



Daffodil
International
University

Computational Intelligence Techniques for Disease Prediction

Supervised by

Dr. Md. Asraf Ali

Associate Professor

Department of Software Engineering

Daffodil International University

Submitted by

Md. Razu Ahmed

ID: 151-35-1072

Department of Software Engineering

Daffodil International University

This Thesis report has been submitted in fulfillment of the requirements for the Degree of Bachelor of Science in Software Engineering.

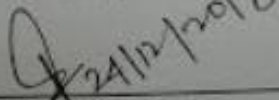
DECLARATION

I hereby declare that, I have taken this thesis under the supervision of **Dr. Md. Asraf Ali**, Associate Professor, Department of Software Engineering, Daffodil International University. I also declare that neither this thesis nor any part of this thesis has been submitted elsewhere for award of any degree or diploma.



Md. Razu Ahmed
ID: 151-35-1072
Program: B.Sc.
Department of Software Engineering
Daffodil International University

Certified By

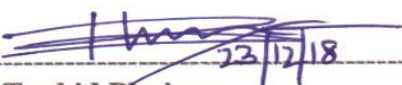


Dr. Md. Asraf Ali
Associate Professor
Department of Software Engineering
Daffodil International University

APPROVAL

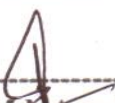
This thesis titled “**Computational Intelligence Techniques for Disease Prediction**”, submitted by **Md. Razu Ahmed, ID: 151-35-1072** to the Department of Software Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approved as to its style and contents.

BOARD OF EXAMINERS



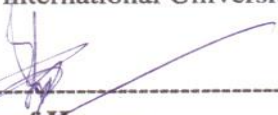
Dr. Touhid Bhuiyan
Professor and Head
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Chairman



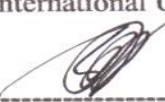
Dr. Md. Asraf Ali
Associate Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 1



Md. Maruf Hassan
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 2



Prof Dr. Mohammad Abul Kashem
Professor
Department of Computer Science and Engineering
Faculty of Electrical and Electronic Engineering
Dhaka University of Engineering & Technology, Gazipur

External Examiner

ACKNOWLEDGEMENT

Firstly, I express my heartiest thanks and gratefulness to almighty Allah for His divine blessing makes us possible to complete this study successfully.

I sincerely and heartily grateful to my advisor, **Dr. Md. Asraf Ali, Associate Professor**, Department of Software Engineering, Daffodil International University, Dhaka. For the support and guidance, he showed me throughout the study. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

I would like to express my heartiest gratitude to other faculty members of Software Engineering department of Daffodil International University.

Finally, I must acknowledgement with due respect the constant support and patients of my parents.

TABLE OF CONTENTS

DECLARATION	i
APPROVAL	ii
ACKNOWLEDGEMENT	iii
TABLE OF CONTENTS	iv
LIST OF FIGURES	vi
LIST OF TABLES	viii
ABSTRACT	ix
CHAPTER 1: INTRODUCTION	1
1.1 Background	1
1.2 Motivation of the Research	1
1.3 Problem Statement	2
1.4 Research Questions	2
1.5 Research Scope	2
1.6 Thesis Organization	3
CHAPTER 2: LITERATURE REVIEW	4
2.1 Article Searching Procedure	4
2.2 Article inclusion and Exclusion criteria	4
2.3 Data Extraction	4
2.4 Related Work	4
CHAPTER 3: MATERIALS AND METHODS	9
3.1 System Architecture	9
3.2 Data Collection	10
3.2.1 Indian Liver Patient Data	10
3.2.2 Breast Cancer Wisconsin Data	11
3.2.3 Chronic Kidney Disease	12
3.3 Data Preprocessing	13
3.3.1 Kidney Disease	13

3.3.2	Breast Cancer	16
3.3.3	Liver Disease	18
3.4	Classification Techniques	26
3.4.1	Logistics Regression	26
3.4.2	Support Vector Machine	26
3.4.3	Decision Tree	26
3.4.4	Random Forest	26
3.4.5	Naïve Bayes	27
3.4.6	K Nearest Neighbors	27
3.5	Evaluation Criteria	27
3.6	Software and Tools	28
	CHAPTER 4: RESULTS AND DISCUSSION	29
4.1	Analysis of the Results	29
4.1.1	Chronic Kidney Disease	29
4.1.2	Breast Cancer	31
4.1.3	Liver Disease	32
4.2	Performance Evaluation	34
	CHAPTER 5: CONCLUSIONS AND RECOMMENDATIONS	37
5.1	Findings and Contributions	37
5.2	Recommendations for Future Works	37
	REFERENCES	38
	APPENDIX A	42
	APPENDIX B	43
	APPENDIX C	44

LIST OF FIGURES

FIGURES	PAGE NO
Figure 3.1 The proposed system architecture	9
Figure 3.2: Numbers of missing values in CKD datasets	14
Figure 3.3: Heat map for checking corelated columns in CKD	15
Figure 3.4: Blood pressure and hemoglobin values in CKD	16
Figure 3.5: The heatmap shown that the accuracy is affected	17
Figure 3.6: No Missing values on breast cancer datasets	17
Figure 3.7: Heat map for checking corelated columns in breast cancer	18
Figure 3.8: Count plot shows the ratio of the liver patient	18
Figure 3.9: Count plot shows the ration of gender the liver patients	19
Figure 3.10: Factor plot of liver disease for both male and female	20
Figure 3.11 Direct relationship between Total and Direct Bilirubin	21
Figure 3.12: Direct relationship between Alamine and Asparate Aminotransferase	22
Figure 3.13: Direct relationship between Albumin and Total Proteins	22
Figure 3.14: Direct relationship between Total Proteins, Albumin and Globulin ratio	23
Figure 3.15: Disease by gender and age (years)	24
Figure 3.16: Number of missing values for liver datasets	24
Figure 3.17: Heat map for checking corelated column for liver datasets	25
Figure 4.1: Performance of six supervised classification techniques	29
Figure 4.2: Confusion matrix of classifiers	30
Figure 4.3: The accuracy of six machine learning	31
Figure 4.4: Classification performance measurements	32
Figure 4.5: Confusion matrix of classification techniques	32
Figure 4.6: The performance comparison of six classification techniques	33
Figure 4.7: The confusion matrix of prediction results	33
Figure 4.8: ROC curve for kidney datasets	34

Figure 4.9: ROC curve for breast cancer datasets 35

Figure 4.10: ROC curve for liver patient datasets 35

Figure 4.11: False discovery rate (FDR) 36

Figure 4.12: False Omission Rate (FOR) 36

LIST OF TABLES

TABLES	PAGE NO
Table 3.1: Parameters for liver data analysis	11
Table 3.2: Parameters for cancer data analysis	12
Table 3.2: Parameters for kidney data analysis	13
Table 4.1: Classification performance measurements	30

ABSTRACT

Objective: The aim of the study is to examine the performance of six Machine Learning algorithms for reducing the complexity and cost of chronic disease diagnosis by prediction.

Methods: I used six machine learning techniques for the classification of chronic disease datasets including Breast Cancer, Chronic Kidney Disease and Liver Patient datasets. SVM, NB, KNN, RF, DT and LR were used for prediction and diagnosis of chronic disease. The performance of the used techniques was evaluated with sensitivity, specificity, f 1 measure and total accuracy.

Results: All the machine learning classifiers show the accuracy level above 95% for both of the kidney disease and breast cancer prediction. Hence, the accuracy level nearly of 75% for liver disease prediction using all classification classifiers. In Kidney disease datasets, NB and RF has achieved the best performance than the other classification techniques in terms of accuracy by obtaining the highest accuracy as 100% respectively. The performance of analyzing breast cancer datasets, SVM achieved the highest performance with maximum classification accuracy of 97.07% while second highest classification accuracy is achieved by NB and RF (97%). Moreover, in the terms of accuracy for analyzing liver disease datasets, LR achieved the highest accuracy (i.e. 0.75%) and NB achieved the worst performance (i.e. 0.53%).

Conclusion: My findings showed that the NB, RF outperformed for analyzing the kidney datasets. NB, RF, SVM achieves best performance for performing experiment on breast cancer. In addition, LR have shown the utmost performance on liver disease datasets. In summary, our study has emphasized the research trends and scope in relation to chronic disease and clinical research fields by machine learning techniques, which has an effective impact in bio-medical fields.

Keywords: Machine Learning, Disease Prediction, Chronic kidney disease, Breast Cancer, Liver Disease, Supervised Learning.

CHAPTER ONE

INTRODUCTION

1.1 Background

Disease prediction is very important for medical institutions and physicians in order to make the best possible medical care decisions. Incorrect decisions are likely to cause delays in medical treatment or even loss of life. We know the medical services is a big commercial viewpoint in every time. The business stream always running in this fields. Patients are always searching there is a good platform for better services. But there is no 100% affordable platform for every patient. Therefore, in this domain need of one excessive platform for problem solves in healthcare and medical fields. Here is my main idea for improving healthcare services is to place more highlighting on early detection of chronic disease and less on treatment and live for better life.

1.2 Motivation of the Research

Chronic diseases are the leading causes of death and disability worldwide. It is estimated that worldwide over 508 000 women died in 2011 (“WHO | Breast cancer: prevention and control,” 2016). Four of the most prominent chronic diseases – cardiovascular diseases (CVD), cancer, kidney disease and type 2 diabetes. There are some common and preventable biological risk factors, notably high blood pressure, high blood cholesterol and overweight, and by related major behavioral risk factors: unhealthy diet, physical inactivity and tobacco use. Action to prevent these major chronic diseases should focus on controlling these and other key risk factors in a well-integrated manner (Luyckx, Tonelli, & Stanifer, 2018). About 17.7 million deaths around the world have been caused from CHD (indicating 31% of all global deaths) in 2015 (“WHO | World Health Organization,” n.d.). World Health Organization projected that more than 23.6 million individuals will be dead by 2030, because of heart disease (Purushottam, Saxena, & Sharma, 2016). Most of the time chronic disease diagnosis is very costly and complicated (Singh, Singh, & Pandi-Jain, 2018). This takes obsessive time, as a result, incorrect or delayed decisions are likely to cause death. However, the cost of kidney dialysis and diagnosis is very high and it can be calls as extreme level of financial expenses. Conferring to the report of Centers for Disease Control and Prevention (CDC), Kidney disease cause the commercial benefits with CKD cost over \$79 billion, and treating people with end stage renal disease cost around \$35 billion (“Chronic Kidney Disease Basics | Chronic Kidney Disease Initiative | CDC,” n.d.). kidney disease is chronic in nature and take long time for

cured. This causes most of the patients cannot afford the cost of the cure for kidney disease. Furthermore, chronic disease prediction is most prominent matter for clinical practitioners and medical services center in order to take accurate decision of such disease. Therefore, machine learning based extensive platform can solve these kidney disease problems through early detection and diagnosis. This works main aspect is to improve early treatment and diagnosis of kidney disease for peoples of low-income and developing countries. Hence, our study can be a significant approach for the detecting kidney disease outbreak with machine learning algorithms.

1.3 Problem Statement

In the last 10 years, medical data has been created in a large volume from various fields including health care services (HCS) (Ahmed, Arifa Khatun, Ali, & Sundaraj, 2018). Machine learning (ML) methods have depicted that aim to solve diverse medical and clinical problem (Hossain, Mahmud, Hossain, Haider Noori, & Jahan, 2018). Many of the studies show that machine learning methods have gained expressively high accuracies in classification-based medical problems. However, supervised learning-based methods are one of the most effective method for the research community and real-life applications on clinical fields. (Dwivedi, 2017). This works main aspect is to improve early treatment and diagnosis of chronic disease for peoples of low-income and developing countries. Hence, our study can be a significant approach for the detecting chronic disease outbreak with machine learning algorithms.

1.4 Research Question

The following research questions related to the prediction of chronic diseases are also addressed:

- (1) Does the different type of diseases affect the difference in performance between the various classifiers?

1.5 Research Scope

To date, machine learning classification techniques have created a significant impact and obligation in the chronic disease research society for early detection of chronic disease. Moreover, machine learning algorithms are given more accurate results in chronic disease prediction as compared to others data classification techniques (Dwivedi, 2017)(Mahmud & Ahmed, 2018). Many of studies already shows that the supervised based classification techniques have obtained excellent accuracies in the field of disease prediction (Heydari, Teimouri, Heshmati, & Alavinia, 2016)(Singh et al., 2018)(Kukar, Kononenko, Grošelj, ..., & 1999, n.d.) Motivated by this, the

authors have used six popular machine learning techniques for early detection and proper treatment of chronic patients. The main goal of this study is to examine the performance measurement of various prominent classification methods and gained more efficient outcome by reducing extremely cost of diagnosis and dialysis of chronic diseases. For this study, six supervised learning techniques were used including KNN, Support Vector Machine, Decision Tree, Random Forest, Naïve Bayes and Logistics Regression. Moreover, the performance of the selected learning techniques is evaluated using the confusion matrix and different statistical methods. Henceforth, the outperform classification technique will donated for the decision support system and diagnosis of chronic disease.

1.6 Thesis Organization

The remainder of the study is ordered as follows, chapter 1 describes the objectives of this thesis, motivation behind this thesis, research scope and thesis organization. Chapter 2 depicted the literature review and related works in these clinical areas. And the materials and methodology are described with the evaluation benchmark of different classification techniques in Section 3. Therefore, the performance results and discussion are illustrated in Section 4. Finally, conclusions and viewpoints for future research and recommendations are deliberated in Section 5

CHAPTER TWO

LITERATURE REVIEW

2.1 Article Searching Procedure

I used a systematic searching procedure to identify all of the available articles that discuss the chronic disease prediction and especially Liver Patient, Breast Cancer and Kidney disease prediction using machine learning techniques. In our systematic procedure, we search two keywords from ScienceDirect online database in order to access the article. We used the keywords, “machine learning” and “Liver Disease or Breast Cancer or Chronic Kidney disease” to find journal articles published in English Language between years 2016 to 2018.

2.2 Article inclusion and Exclusion Criteria

I used some benchmark to include and exclude articles from the set of articles that were selected through the search of ScienceDirect databases. I have selected ScienceDirect databases for best search results and quality research works. To include and exclude articles from the set of articles found through our systematic searching technique, we read the title, abstract, methodology and results of each article. We considered only those article that were written in English and that used machine learning. The exclusion criteria were the following: 1) Article that applied Machine learning for disease prediction, 2) And Liver Patient, Breast Cancer and Chronic Kidney prediction using Machine Learning.

2.3 Data Extraction

We carefully read and analyzed all of include articles to collect the key information. We followed the standard data extraction from for the particular analysis of each article. Each article was evaluated for the following information: 1) Performance comparison for different Machine Learning algorithms, 2) Disease prediction using Machine Learning technique, 3) Working with Liver Patient, Breast Cancer and Chronic Kidney datasets.

2.4 Related Work

Through related work, 21 studies were done on applying and using different machine learning techniques to determine early detection and diagnosis of different chronic diseases. Previous work also introduces a set of studies-based prediction and detection of chronic diseases using machine

learning algorithms. However, the outcomes of the 21 articles on machine learning used in disease prediction as follows:

(Abdelaziz, Elhoseny, Salama, & Riad, 2018), introduced a model for HCS based on cloud computing environment using particle Swarm optimization to optimize the VMs selection. The projection model of chronic kidney disease (CKD) is implemented using 2 machine learning techniques, (i) LR regulate critical factors that influence on CKD, (ii) Neural Network (NN) is used to forecast of CKD. The contribution of this study, a proposed system efficiency considering real time data retrieval is significantly enhanced by 5.2%. And accuracy of the hybrid model is 97.8%.

