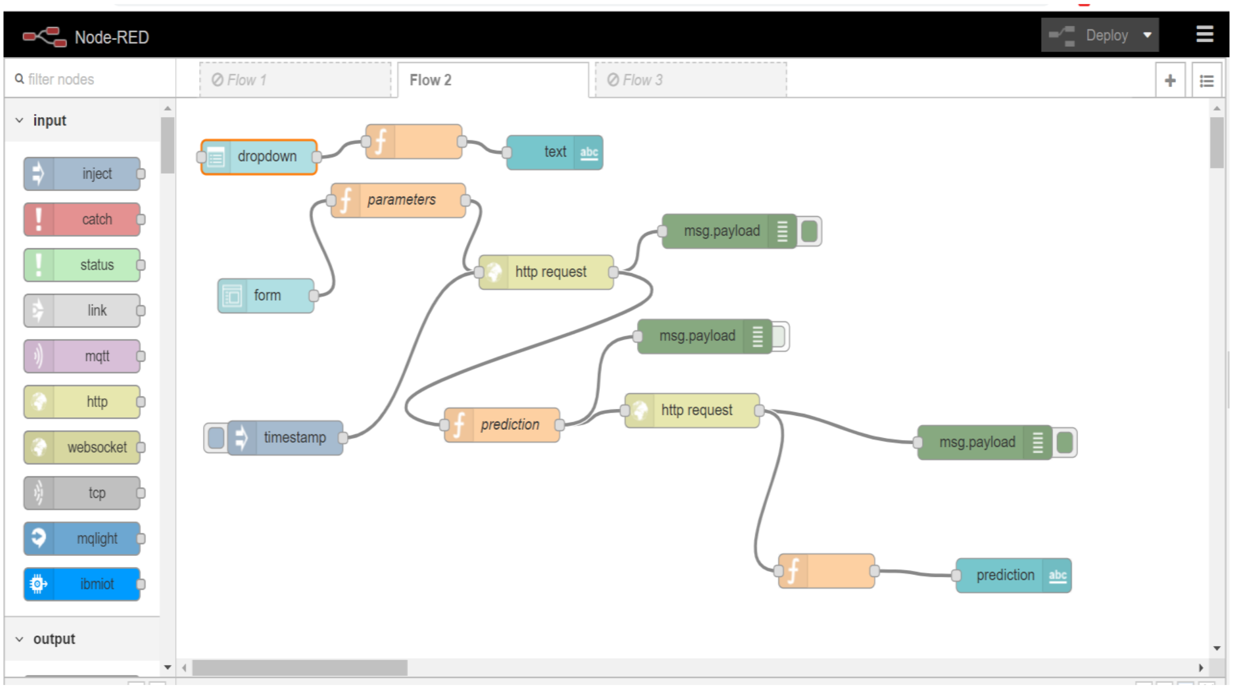


```
In [23]: x=["LR","KNN","Linear SVM","DT","Gaussian SVM","RF"]
         y=[ac_lr,ac_knn,ac_svm,ac_dt,ac_sv,ac_rfc]
         plt.bar(x,y)
         plt.title("Accuracy",color='r')
         plt.legend

Out[23]: <function matplotlib.pyplot.legend>
```



```
In [22]: import matplotlib.pyplot as plt
         plt.title('AUC')
         plt.bar(x, roc_auc)
         plt.show()
```



4.2 Data Modeling:

Data pre-processing is an important step of solving every machine learning problem. Most of the datasets used with Machine Learning problems need to be processed / cleaned / transformed so that a Machine Learning algorithm can be trained on it. Most commonly used pre-processing techniques are very few like missing value imputation, encoding categorical variables, scaling, etc. These techniques are easy to understand. But when we actually deal with the data, things often get clunky. Every dataset is different and poses unique challenges. All features, except Gender are real valued integers. The last column, Disease, is the label (with ’1’ representing presence of disease and ’2’ representing absence of disease). Total number of data points is 583, with 416 liver patient records and 167 non-liver patient records. In the description of this dataset, it is observed that some values are Null for the Albumin and Globulin Ratio column. The columns which contain null values are replaced with mean values of the column.