(Chen, Zhang, Zhu, Xiang, & Harrington, 2016), presented a new way to evaluate the two fuzzy approaches for diagnosis of patients with kidney disease, FURES and FOAM, are viable for the classification of biomedical arena. Moreover, the average prediction rates of FURES and FOAM obtained from 200 bootstrapped evaluations were $99.2 \pm 0.3\%$ and $99.0 \pm 0.3\%$. In addition, PLS-DA yields slightly worse accuracy with $95.9 \pm 0.6\%$. The contribution of this study is to evaluate the two classifiers.

(Kazemi & Mirroshandel, 2018), introduced a new model to obtain the early detection of the type of kidney stone and provide a decision support system. This paper uses different classification techniques to examine the quality of the proposed model. The final ensemble-based model showed that the accuracy is 97.1%. The contribution of this study is to use an early identification and reduction in diagnosis time.

(Jain & Singh, 2018), presented a survey to feature selection and machine learning techniques for diagnosis and prediction of chronic disease. This works focused on a comprehensive review of different feature selection methods and their advantage and limitation. The contribution of this study is to use adaptive and parallel classification systems for chronic disease prediction.

(Bartz-Kurycki et al., 2018), introduced a new model to predict neonatal surgical site infections (SSI) using different classification methods. Accuracy of area under the curve for each model was similar (full model 0.65, clinical model 0.67, RF 0.68, hybrid LR 0.67). The contribution of this study is to examine the hybrid model and other models with fewer and more clinically relevant variables.

(Carvalho, Pinheiro, & Pinheiro, 2016), presented a new hybrid model to support the early diagnosis of breast cancer. This study seeks to find optimal accuracy to give support to diagnosis in cases where the Bayesian Network does not provide a satisfactory result. The contribution of the study is to develop an automated tool to give an accurate diagnosis and prediction of breast cancer.

(Kumari, 2018), presented a new prediction system that can predict the occurrence of breast cancer at early stage by analyzing nominal set of attributes that has been selected from medical datasets. The KNN classifier obtain the best performance (99.28%) than others classifiers. The contribution of this study is to use the proposed system to predict the breast cancer at early stage with greatly reduces the cost of treatment and improves the quality of life.

(Tapak et al., 2018), introduced a comparative study between Naïve Bayes, Random Forest, AdaBoost, Support Vector Machine, LSSVM, Adabag, Logistics Regression and LDA to predict breast cancer survival and metastasis. LR and LDA were achieved the highest accuracy (86%). And the SVM and LDA have superior sensitivity in comparison to other classifiers. The contribution of this study is to use SVM to predict survival of breast cancer.

(Asri, Mousannif, Moatassime, & Noel, 2016), presented a comparative study between SVM, DT(C4.5), NB, K-NN to predict early stage of breast cancer. The intelligent techniques are applied on WEKA data mining tool. Experiment results show that the SVM have the best performance accuracy, it is 97.13%. The contribution of this study to use SVM to predict the early stage of breast cancer.

(Chougrad, Zouaki, & Alheyane, 2018), developed a computer aided diagnosis (CAD) system based on deep convolutional neural networks (CNN). The CNN model achieved the best performance, it is 98.94%. And they tested the CNN model on independent database and they've got the accuracy 98.23% and 0.99 AUC. The contribution of this study to use the high performer classifiers within the proposed structure and that can be used to predict the mass lesions are benign or malignant.

(Jafari-Marandi, and et al 2018), introduced a new approach to solve the problem of extracting patient-reported symptoms from free-text electronics health records. Moreover, the final model achieved the precision of 0.82, 0.86, 0.99 and recall of 0.56, 0.69, and 1.00. The contribution of

this study is to use the machine learning techniques for tracking and analyzing symptoms experienced by the cancer patient.

(Nagarajan & Upreti, 2017), presented a new framework to find the best performance of multiple base classifiers. The contribution of this study is to investigate an ensemble predictive modeling structure for predicting good-prognosis and poor-prognosis breast cancer samples from their 70-gene signatures using openly available data.

(Shukla, Hagenbuchner, Win, & Yang, 2018), introduced a robust data analytical model which can perform the better understanding about breast cancer survivability in presence of missing data and provide a better insight about survival data. The contribution of this paper is used to segment the historical patient data into clusters or subsets. The survivability forecast accuracy of an MLP is enhanced by using recognized patient groups as different to using raw historical data.

(Wang, Zheng, Yoon, & Ko, 2018), presented a new model to use breast cancer diagnosis based on patient's historical data from clinical data. The proposed WAUCE model reduces the variance by around 97.98% and increase accuracy by 33.34%. The contribution of this study can be further applied to safer, more reliable illness diagnosis process.

(Xiao, Wu, Lin, & Zhao, 2018), introduce a new strategy for gene expression analysis to five different classification algorithms with deep learning methods. The contribution of this study is shown to be accurate and effective prediction results for cancer prediction.

(Baitharu & Pani, 2016), presented a comparative study between DT(J48), NB, ANN, ZeroR, 1BK and VFI algorithm to predict the early stage of liver disease. By analyzing the results, Multilayer perception shows the outer perform than other classifiers. The contribution of this study to comparison different classification algorithms for liver disease predictions.

(Nilashi, Ahmadi, Shahmoradi, Ibrahim, & Akbari, 2018), presented an accurate model for the hepatitis disease diagnosis by taking the benefits of ensemble learning. The results of this analyses on the dataset displayed that the method performance is excellent to the Neural Network, ANFIS, K-Nearest Neighbors and Support Vector Machine. The contribution of this study is used to intelligent learning system for hepatitis disease diagnosis in the medical fields.

(Abdar, Zomorodi-Moghadam, Das, & Ting, 2017), introduced a computer-aided diagnostic approaches as an intelligent system have notable impression on liver disease detection. The contribution of this article is observing the classifiers performance for predict early stage of liver disease.

(Acharya et al., 2016), presented a study on the detection of steatosis and classification of steatotic livers from the normal using ultrasound images. Moreover, the important information from the image is extracted using GIST descriptor models. An average classification accuracy of PNN classifier can diagnose FLD with 98%, 96% sensitivity, 100% specificity and Area

Under Curve (AUC) of 0.9674. The contribution of this study will reduce the burden of the radiologists by 50% for early detection of fatty liver disease.

(Lorente et al., 2017), The contribution of this study is comparative study on different machine learning based classifiers. Random Forest shows the best performance with a mean error of 0.07 mm. Finally, the best results to place the future development of real-time software capable of early prediction of liver disease.

(Rau et al., 2016), presented a new model for predicting the development of liver cancer within 6 years of diagnosis with type II diabetes. In this study, used two classifications techniques, such as the artificial neural network (ANN) and logistic regression (LR). The contribution of this study is use to detect liver cancer patients and also helpful to cancer treatment.

CHAPTER THREE

MATERIALS AND METHODS

3.1 System Architecture

In this present study, the proposed system architecture including machine learning techniques has been presented. The proposed disease detection and Monitoring System consists of a four-tier system architecture to store and process a huge volume of wireless sensors and device data. Tier 1 focuses on collecting and combined data from different health tracking sensors and devices. In tier 2 uses to store huge amount of medical data. Therefore, tier 3 uses machine learning classification algorithm for training data of chronic disease datasets. In addition, tier 4 represents the results of the whole system for users. The proposed system architecture for early detection of chronic disease is shown in Figure 3.1. In this study, I focused only the machine learning phase (tier 3). For future study, I am currently developing in this full system architecture.

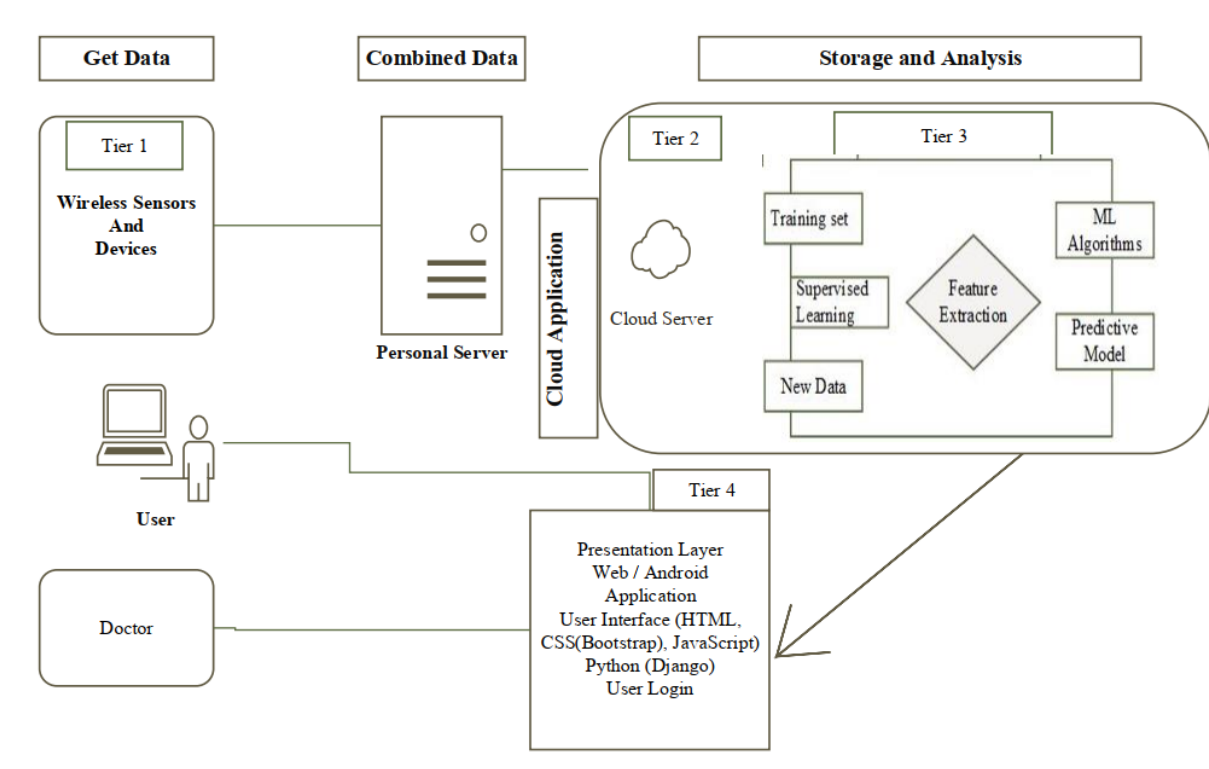


Figure 3.1 The proposed system architecture

The proposed system architecture consists of four modules, specifically, the data collection module, the data storage module, analysis module, and application presentation module. In data collection module is used for extracting the particular patient or person data using health tracking sensors (i.e. heart rate sensor, fitness wearables, infant monitoring etc.) and devices (i.e. smart watch, smartphones, medical IoT based tools etc.). Health tracking devices are integrated into the human body to collect the particular person's health data in a continuous manner. In addition, tracking sensors and devices are sending health data uninterruptedly. For this vast amount of data, it is very tough to store and analysis by traditional database tools and techniques. The proposed architecture uses cloud storage and NoSQL database technologies to store the continuous healthcare data using machine learning classification techniques. Moreover, in the application module users can see their health report through the mobile application.

3.2 Data Collection

3.2.1 Indian Liver Patient Data

In this study, we use the liver patient data from provided by the University of California, Irvine (also known as UCI Machine Learning Repository). In addition, this dataset is originally from the north east of Andhra Pradesh, India (Ramana, Babu, of, & 2012, n.d.). The objective of the dataset is to diagnostically predict whether or not a patient has liver disease based on certain diagnostic parameters. Moreover, in the dataset contains 583 liver patient's data whereas 416 samples are liver patients and 167 samples are non-liver patient. This data set contains 75.64% samples are male patients and 24.36% samples are female patient. Furthermore, if any patient whose age beaten 89 is listed as being of age "90". However, I chose the particular parameters for data analysis which are summarized in table 3.1

Table 3.1 Parameters for data analysis

No	Attributes	Indication	Description
1	Age	Numerical	Years
2	Gender	Male or Female	Sex of the patients
3	Total_Bilirubin	Numerical	mg/dL
4	Direct_Bilirubin	Numerical	mg/dL
5	Alkaline_Phosphotase	Numerical	ALP in IU/L
6	Alamine_Aminotransferase	Numerical	ALT in IU/L
7	Aspartate_Aminotransferase	Numerical	AST in IU/L
8	Total_Protiens	Numerical	g/dL
9	Albumin	Numerical	g/dL
10	Albumin_and_Globulin_Ratio	Numerical	A/G ratio
11	Dataset	Yes or No	patient has liver disease or not

3.2.2 Breast Cancer Wisconsin Data Set

In this study, we use the Wisconsin Breast Cancer data (Original) from provided by the University of California, Irvine (also known as UCI Machine Learning Repository). In addition, this dataset is originally from the University of Wisconsin Hospitals Madison, Wisconsin, USA (Wolberg & Mangasarian, 1990). In this dataset contains 699 breast cancer patients' records. Moreover, the datasets contain the Benign: 458 (65.5%) samples and Malignant: 241 (34.5%) samples. However, I chose the particular parameters for data analysis which are summarized in table 3.2

Table 3.2 Parameters for data analysis

No	Factor	Information Factor	Description
1	Id	Numerical	Id
2	Clump Thickness	Numerical	(1-10)
3	Uniformity of Cell Size	Numerical	(1-10)
4	Uniformity of Cell Shape	Numerical	(1-10)
5	Marginal Adhesion	Numerical	(1-10)
6	Single Epithelial Cell Size	Numerical	(1-10)
7	Bare Nuclei	Numerical	(1-10)
8	Bland Chromatin	Numerical	(1-10)
9	Normal Nucleoli	Numerical	(1-10)
10	Mitoses	Numerical	(1-10)
11	Class	Benign or Malignant	2 for benign, 4 for malignant

3.2.3 Chronic Kidney Disease

In this literature, we use the kidney patient data from provided by the University of California, Irvine. In addition, this dataset is originally from the Apollo Hospitals, Managiri, Madurai Main Road, Karaikudi, Tamilnadu, India (“UCI Machine Learning Repository: Chronic_Kidney_Disease Data Set,” n.d.). For early detection of kidney disease, we use 24 + class = 25 (11 numeric ,14 nominal). shows the number of missing values in this dataset. However, I chose the particular parameters for data analysis which are summarized in table 3.3

Table 3.3 Parameters for data analysis

No	Attributes	Indication	Description
1	Age	Numerical	Years
2	Blood pressure	Numerical	Mm/Hg
3	Specific gravity	Nominal	1.005,1.010,1.015,1.020,1.025
4	Albumin	Nominal	0.1.2.3.4.5
5	Sugar	Nominal	0.1.2.3.4.5
6	Red blood cells	Nominal	Normal or Abnormal
7	Pus cell	Nominal	Normal or Abnormal
8	Pus cell clumps	Nominal	Present or Not present
9	Bacteria	Nominal	Present or Not present
10	Blood glucose random	Numerical	Mgs/dl
11	Blood urea	Numerical	Mgs/dl
12	Serum Creatinine	Numerical	Mgs/dl
13	Sodium	Numerical	mEq/L
14	Potassium	Numerical	mEq/L
15	Haemoglobin	Numerical	Gms
16	Packed cell volume	Numerical	Gms
17	White blood cell count	Numerical	Cells/cmm
18	Red blood cell count	Numerical	Millions/cmm
19	Hypertension	Nominal	Yes or No
20	Diabetes mellitus	Nominal	Yes or No
21	Coronary artery disease	Nominal	Yes or No
22	Appetite	Nominal	Good or Poor
23	Pedal edema	Nominal	Yes or No
24	Anemia	Nominal	Yes or No
25	Class	Nominal	CKD or NOCKD

3.3 Data Preprocessing

In this section, I conducted several experiments to reduce missing values, redundant values. Therefore, analyzing the attributes of the selected chronic datasets, some of them presented a very few values whereas others appeared not correlated with the specific medical event. Hence, I removed those elements from the initial features. However, after the preliminary cleansing phase, some of the datasets showed many missing values in presence of the remaining attributes of the chronic disease datasets. The overall process goes through a sequence of data preprocessing steps, as detailed in the following.

3.3.1 Kidney Disease

In this datasets (CKD), contains 400 Kidney patient's data including 25 parameters. In the first step, I removed 'Id' parameter from the data set. The reason of remove this parameter is it will not work for classification methods. However, there is no value called "ckd\t " within this dataset so

that I replace it by 'ckd' value. Therefore, I changed the target factor values to 1(ckd) and 0 (not ckd) to be able to get the good performance of machine learning algorithms. Figure 3.3 shows the number of missing values in this dataset. Hence, I used dropna() function to remove all rows with missing values. Then, Figure 3.2 shows the number of missing values is empty. After cleaning the missing value, there are 158 samples including 25 parameters. The number of samples reduced but the consistency of the classification model increased. I corrected and changed some of the parameters and data types to create tidy datasets. Moreover, the CKD datasets was also checked to verify the correlation of parameters. The heatmap shown in figure 3.3 appear to have no correlated parameters. And figure 3.4 shows the Blood Press and Hemoglobin values according the age from the CKD datasets.

```
In [4]: data.isnull().sum()
Out[4]: id          0
        age         9
        bp         12
        sg         47
        al         46
        su         49
        rbc        152
        pc         65
        pcc         4
        ba          4
        bgr        44
        bu         19
        sc         17
        sod        87
        pot        88
        hemo       52
        pcv        70
        wc         105
        rc         130
        htn         2
        dm          2
        cad         2
        appet       1
        pe          1
        ane         1
        classification 0
        dtype: int64
```

Figure 3.2 Number of missing values in CKD datasets

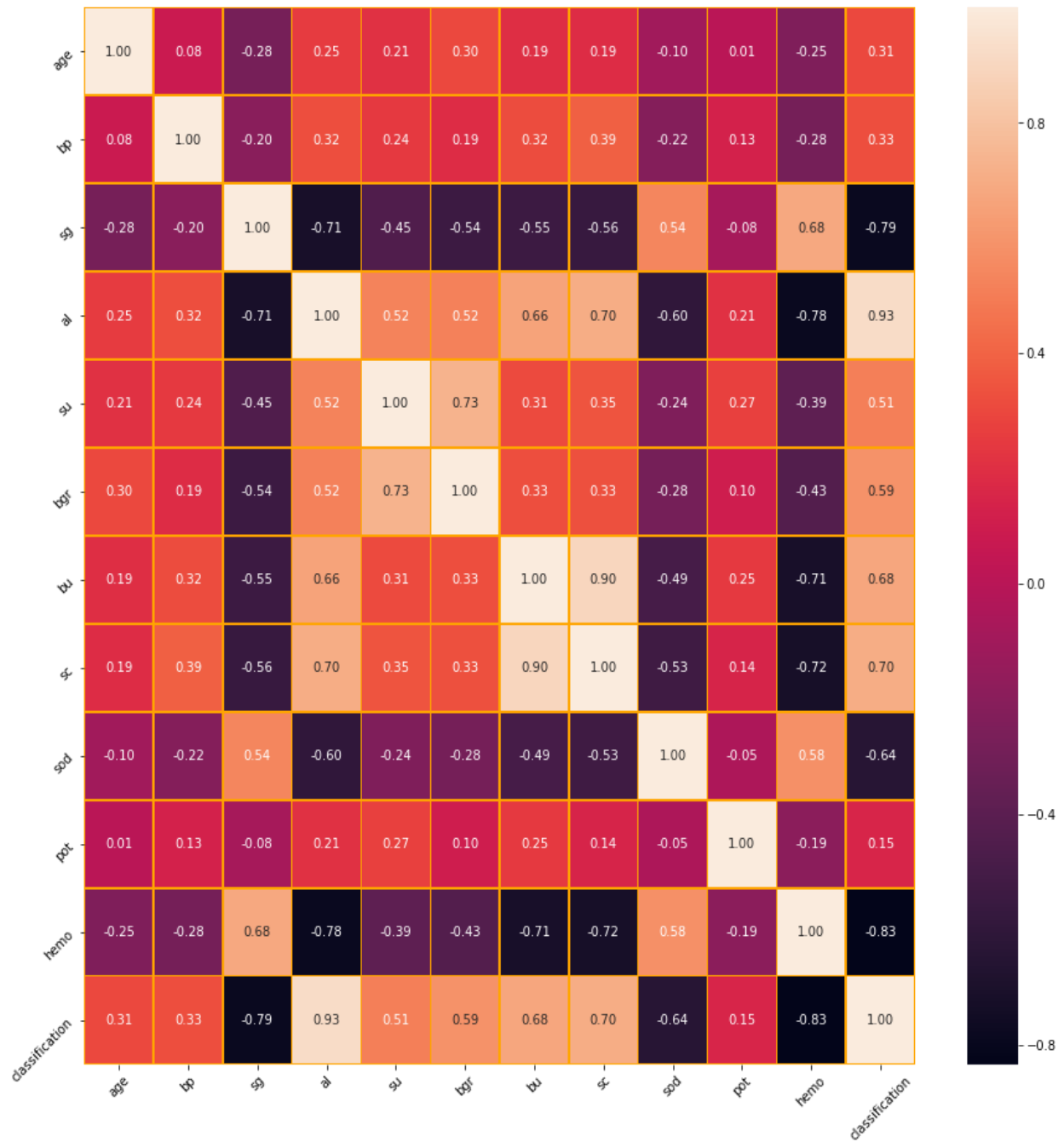


Figure 3.3 Heat map for checking correlated columns in CKD

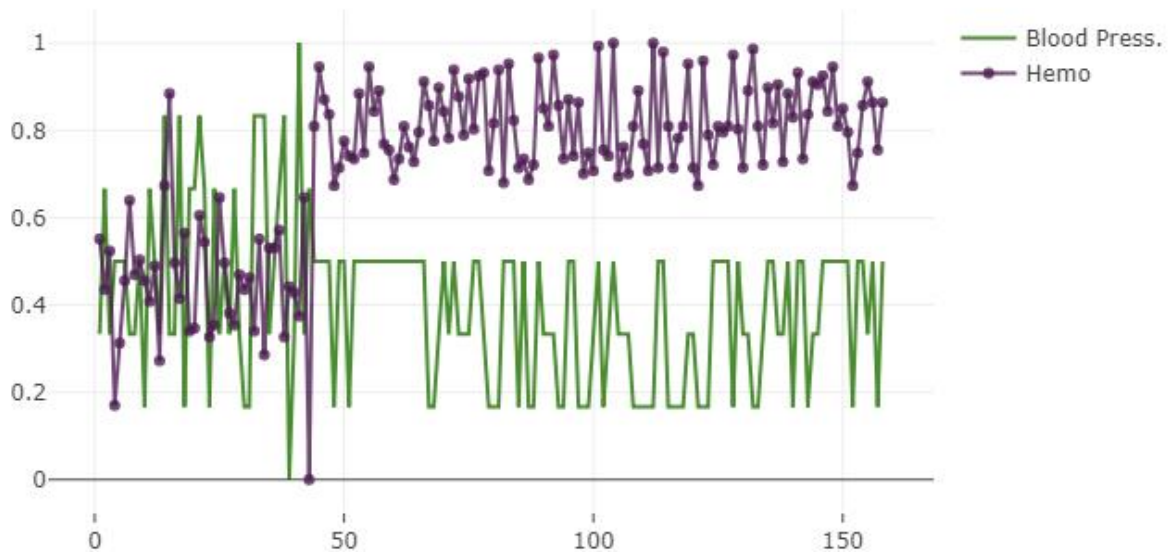


Figure 3.4 Blood Press and Hemoglobin values in CKD

3.3.2 Breast Cancer

From the Wisconsin Breast Cancer data, it contains 699 breast cancer patient's data including 11 parameters. I see that the column 'class' has high correlation with all columns except ID Number which has no significance and needs to be removed. Therefore, I removed 'Id' parameter from the data set. If the 'ID number' column is not removed, the heatmap shown in the figure 3.5 that the accuracy is affected when I conducted the analysis. In this dataset, there is no missing data that shown in the figure 3.6. But there are some NaN values, it is denoted by '?'. Therefore, we used the `dropna()` function to remove the NaN values. After cleaning the datasets, we have 683 entries including 10 parameters. Hence, the heatmap shown in figure 3.7 appear to have no correlated parameters.

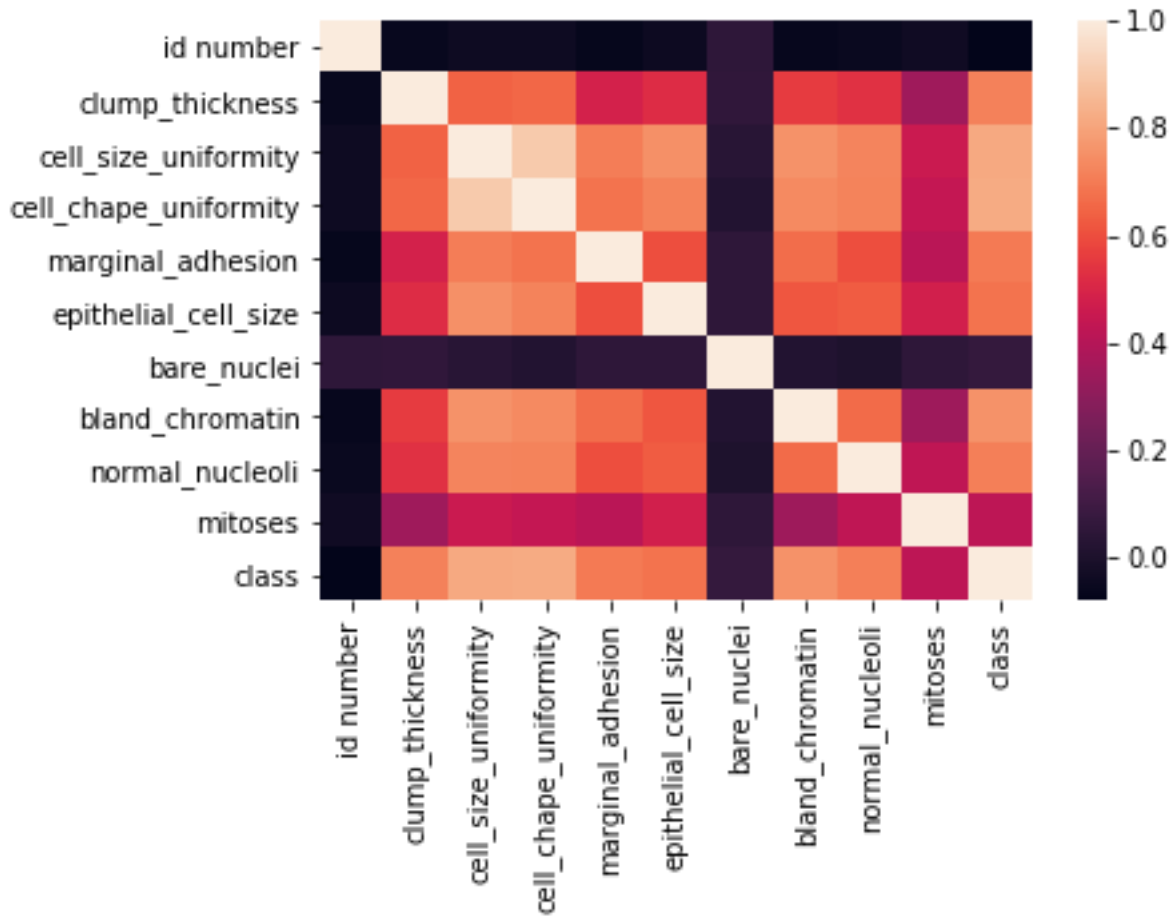


Figure 3.5 The heatmap shown that the accuracy is affected for the 'ID number'

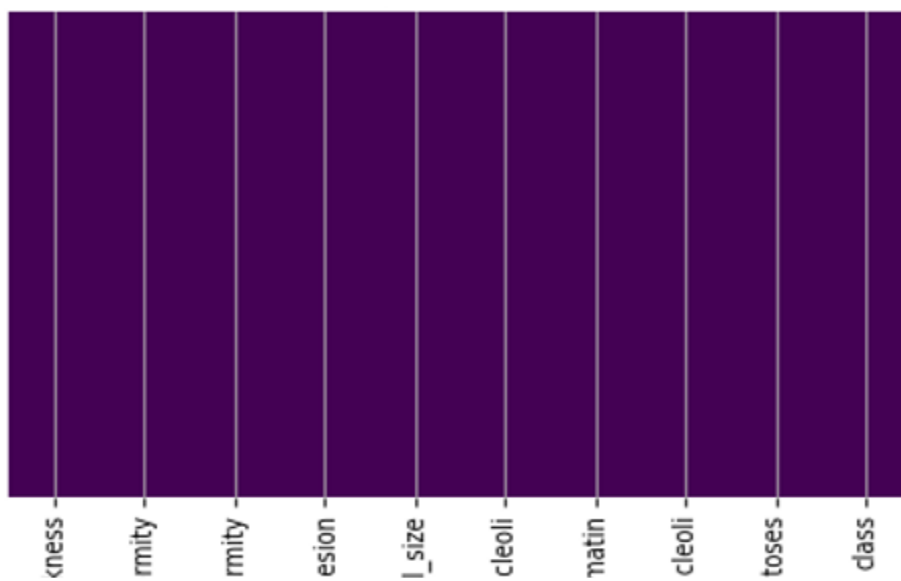


Figure 3.6 No missing values on breast cancer datasets

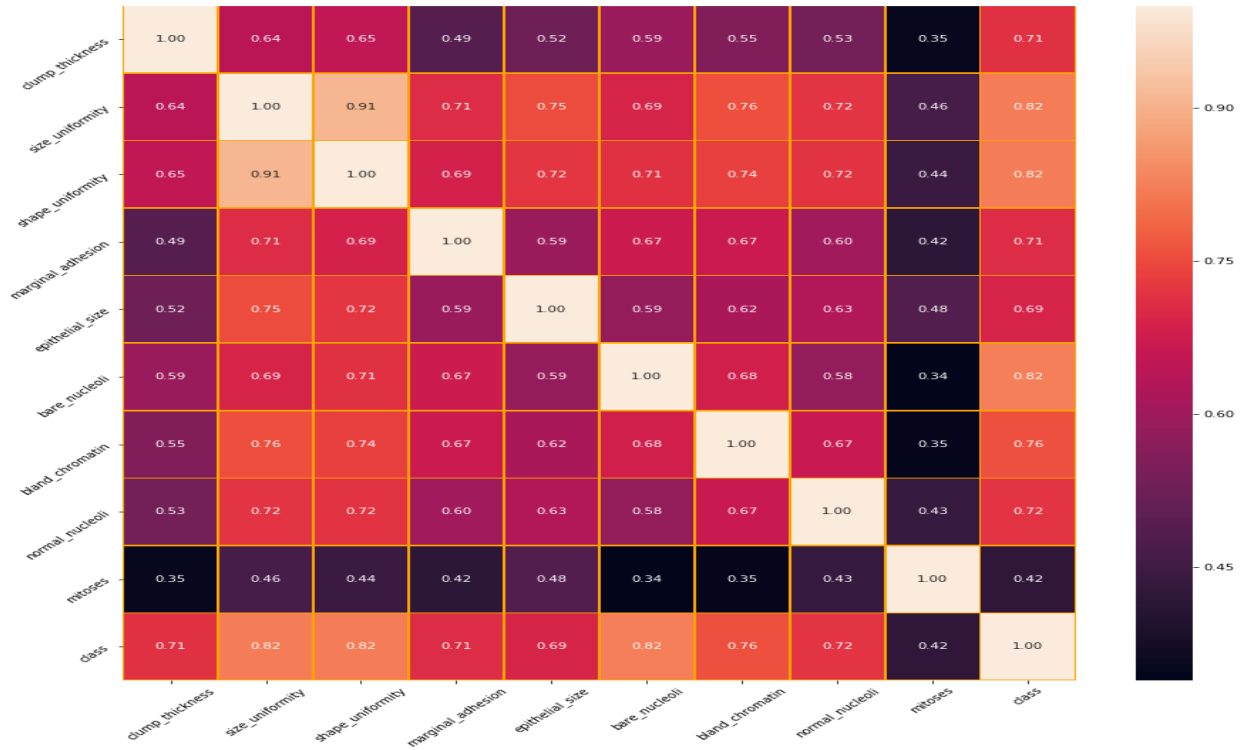


Figure 3.7 Heat map for checking correlated columns for breast cancer

3.3.3 Liver Disease

This data set contains 416 patients diagnosed with liver disease and 167 patients not diagnosed with liver disease. Figure 3.8 shown that the ratio of the liver patients. In this dataset, there is most of the patient are male, Fig 3.9 shown that the ratio of gander the liver patients (the number of male patients is 441 and 142 are female).