1. CLASSIFICATION TECHNIQUES:

a) **SVM (SUPPORT VECTOR MACHINE)**

SVM aims to find an optimal hyperplane that separates the data into different classes. The scikit-learn package in python is used for implementing SVM. The pre-processed data is split into test data and training set which is of 25% and 75% of the total dataset respectively. A support vector machine constructs a hyper plane or set of hyper planes in a high- or infinite-dimensional space. A good separation is achieved by the hyper plane that has the largest distance to the nearest training data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier.

b) **LOGISTIC REGRESSION**

Logistic regression is one of the simpler classification models. Because of its parametric nature it can to some extent be interpreted by looking at the parameters making it useful when experimenters want to look at relationships between variables. A parametric model can be described entirely by a vector of parameters = (0, 1... p). An example of a parametric model would be a straight-line y = kx + m where the parameters are k and m. With known parameters the entire model can be recreated. Logistic regression is a parametric model where the parameters are coefficients to the predictor variables written as 0 +1 +X1 + ...PXp Where 0 is called the intercept. For convenience we instead write the above sum of the parameterized predictor variables in vector form as X. The name logistic regression is a bit unfortunate since a regression model is usually used to find a continuous response variable, whereas in classification the response variable is discrete. The term can be motivated by the fact that we in logistic regression found the probability of the response variable belonging to a certain class, and this probability is continuous.

c) **K-NN**

This section describes the implementation details of KNN algorithm. The model for KNN is the entire training dataset. When a prediction is required for a unseen data instance, the KNN algorithm will search through the training dataset for the k-most similar instances. The prediction attribute of the most similar instances is summarized and returned as the prediction for the unseen instance. The similarity measure is dependent on the type of data. For real-valued data, the Euclidean distance can be used. Other types of data such as categorical or binary data, Hamming distance can be used. The KNN algorithm is belongs to the family of instance-based, competitive learning and lazy learning algorithms. Instance-based algorithms are those algorithms that model the problem using data in-stances (or rows) in order to make predictive decisions.

d**) Decision tree** : Algorithm falls under the category of supervised learning. They can be used to solve both regression and classification problems.

Decision tree uses the tree representation to solve the problem in which each leaf node corresponds to a class label and attributes are represented on the internal node of the tree.We can represent any boolean function on discrete attributes using the decision tree.

5. Findings and Suggestions:

There are many criterions for evaluating the selected feature subset, here this thesis used features such as Total bilirubin, Direct_ bilirubin, Total_protiens, Albumin, A/G ratio, SGPT, SGOT, Alkphos to evaluate the performance of different classification algorithm. In future, we have attempted toclassify different feature selection algorithms into four groups: complete search, heuristic search, meta-heuristic methods and methods that use artificial neural network.

The future methodology is used to analyze the liver region into separable compartments i.e. liver etc. However, the method requires further improvement mostly regarding feature selection of the liver into multiple components: renal cortex, renal column, renal medulla and renal pelvis. Apart from that, it is planned to expand the database on which the system will be tested. And also the proposed method in this thesis can be employed for detecting the heart diseases in future with the heart dataset and classification of the diseases

6.References:

[1] Bendi Venkata Ramana, Surendra. Prasad Babu. M, Venkateswarlu. N.B, A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis, International Journal of Database Management Systems (IJDMS), Vol.3, No.2, May 2011 page no 101-114

[2] Sebastian, Anu, and Surekha Mariam Varghese. "Fuzzy logic for child-pugh classification of patients with cirrhosis of the liver." 2016 International Conference on Information Science (ICIS). IEEE, 2016.

[3] Arshad, Insha, et al. "Liver disease detection due to excessive alcoholism using data mining techniques." 2018 International Conference on Advances in Computing and Communication Engineering (ICACCE). IEEE, 2018.

[4] Ramkumar, N., et al. "Prediction of liver cancer using Conditional probability Bayes theorem." 2017 International Conference on Computer Communication and Informatics (ICCCI). IEEE, 2017.

[5] Hassoon, Mafazalyaqeen, et al. "Rule optimization of boosted c5. 0 classification using a genetic algorithm for liver disease prediction." 2017 International Conference on Computer and Applications (ICCA). IEEE, 2017.

[6] Karthik. S, Priyadarshini. A. Anuradha J. and Tripathi B. K, Classification and Rule Extraction using Rough Set for Diagnosis of Liver Disease and its Types, Ad.

[7] Thapa, B. R., and Anuj Walia. "Liver function tests and their interpretation." The Indian Journal of Pediatrics 74.7 (2007): 663-671.

6.Conclusion:

In this project, we have proposed methods for diagnosing liver disease in patients using machine learning techniques. The four machine learning techniques that were used include SVM, Logistic Regression, KNN and Artificial Neural Network. The system was implemented using all the models and their performance was evaluated. Performance evaluation was based on certain performance metrics. ANN was the model that resulted in the highest accuracy with an accuracy of 98%. Comparing this work with the previous research works, it was discovered that ANN proved highly efficient. A GUI, which can be used as a medical tool by hospitals and medical staff was implemented using ANN. R