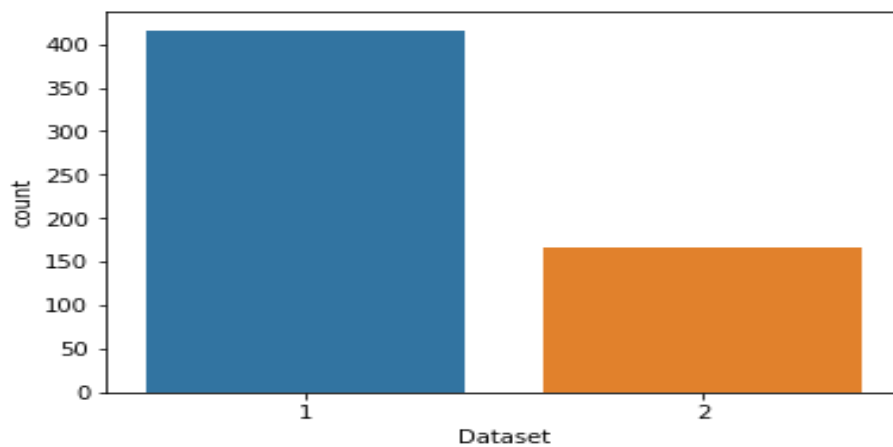


Figure 3.8 Count plot shows the ratio of the liver patients

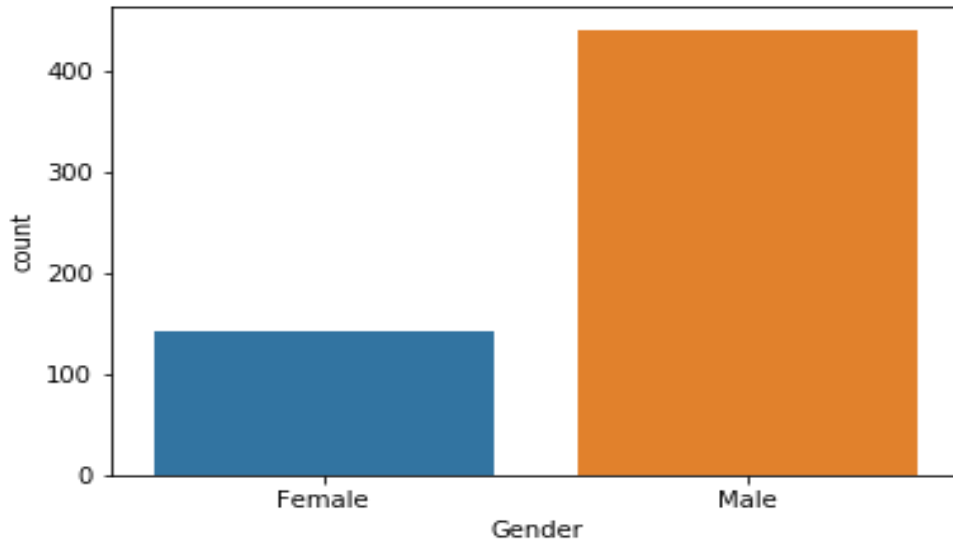


Figure 3.9 Count plot shows the ratio of gender the liver patients

There are 10 features and 1 output - dataset. Value 1 indicates that the patient has liver disease and 0 indicates the patient does not have liver disease. Figure 3.10 shown that the age seems to be a factor for liver disease for both male and female genders. Therefore, figure 3.11 shown that there seems to be direct relationship between Total Bilirubin and Direct Bilirubin. We have the possibility of removing one of these features.

```
In [8]: sns.catplot(x="Age", y="Gender", hue="Dataset", data=liver_df)
```

```
Out[8]: <seaborn.axisgrid.FacetGrid at 0x8f9cc71a58>
```

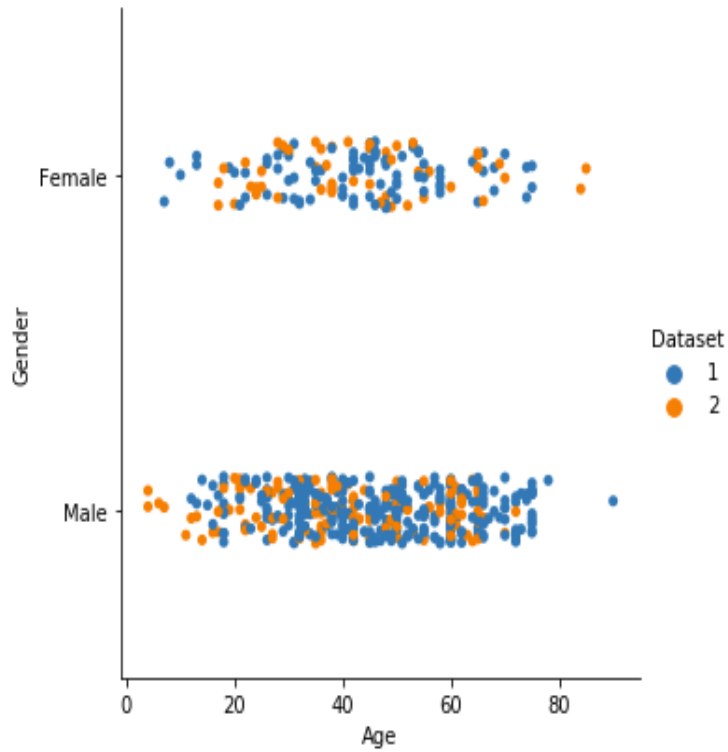


Figure 3.10 Factor plot of liver disease for both male and female

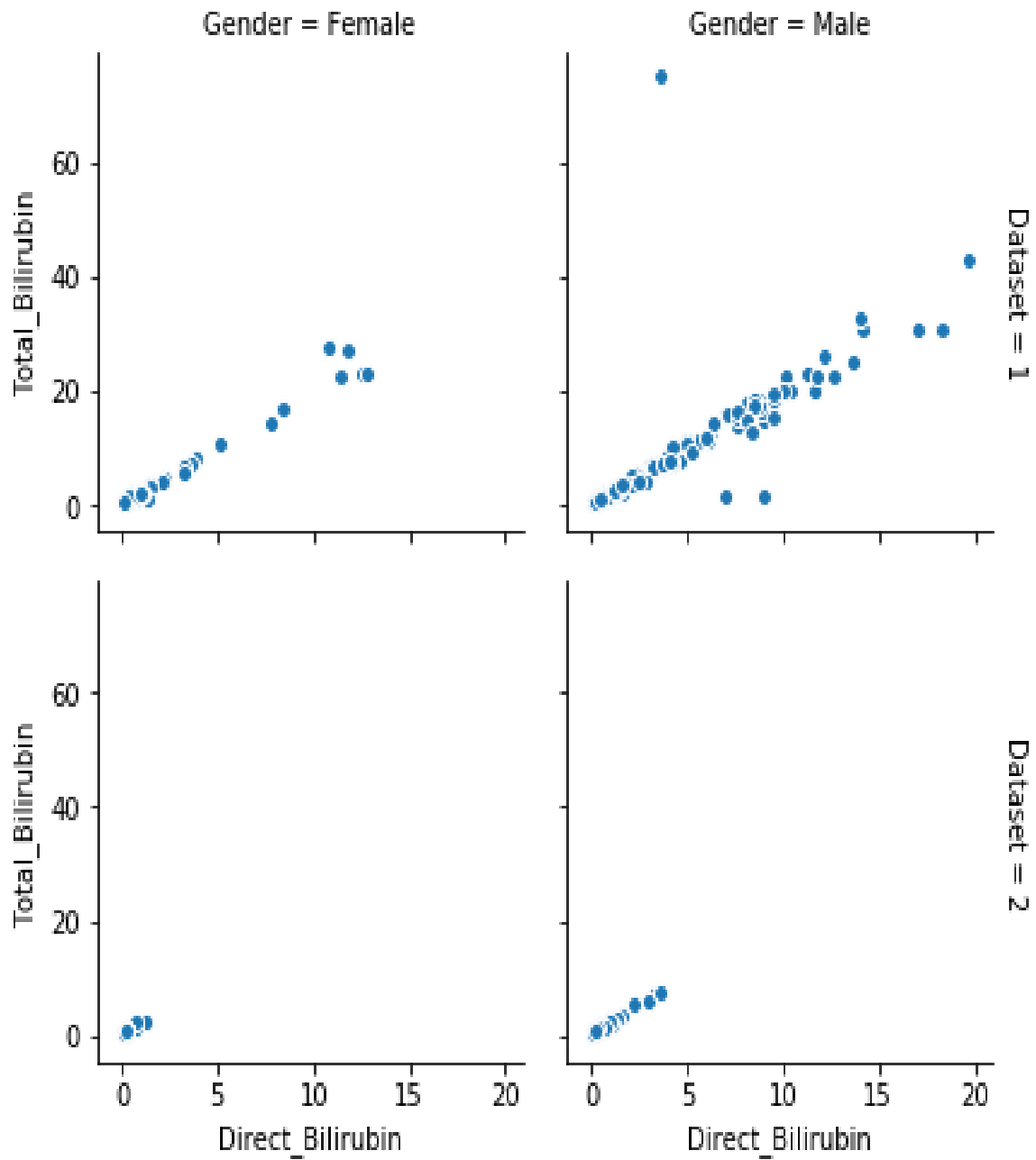


Figure 3.11 Direct relationship between Total Bilirubin (mg/dl) and Direct Bilirubin(mg/dl)

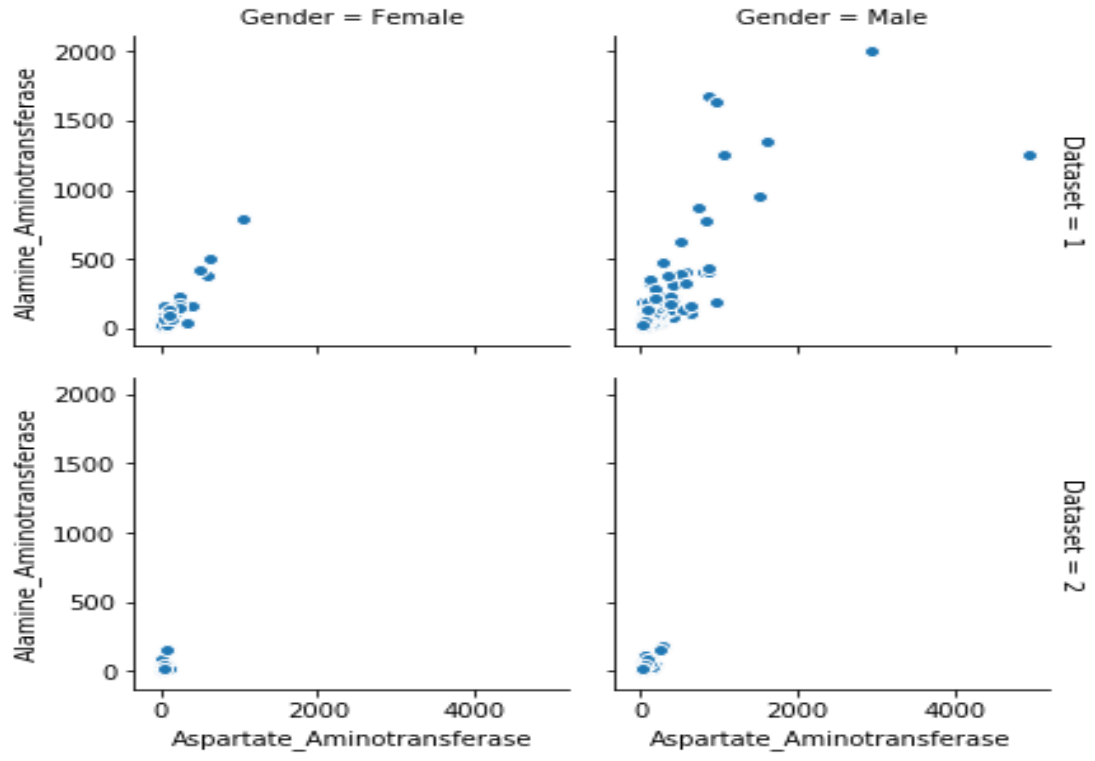


Figure 3.12 Direct relationship between Alamine_Aminotransferase (ALT in IU/L) and aspartate_Aminotransferase (ALT in IU/L)

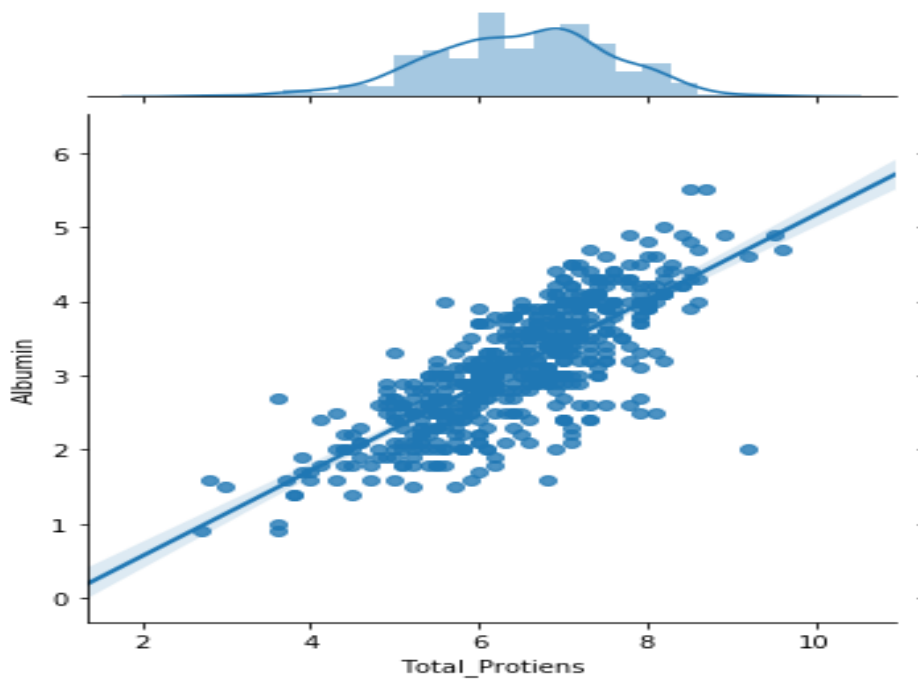


Figure 3.13. Direct relationship between Albumin (g/dl) and Total_Protiens (g/dl)

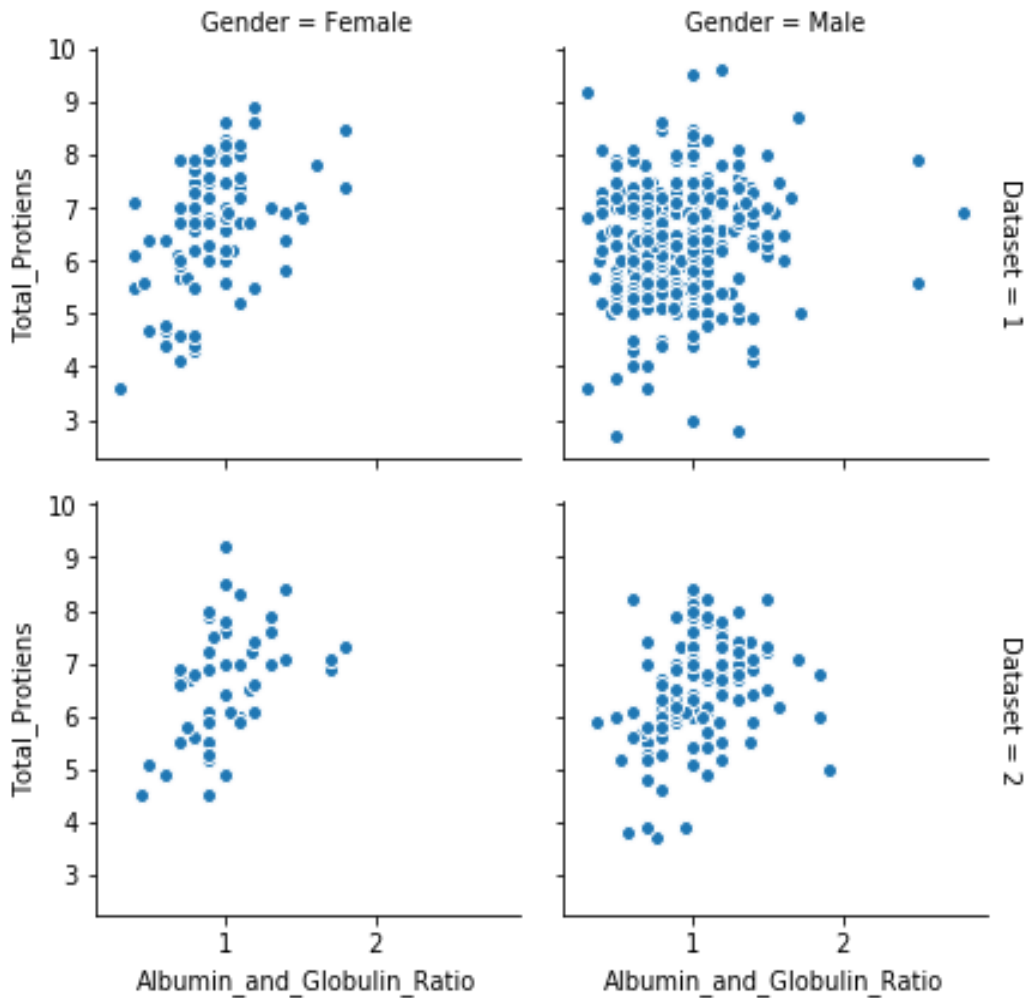


Figure 3.14. Direct relationship between Total_Protiens (g/dl) and Albumin_and_Globulin_Ratio (A/G ratio)

From the above joint plots and scatterplots (Figure 3.11 - 3.14), I found direct relationship between the different features. Hence, I omitted some of the features for better prediction of liver disease and the following features are consider for model: Alamine_Aminotransferase, Total_Protiens, Albumin_and_Globulin_Ratio, Albumin, Total Bilirubin. Figure 3.15 shows that the liver disease by gender and age. And Figure 3.16 shows the number of missing values in this dataset.

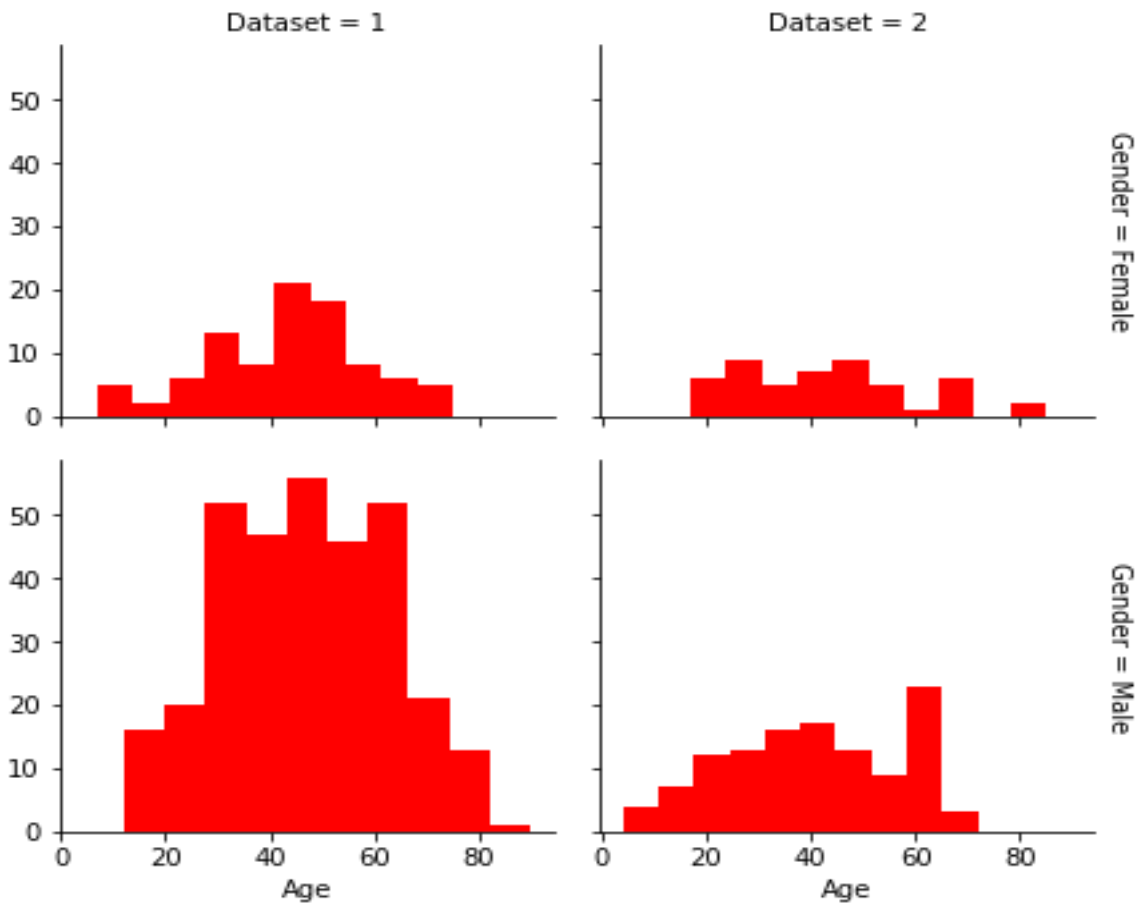


Figure 3.15 Disease by Gender and Age (years)

```
In [14]: liver_df.isnull().sum()
Out[14]: Age                0
         Gender              0
         Total_Bilirubin    0
         Direct_Bilirubin   0
         Alkaline_Phosphotase 0
         Alamine_Aminotransferase 0
         Aspartate_Aminotransferase 0
         Total_Protiens     0
         Albumin            0
         Albumin_and_Globulin_Ratio 4
         Dataset            0
         dtype: int64
```

Figure 3.16 Number of missing values for liver datasets

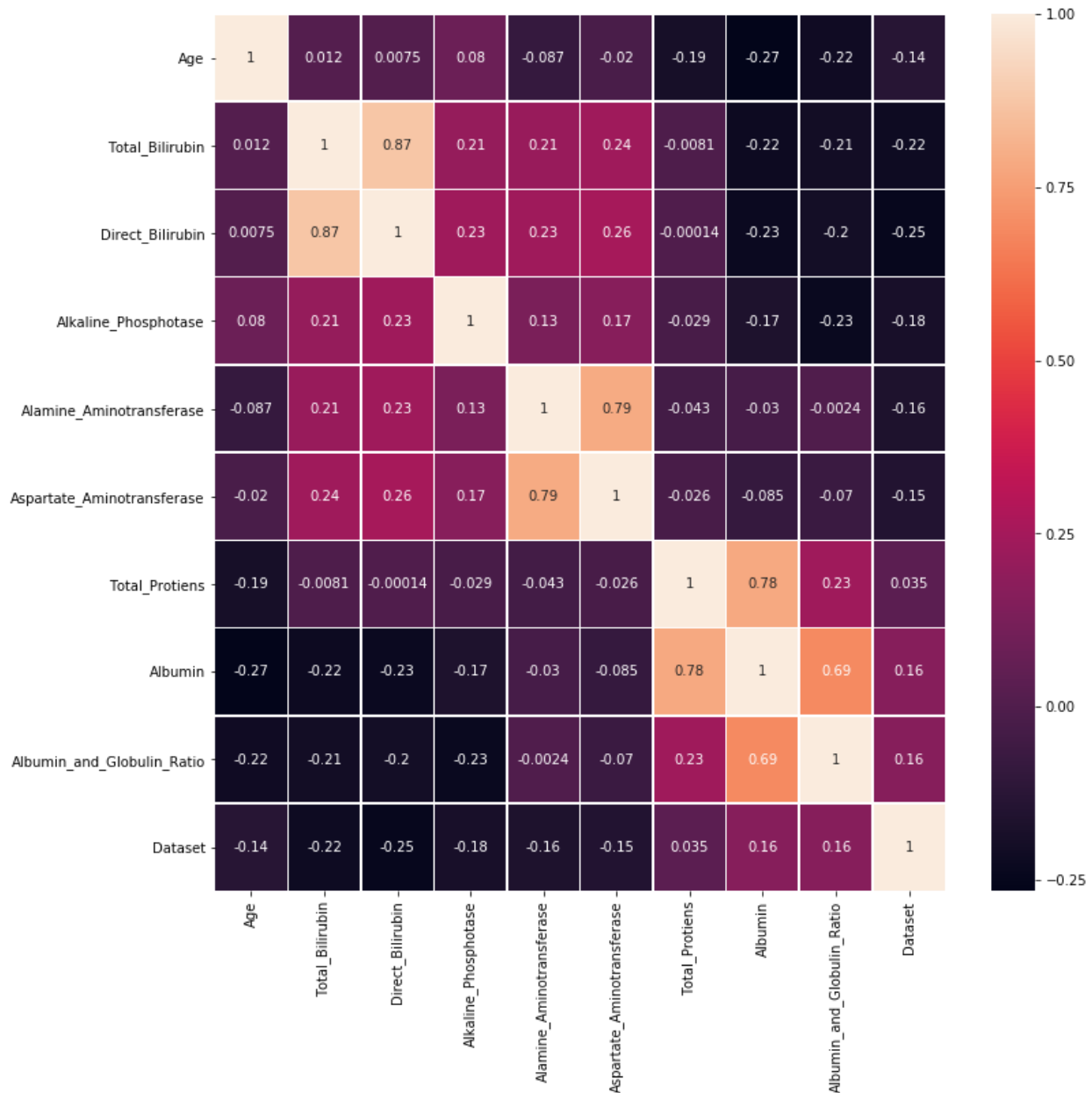


Figure 3.17. Heat map for checking correlated columns for liver datasets

Hence, the heatmap shown in figure 3.17 appear to have some correlated parameters. The heatmap indicates the following correlation: Total_Protiens and Albumin, Alamine_Aminotransferase and Aspartate_Aminotransferase, irect_Bilirubin and Total_Bilirubin, Total_Protiens & Albumin. There is some correlation between Albumin_and_Globulin_Ratio and Albumin. I see that the experiment, some of the column has low correlation.

3.4 Classification Techniques

3.4.1 Logistics Regression

Logistic Regression was mostly used in the biological research and applications in the early 20th century (Jr, Lemeshow, & Sturdivant, 2013). Logistic Regression (LR) is one of the most used machine learning algorithms that is used where the target variable is categorical. Recently, LR is a popular method for binary classification problems. Moreover, it presents a discrete binary product between 0 and 1. Logistic Regression computes the relationship between the feature variables by assessing probabilities (p) using underlying logistic function.

3.4.2 Support Vector Machine (SVM)

Support vector machine has been first introduced by Vladimir Vapnik and Alexey Chervonenkis (Chervonenkis, 2013)(Vapnik, Guyon, Learn, & 1995, n.d.). SVM is a method of machine learning that can solve both linear and nonlinear problems. It provides good performance to solve both regression and classification problem. The SVM classification technique inspects for the optimal separable hyperplane in order to classify the dataset between two classes (Smola & Schölkopf, 2004). Finally, the model can estimate noisy data problems for new cases.

3.4.3 Decision Tree (DT)

Decision tree is one of the well-known supervised learning algorithm of machine learning. DT is a classification technique that divides the dataset into smaller subsets. Here, each branch represents a decision and each leaf represents an outcome. In decision tree, root of the tree is used to estimate the entropy, i.e., the information gain (Safavian & Landgrebe, 1991).

3.4.4 Random Forest (RF)

Leo Breiman first introduced random forest in his study (Breiman, 2001). Random Forest algorithm is a popular algorithm in machine learning. RF work well for many clinical and biological problems. It is able to solve both classification and regression problems in health care services. It creates a forest by different ways and make it random. In the forest of trees has been

the direct relationship between the combine trees and the result it can get. To acquire more efficient and accurate prediction, random forest inserts an extra layer of randomness to bagging.

3.4.5 Naïve Bayes (NB)

Naive Bayes is one of the simple, most effective and commonly-used, machine learning technique. It is a probabilistic classifier that classifies using the hypothesis of a conditional independence with the pre-trained datasets (Leung, 2007).Henceforth, Naive Bayes classifiers are techniques for finding the traditional solution of classification problems, such as spam detection, and also well fit for medical problems.

3.4.6 k- Nearest Neighbors (KNN)

The K-Nearest Neighbors is one of the most basic instance-based classification algorithms in Machine Learning. However, the KNN works on the concept that samples are close to fit in the same samples class (Zhang & Zhou, 2007). A KNN categorizes a sample to the class that is most determined among K neighboring. K is constraint for fine-tuning the classification algorithms (Guo, Wang, Bell, Bi, & Greer, 2003).

3.5 Evaluation Criteria

In this thesis, we used six machine learning techniques for the early detection of chronic disease. Therefore, the performance measurements of the classifiers is appraised by different statistical procedures. Such as confusion matrix (True Positive, False Positive, True Negative, False Negative), Recall, Precision, f1- measure etc (“Confusion Matrix,” n.d.). Hence, the validation matrix is defined by,

True Positive (TP): Prediction results are true and the patient has chronic disease

True Negative (TN): Prediction results are false and the patient does not have chronic disease.

False Positive (FP): Prediction results are true but the patient does not have chronic disease.

False Negative (FN): Prediction results are false and the patient has chronic disease.

The computation method of the measurement considerations are as follows,

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{FP} + \text{TN} + \text{FN}) \quad (1)$$

$$\text{True Positive Rate or Sensitivity or Recall} = \text{TP} / (\text{TP} + \text{FN}) \quad (2)$$

$$\text{Specificity or True Negative Rate or TN} / (\text{TN} + \text{FP}) \quad (3)$$

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \quad (4)$$

$$f1 = 2 * (\text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision}) \quad (5)$$

$$\text{False Positive Rate} = 1 - \text{Specificity} \quad (6)$$

The f1-measure is stated by the weighted norm of the recall & precision. For the better performance of the supervised classifiers, the value must be 1 and for the poorest performance, it must be zero

3.6 Software and Tools

In the present study all analysis was implemented in Python version 3.7.0 using Anaconda Distribution including Jupyter Notebook. The version of the notebook server is: 5.6.0-3badce9.

CHAPTER FOUR

RESULTS & DISCUSSION

4.1 Analysis of the Results

In this experiment, I conducted different analysis to evaluate the six machine learning classification algorithms for diagnosis and prediction of chronic disease. The performance comparison and performance measure of six machine learning classifiers for chronic disease prediction as detailed in the following.

4.1.1 Chronic Kidney Disease

The prediction of six computational intelligence techniques were examined for the classification of chronic kidney sample data. Figure 4.1 shows the performance of six supervised classification techniques. Here, NB and RF outperformed than the other classification techniques in terms of accuracy by obtaining the highest accuracy as 100% respectively.

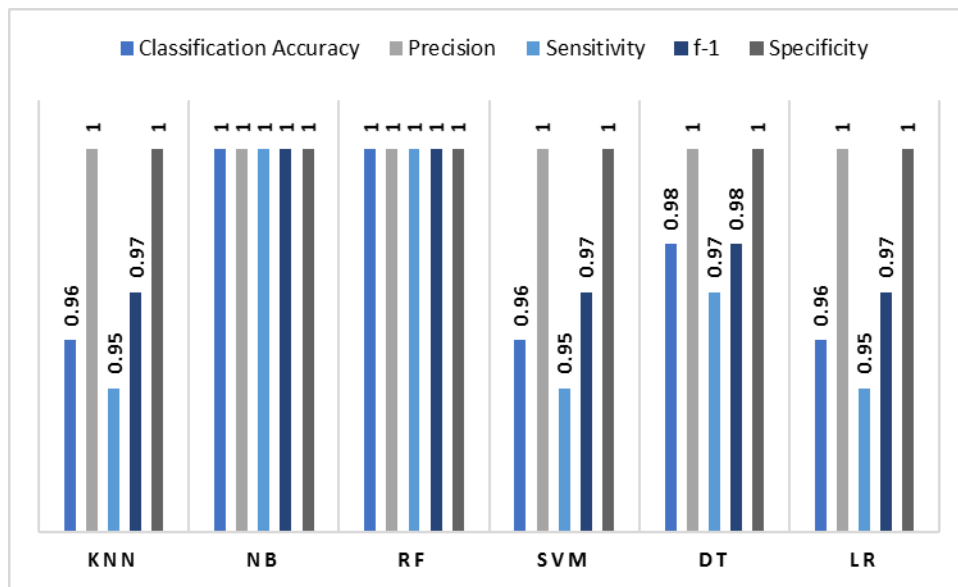


Figure 4.1 Performance of six supervised classification techniques

However, the other classification techniques achieved the quite good performance, having the accuracy of 96% achieved by KNN, SVM, LR. In this experiment, DT achieved the second

highest accuracy (98%). Table 4.1 shows the classification performance measurements of six classification techniques.

Table 4.1 Classification performance measurements

	KNN	NB	RF	SVM	DT	LR
Classification Accuracy	0.96	1	1	0.96	0.98	0.96
Precision	1	1	1	1	1	1
Sensitivity	0.95	1	1	0.95	0.97	0.95
f-1	0.97	1	1	0.97	0.98	0.97
Specificity	1	1	1	1	1	1

algorithms with maximum precision, sensitivity, f-1, and specificity. However, Other classifiers show the performance in every measurement above 90% respectively. Confusion matrix of prediction results is presented in figure 4.2.

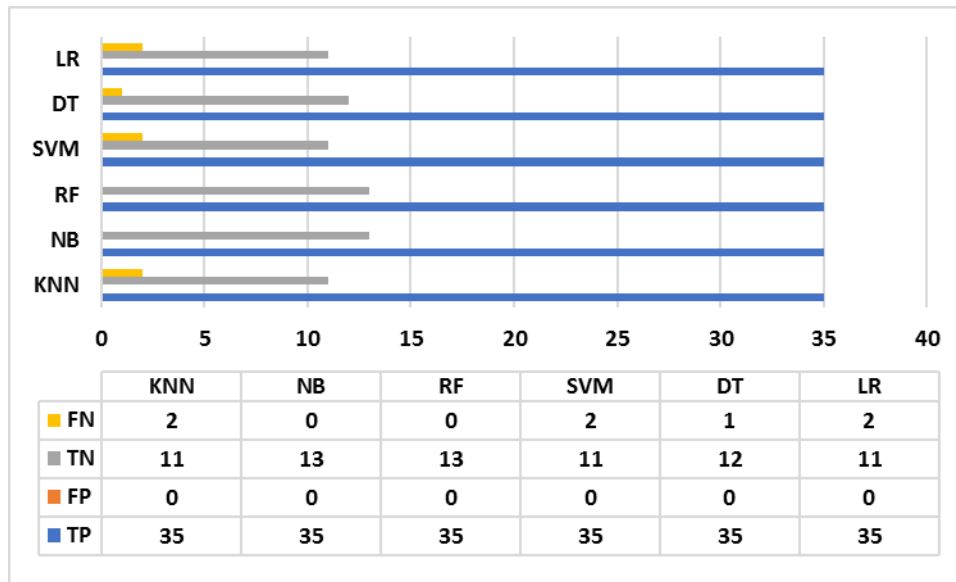


Figure 4.2 Confusion matrix of classifiers

4.1.2 Breast Cancer

The performance of six machine learning shows in figure 4.3. SVM achieved the highest performance with maximum classification accuracy of 97.07% while second highest classification accuracy is achieved by NB and RF (97%) whereas KNN, DT and LR shows the almost same performance by attaining 96% accuracy.

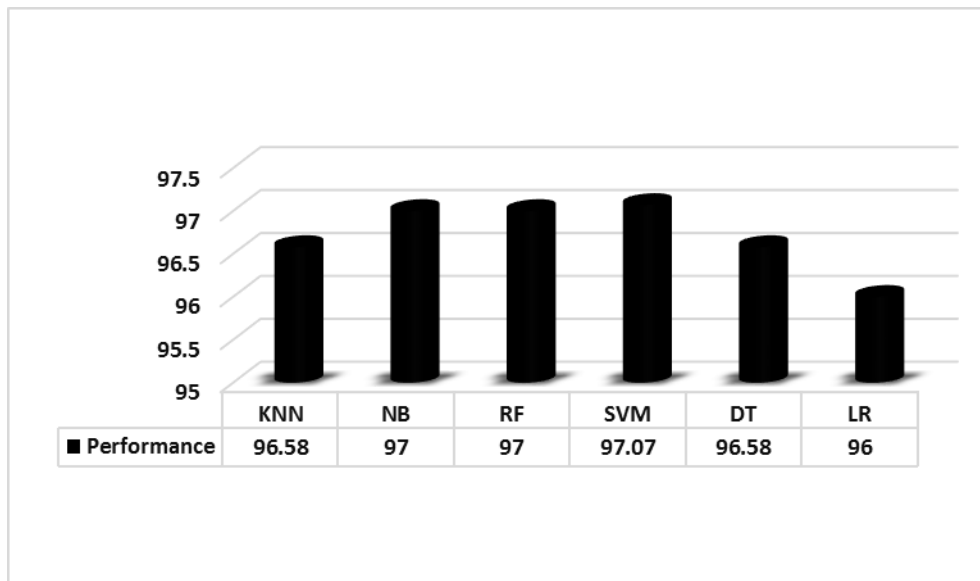


Figure 4.3 The accuracy of six machine learning (Breast Cancer)

According to the performance measurements of six classification algorithms are presented in figure 4.4. The results clearly show that the DT and LR reached to the highest precision (97%). NB achieved the highest sensitivity, it's 100%. And NB also achieved the worst specificity (92%). Considering f1 measure, all of classifiers shows the same performance, it's above 95%, respectively. Figure 4.5 shows the confusion matrix of prediction results for Naïve Bayes, Random Forest, Support Vector Machine, Decision Tree, KNN and Logistics Regression algorithms.

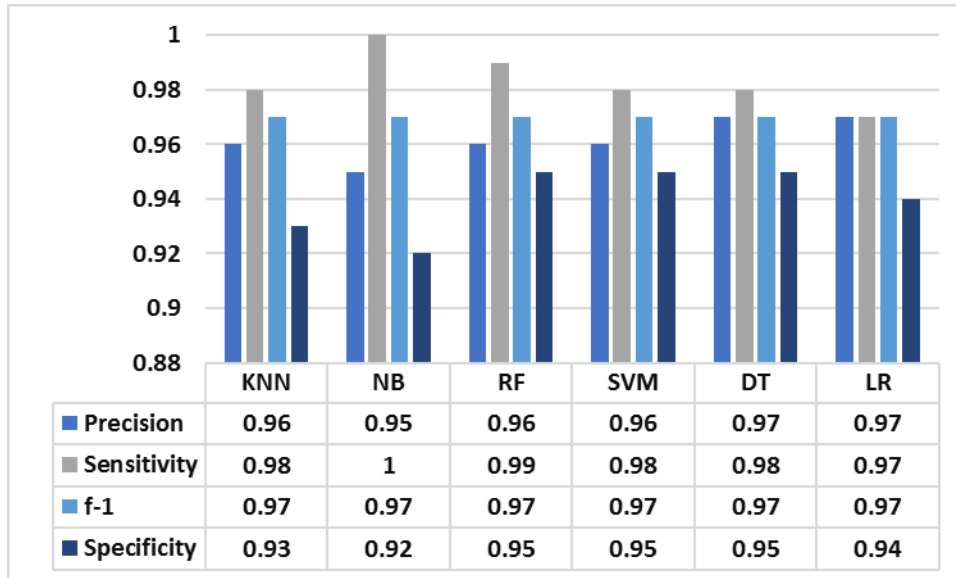


Figure 4.4. Classification Performance Measurements (Breast cancer)

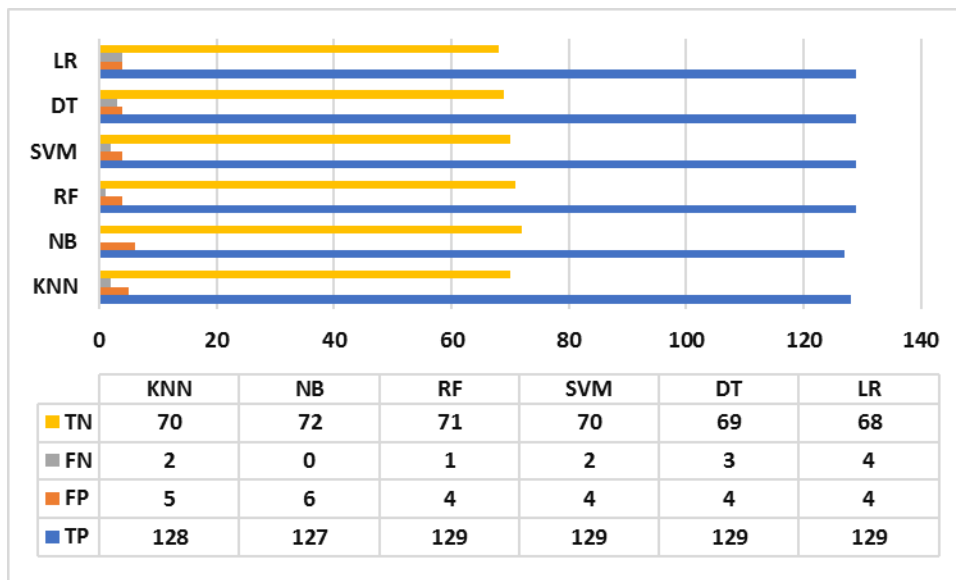


Figure 4.5. Confusion matrix of classification techniques

4.1.3 Liver Disease

In this experiment, we consider different analysis to examine the six machine learning classification techniques for the classification of chronic kidney datasets (CKD). Figure 4.6 shows the performance comparison of six supervised learning techniques for kidney disease prediction. With respect to precision LR achieved the highest score, 91% and NB perform worst (i.e. 0.36%). However, when considering the sensitivity, SVM achieved the highest value (i.e. 0.88%) and KNN

obtained the worst (i.e. 0.76%). Logistics Regression was also the best performer in terms of fi measure (i.e. 0.83%) and NB obtained the worst performance (i.e. 0.53%). In the terms of accuracy, LR achieved the highest accuracy (i.e. 0.75%) and NB achieved the worst performance (i.e. 0.53%). LR indicates that this classification technique is more effective than the other classifiers for predicting chronic kidney disease. Figure 4.7 shows the confusion matrix of prediction results for the six classifiers.

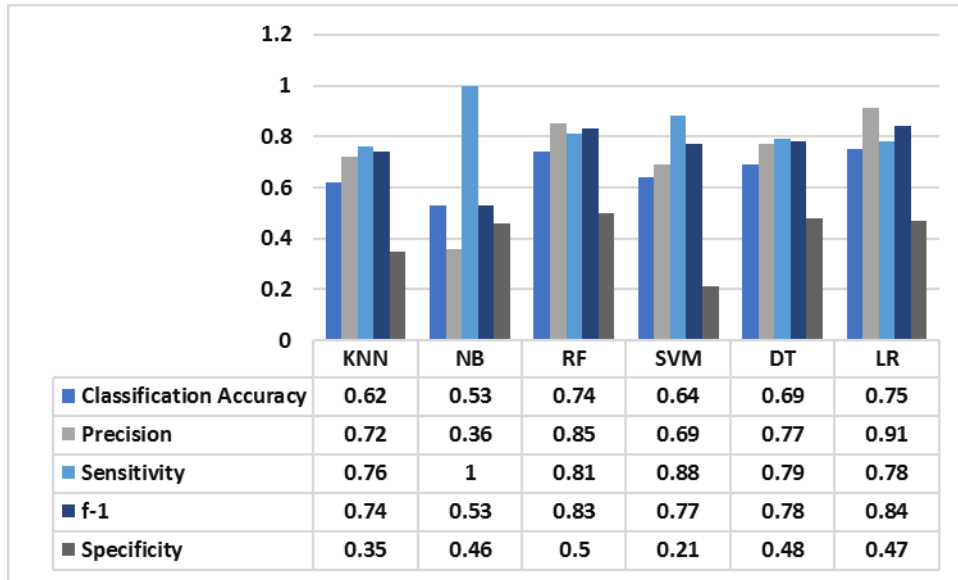


Figure 4.6 The performance comparison of six supervised learning techniques

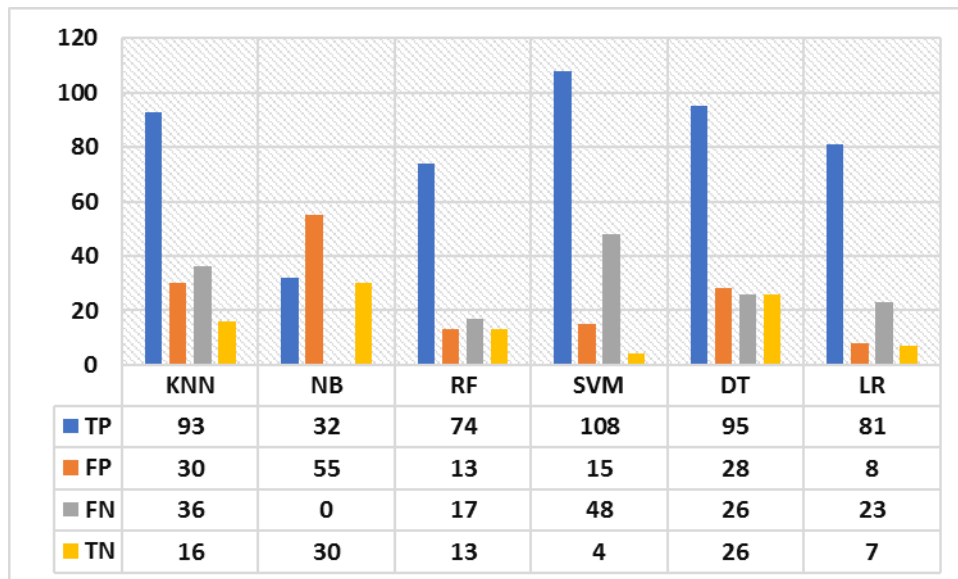


Figure 4.7 The confusion matrix of prediction results

4.2 Performance Evaluation

All the machine learning classifiers show the accuracy level above 95% for kidney disease prediction and breast cancer classification. Hence, the accuracy level nearly of 75% for liver disease prediction using all classification classifiers. Which indicates that the performance of these classification techniques is excellent for breast cancer and kidney disease prediction. Moreover, the six machine learning algorithms are pretty good for liver disease prediction. From the above discussion, it is very important to know about the receiver operating characteristics (ROC) curve, which is based on true positive rate (TPR) and false positive rate (FPR) of these classification results. According to ROC curve, RF and NB outperformed (Kidney Disease) all other techniques. Furthermore, KNN (Breast Cancer) and SVM (Liver Disease) achieved highest AUC (area under curve) for ROC. The ROC curve is presented in figure 4.8-4.10. In addition, The false discovery rate (FDR) is the projected proportion of type I errors (“False Discovery Rate: Simple Definition, Adjusting for FDR - Statistics How To,” n.d.). A type I error is where the value incorrectly discards the null hypothesis. Moreover, False omission rate (FOR) is a statistical method used in multiple hypothesis testing to correct for multiple comparisons and it is the complement of the negative predictive value. FOR measures the proportion of false negatives which are incorrectly rejected (“False Omission Rate - calculator - fxSolver,” n.d.). Figure 4.11 and 4.12 shows the FDR and FOR, respectively.

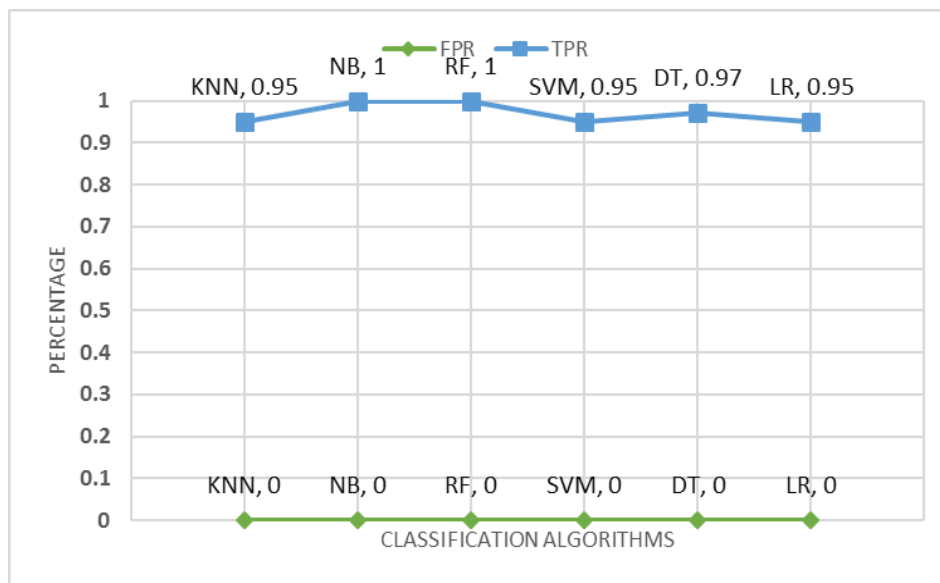


Figure 4.8 Receiver Operating Characteristics curve for kidney datasets

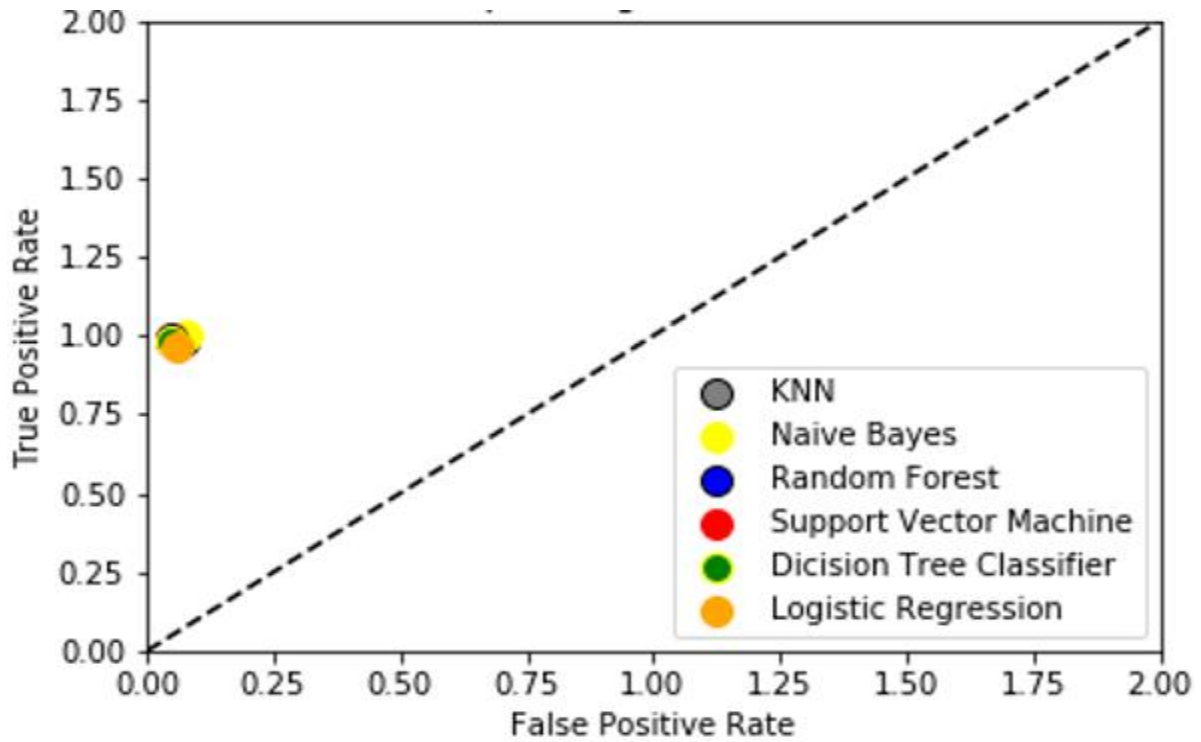


Figure 4.9 Receiver Operating Characteristics curve for Breast Cancer datasets

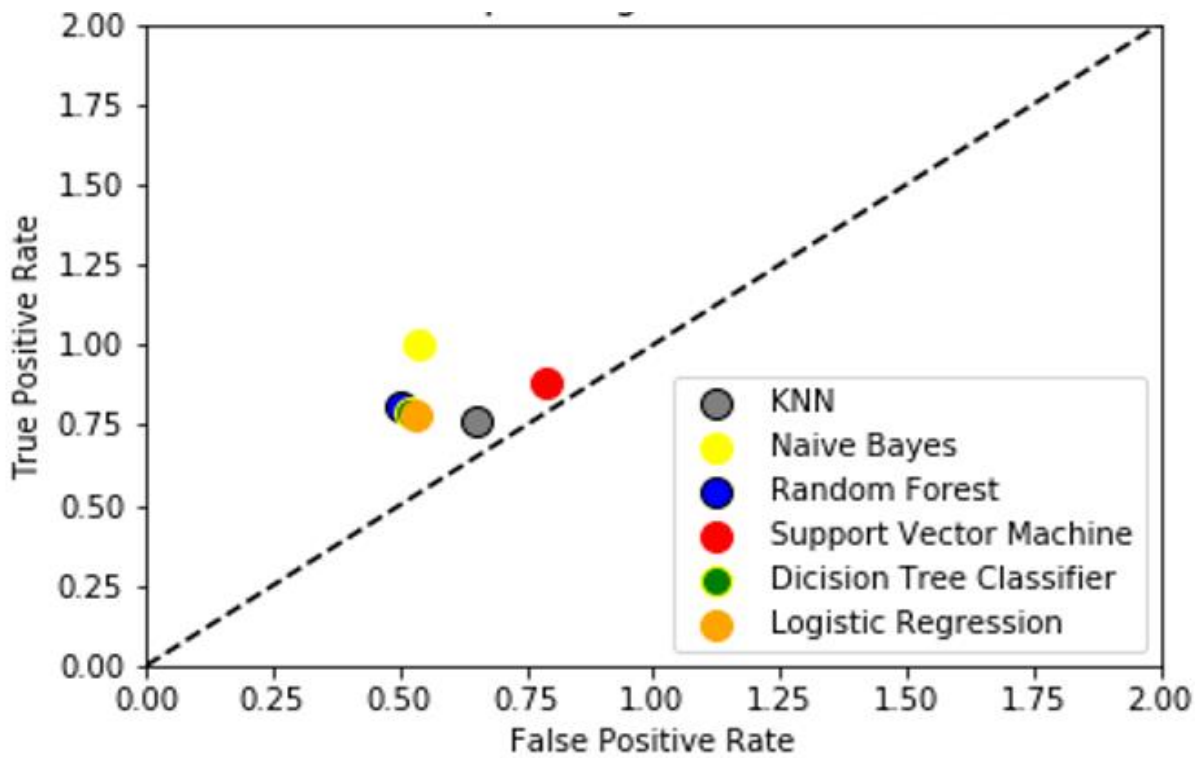


Figure 4.10 Receiver Operating Characteristics curve for Liver datasets

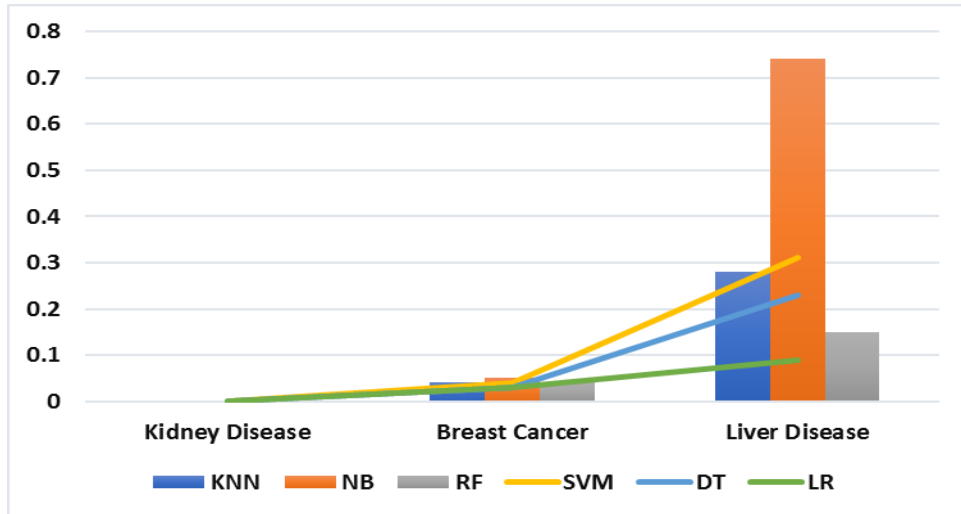


Figure 4.11 False Discovery Rate (FDR)

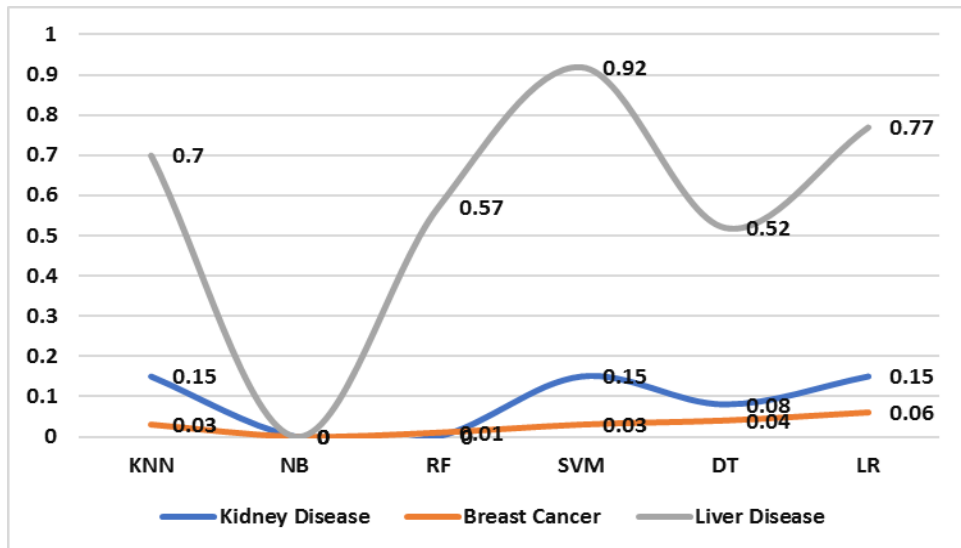


Figure 4.12 False Omission Rate (FOR)

CHAPTER FIVE

CONCLUSION & RECOMMENDATION

5.1 Findings and Contribution

In this study, I have described six supervised learning-based machine learning techniques. Therefore, I provide a workflow on machine learning based application for the early detection and monitoring of chronic disease. Afterwards, I compared the performance of the six classifiers which are used in the prediction of chronic kidney diseases and evaluated their performance using confusion matrix. The experimental performance shows that the Naïve Bayes and Random Forest has achieved the outperform than the other classifiers within kidney and cancer datasets. It is 100%, respectively. Hence, LR have achieved the best accuracy (i.e. 75%) on liver datasets. This examination has used six machine learning techniques for the early detection of chronic disease based on several parameters. In addition, this study is part of a project that has the purpose to develop a real time-based computerized tool to give more precise treatment to normal events and make a superior decision to complex situations. The application will be able to early detect in chronic disease in a few minutes and notify the real condition with extreme likelihood of having disease. This application can be remarkably beneficial in low-income countries where is lack of medical institutions and as well as specialized practitioners.

5.2 Recommendation for Future Work

In my experiments, related to most work in the study, each classification algorithms were trained and evaluated on a training set that includes both positive and negative samples. Moreover, the work can be helpful for chronic disease diagnosis and detection by collecting data from different devices and health related sensors, clinical and medical center and can deliver more accurate results for disease prediction and diagnosis. In my research perspective, there are several directions for the future work in this area of research. I only investigated to some popular supervised machine learning algorithms, it can be choosing more algorithm for build the accurate model of these chronic disease prediction and performance can be more improved. In summary, I have highlighted the research direction and scope in relation to Health Care Services and Bio-medical fields by machine learning techniques, which has emerging impact in medical fields.

References

- Abdar, M., Zomorodi-Moghadam, M., Das, R., & Ting, I.-H. (2017). Performance analysis of classification algorithms on early detection of liver disease. *Expert Systems with Applications*, *67*, 239–251. <https://doi.org/10.1016/J.ESWA.2016.08.065>
- Abdelaziz, A., Elhoseny, M., Salama, A. S., & Riad, A. M. (2018). A machine learning model for improving healthcare services on cloud computing environment. *Measurement*, *119*, 117–128. <https://doi.org/10.1016/J.MEASUREMENT.2018.01.022>
- Acharya, U. R., Fujita, H., Bhat, S., Raghavendra, U., Gudigar, A., Molinari, F., ... Hoong Ng, K. (2016). Decision support system for fatty liver disease using GIST descriptors extracted from ultrasound images. *Information Fusion*, *29*, 32–39. <https://doi.org/10.1016/J.INFFUS.2015.09.006>
- Ahmed, M. R., Arifa Khatun, M., Ali, A., & Sundaraj, K. (2018). A literature review on NoSQL database for big data processing. *International Journal of Engineering & Technology*, *7*(2), 902–906. <https://doi.org/10.14419/ijet.v7i2.12113>
- Asri, H., Mousannif, H., Moatassime, H. Al, & Noel, T. (2016). Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. *Procedia Computer Science*, *83*, 1064–1069. <https://doi.org/10.1016/J.PROCS.2016.04.224>
- Baitharu, T. R., & Pani, S. K. (2016). Analysis of Data Mining Techniques for Healthcare Decision Support System Using Liver Disorder Dataset. *Procedia Computer Science*, *85*, 862–870. <https://doi.org/10.1016/J.PROCS.2016.05.276>
- Bartz-Kurycki, M. A., Green, C., Anderson, K. T., Alder, A. C., Bucher, B. T., Cina, R. A., ... Tsao, K. (2018). Enhanced neonatal surgical site infection prediction model utilizing statistically and clinically significant variables in combination with a machine learning algorithm. *The American Journal of Surgery*, *216*(4), 764–777. <https://doi.org/10.1016/J.AMJSURG.2018.07.041>
- Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Carvalho, D., Pinheiro, P. R., & Pinheiro, M. C. D. (2016). A Hybrid Model to Support the Early Diagnosis of Breast Cancer. *Procedia Computer Science*, *91*, 927–934. <https://doi.org/10.1016/J.PROCS.2016.07.112>
- Chen, Z., Zhang, Z., Zhu, R., Xiang, Y., & Harrington, P. B. (2016). Diagnosis of patients with chronic kidney disease by using two fuzzy classifiers. *Chemometrics and Intelligent Laboratory Systems*, *153*, 140–145. <https://doi.org/10.1016/J.CHEMOLAB.2016.03.004>
- Chervonenkis, A. Y. (2013). Early History of Support Vector Machines. In *Empirical Inference* (pp. 13–20). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-41136-6_3
- Chougrad, H., Zouaki, H., & Alheyane, O. (2018). Deep Convolutional Neural Networks for breast cancer screening. *Computer Methods and Programs in Biomedicine*, *157*, 19–30. <https://doi.org/10.1016/J.CMPB.2018.01.011>

- Chronic Kidney Disease Basics | Chronic Kidney Disease Initiative | CDC. (n.d.). Retrieved December 12, 2018, from <https://www.cdc.gov/kidneydisease/basics.html>
- Confusion Matrix. (n.d.). Retrieved December 20, 2018, from http://www2.cs.uregina.ca/~hamilton/courses/831/notes/confusion_matrix/confusion_matrix.html
- Dwivedi, A. K. (2017). Analysis of computational intelligence techniques for diabetes mellitus prediction. *Neural Computing and Applications*, 1–9. <https://doi.org/10.1007/s00521-017-2969-9>
- False Discovery Rate: Simple Definition, Adjusting for FDR - Statistics How To. (n.d.). Retrieved December 9, 2018, from <https://www.statisticshowto.datasciencecentral.com/false-discovery-rate/>
- False Omission Rate - calculator - fxSolver. (n.d.). Retrieved December 9, 2018, from <https://www.fxsolver.com/browse/formulas/False+Omission+Rate>
- Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). KNN Model-Based Approach in Classification (pp. 986–996). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-39964-3_62
- Heydari, M., Teimouri, M., Heshmati, Z., & Alavinia, S. M. (2016). Comparison of various classification algorithms in the diagnosis of type 2 diabetes in Iran. *International Journal of Diabetes in Developing Countries*, 36(2), 167–173. <https://doi.org/10.1007/s13410-015-0374-4>
- Hossain, R., Mahmud, S. M. H., Hossin, M. A., Haider Noori, S. R., & Jahan, H. (2018). PRMT: Predicting Risk Factor of Obesity among Middle-Aged People Using Data Mining Techniques. *Procedia Computer Science*, 132, 1068–1076. <https://doi.org/10.1016/J.PROCS.2018.05.022>
- Jafari-Marandi, R., Davarzani, S., Soltanpour Gharibdousti, M., & Smith, B. K. (2018). An optimum ANN-based breast cancer diagnosis: Bridging gaps between ANN learning and decision-making goals. *Applied Soft Computing*, 72, 108–120. <https://doi.org/10.1016/J.ASOC.2018.07.060>
- Jain, D., & Singh, V. (2018). Feature selection and classification systems for chronic disease prediction: A review. *Egyptian Informatics Journal*, 19(3), 179–189. <https://doi.org/10.1016/J.EIJ.2018.03.002>
- Jr, D. H., Lemeshow, S., & Sturdivant, R. (2013). *Applied logistic regression*. Retrieved from https://books.google.com/books?hl=en&lr=&id=64JYAwwAAQBAJ&oi=fnd&pg=PR13&dq=applied+logistics+regression+wiley&ots=DsfO4Z4rkK&sig=qrAYpO4bqZMuaADsHhl_mjXWZFE
- Kazemi, Y., & Mirroshandel, S. A. (2018). A novel method for predicting kidney stone type using ensemble learning. *Artificial Intelligence in Medicine*, 84, 117–126. <https://doi.org/10.1016/J.ARTMED.2017.12.001>
- Kukar, M., Kononenko, I., Grošelj, C., ... K. K.-A. intelligence in, & 1999, undefined. (n.d.). Analysing and improving the diagnosis of ischaemic heart disease with machine learning.

Elsevier. Retrieved from

<https://www.sciencedirect.com/science/article/pii/S0933365798000633>

- Kumari, M. (2018). Breast Cancer Prediction system. *Procedia Computer Science*, 132, 371–376. <https://doi.org/10.1016/J.PROCS.2018.05.197>
- Leung, K. M. (2007). Naive bayesian classifier. *Polytechnic University Department of Computer Science/Finance and Risk Engineering*.
- Lorente, D., Martínez-Martínez, F., Rupérez, M. J., Lago, M. A., Martínez-Sober, M., Escandell-Montero, P., ... Martín-Guerrero, J. D. (2017). A framework for modelling the biomechanical behaviour of the human liver during breathing in real time using machine learning. *Expert Systems with Applications*, 71, 342–357. <https://doi.org/10.1016/J.ESWA.2016.11.037>
- Luyckx, V. A., Tonelli, M., & Stanifer, J. W. (2018). The global burden of kidney disease and the sustainable development goals. *Bulletin of the World Health Organization*, 96(6), 414–422D. <https://doi.org/10.2471/BLT.17.206441>
- Mahmud, S. M. H., & Ahmed, R. (2018). Machine Learning Based Unified Framework for Diabetes Prediction.
- Nagarajan, R., & Upreti, M. (2017). An ensemble predictive modeling framework for breast cancer classification. *Methods*, 131, 128–134. <https://doi.org/10.1016/J.YMETH.2017.07.011>
- Nilashi, M., Ahmadi, H., Shahmoradi, L., Ibrahim, O., & Akbari, E. (2018). A predictive method for hepatitis disease diagnosis using ensembles of neuro-fuzzy technique. *Journal of Infection and Public Health*. <https://doi.org/10.1016/J.JIPH.2018.09.009>
- Purushottam, Saxena, K., & Sharma, R. (2016). Efficient Heart Disease Prediction System. *Procedia Computer Science*, 85, 962–969. <https://doi.org/10.1016/J.PROCS.2016.05.288>
- Ramana, B., Babu, M., of, N. V.-I. J., & 2012, undefined. (n.d.). A critical comparative study of liver patients from USA and INDIA: an exploratory analysis. *Citeseer*. Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.401.9235&rep=rep1&type=pdf>
- Rau, H.-H., Hsu, C.-Y., Lin, Y.-A., Atique, S., Fuad, A., Wei, L.-M., & Hsu, M.-H. (2016). Development of a web-based liver cancer prediction model for type II diabetes patients by using an artificial neural network. *Computer Methods and Programs in Biomedicine*, 125, 58–65. <https://doi.org/10.1016/J.CMPB.2015.11.009>
- Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics*, 21(3), 660–674. <https://doi.org/10.1109/21.97458>
- Shukla, N., Hagenbuchner, M., Win, K. T., & Yang, J. (2018). Breast cancer data analysis for survivability studies and prediction. *Computer Methods and Programs in Biomedicine*, 155, 199–208. <https://doi.org/10.1016/J.CMPB.2017.12.011>
- Singh, P., Singh, S., & Pandi-Jain, G. S. (2018). Effective heart disease prediction system using data mining techniques. *International Journal of Nanomedicine*, 13, 121–124.

<https://doi.org/10.2147/IJN.S124998>

- Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14(3), 199–222. <https://doi.org/10.1023/B:STCO.0000035301.49549.88>
- Tapak, L., Shirmohammadi-Khorram, N., Amini, P., Alafchi, B., Hamidi, O., & Poorolajal, J. (2018). Prediction of survival and metastasis in breast cancer patients using machine learning classifiers. *Clinical Epidemiology and Global Health*. <https://doi.org/10.1016/J.CEGH.2018.10.003>
- UCI Machine Learning Repository: Chronic_Kidney_Disease Data Set. (n.d.). Retrieved December 8, 2018, from https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease#
- Vapnik, V., Guyon, I., Learn, T. H.-M., & 1995, undefined. (n.d.). Support vector machines. *Statweb.Stanford.Edu*. Retrieved from <http://statweb.stanford.edu/~tibs/sta306bfiles/svmtalk.pdf>
- Wang, H., Zheng, B., Yoon, S. W., & Ko, H. S. (2018). A support vector machine-based ensemble algorithm for breast cancer diagnosis. *European Journal of Operational Research*, 267(2), 687–699. <https://doi.org/10.1016/J.EJOR.2017.12.001>
- WHO | Breast cancer: prevention and control. (2016). *WHO*. Retrieved from <https://www.who.int/cancer/detection/breastcancer/en/index1.html>
- WHO | World Health Organization. (n.d.). Retrieved November 5, 2018, from <http://www.who.int/>
- Wolberg, W. H., & Mangasarian, O. L. (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proceedings of the National Academy of Sciences*, 87(23), 9193–9196. <https://doi.org/10.1073/pnas.87.23.9193>
- Xiao, Y., Wu, J., Lin, Z., & Zhao, X. (2018). A deep learning-based multi-model ensemble method for cancer prediction. *Computer Methods and Programs in Biomedicine*, 153, 1–9. <https://doi.org/10.1016/J.CMPB.2017.09.005>
- Zhang, M.-L., & Zhou, Z.-H. (2007). ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7), 2038–2048. <https://doi.org/10.1016/J.PATCOG.2006.12.019>

APPENDIX

Appendix A: Liver Patient Dataset

Sample of Indian Liver Patient Data (Ramana et al., n.d.)

Age	Gender	Total_Bilir	Direct_Bil	Alkaline_P	Alamine_P	Aspartate	Total_Proi	Albumin	Albumin_	Dataset
65	Female	0.7	0.1	187	16	18	6.8	3.3	0.9	1
62	Male	10.9	5.5	699	64	100	7.5	3.2	0.74	1
62	Male	7.3	4.1	490	60	68	7	3.3	0.89	1
58	Male	1	0.4	182	14	20	6.8	3.4	1	1
72	Male	3.9	2	195	27	59	7.3	2.4	0.4	1
46	Male	1.8	0.7	208	19	14	7.6	4.4	1.3	1
26	Female	0.9	0.2	154	16	12	7	3.5	1	1
29	Female	0.9	0.3	202	14	11	6.7	3.6	1.1	1
17	Male	0.9	0.3	202	22	19	7.4	4.1	1.2	2
55	Male	0.7	0.2	290	53	58	6.8	3.4	1	1
57	Male	0.6	0.1	210	51	59	5.9	2.7	0.8	1
72	Male	2.7	1.3	260	31	56	7.4	3	0.6	1
64	Male	0.9	0.3	310	61	58	7	3.4	0.9	2
74	Female	1.1	0.4	214	22	30	8.1	4.1	1	1
61	Male	0.7	0.2	145	53	41	5.8	2.7	0.87	1
25	Male	0.6	0.1	183	91	53	5.5	2.3	0.7	2
38	Male	1.8	0.8	342	168	441	7.6	4.4	1.3	1
33	Male	1.6	0.5	165	15	23	7.3	3.5	0.92	2
40	Female	0.9	0.3	293	232	245	6.8	3.1	0.8	1
40	Female	0.9	0.3	293	232	245	6.8	3.1	0.8	1
51	Male	2.2	1	610	17	28	7.3	2.6	0.55	1
51	Male	2.9	1.3	482	22	34	7	2.4	0.5	1
62	Male	6.8	3	542	116	66	6.4	3.1	0.9	1
40	Male	1.9	1	231	16	55	4.3	1.6	0.6	1

Appendix B: Breast Cancer Dataset

Sample of Wisconsin Breast Cancer Data (Wolberg & Mangasarian, 1990)

id	clump_thi	size_unifc	shape_un	marginal_	epithelial	bare_nucl	bland_chr	normal_n	mitoses	class
1000025	5	1	1	1	2	1	3	1	1	2
1002945	5	4	4	5	7	10	3	2	1	2
1015425	3	1	1	1	2	2	3	1	1	2
1016277	6	8	8	1	3	4	3	7	1	2
1017023	4	1	1	3	2	1	3	1	1	2
1017122	8	10	10	8	7	10	9	7	1	4
1018099	1	1	1	1	2	10	3	1	1	2
1018561	2	1	2	1	2	1	3	1	1	2
1033078	2	1	1	1	2	1	1	1	5	2
1033078	4	2	1	1	2	1	2	1	1	2
1035283	1	1	1	1	1	1	3	1	1	2
1036172	2	1	1	1	2	1	2	1	1	2
1041801	5	3	3	3	2	3	4	4	1	4
1043999	1	1	1	1	2	3	3	1	1	2
1044572	8	7	5	10	7	9	5	5	4	4
1047630	7	4	6	4	6	1	4	3	1	4
1048672	4	1	1	1	2	1	2	1	1	2
1049815	4	1	1	1	2	1	3	1	1	2
1050670	10	7	7	6	4	10	4	1	2	4
1050718	6	1	1	1	2	1	3	1	1	2
1054590	7	3	2	10	5	10	5	4	4	4
1054593	10	5	5	3	6	7	7	10	1	4
1056784	3	1	1	1	2	1	2	1	1	2
1057013	8	4	5	1	2 ?		7	3	1	4
1059552	1	1	1	1	2	1	3	1	1	2
1065726	5	2	3	4	2	7	3	6	1	4
1066373	3	2	1	1	1	1	2	1	1	2
1066979	5	1	1	1	2	1	2	1	1	2
1067444	2	1	1	1	2	1	2	1	1	2
1070935	1	1	3	1	2	1	1	1	1	2
1070935	3	1	1	1	1	1	2	1	1	2
1071760	2	1	1	1	2	1	3	1	1	2
1072179	10	7	7	3	8	5	7	4	3	4

Appendix C: Chronic Kidney Disease Dataset

Sample of Chronic Kidney Disease Data (“UCI Machine Learning Repository:

Chronic_Kidney_Disease Data Set,” n.d.): (Ramana et al., n.d.)

id	age	bp	sg	al	su	rbc	pc	pcc	ba	bgr	bu	sc	sod	pot	hemo	pcv	wc	rc	htn	dm	cad	appet	pe	ane	classificati	
0	48	80	1.02	1	0		normal	notpresen	notpresen		121	36	1.2		15.4	44	7800	5.2	yes	yes	no	good	no	no	ckd	
1	7	50	1.02	4	0		normal	notpresen	notpresen			18	0.8		11.3	38	6000		no	no	no	good	no	no	ckd	
2	62	80	1.01	2	3	normal	normal	notpresen	notpresen		423	53	1.8		9.6	31	7500		no	yes	no	poor	no	yes	ckd	
3	48	70	1.005	4	0	normal	abnormal	present	notpresen		117	56	3.8	111	2.5	11.2	32	6700	3.9	yes	no	no	poor	yes	yes	ckd
4	51	80	1.01	2	0	normal	normal	notpresen	notpresen		106	26	1.4		11.6	35	7300	4.6	no	no	no	good	no	no	ckd	
5	60	90	1.015	3	0			notpresen	notpresen		74	25	1.1	142	3.2	12.2	39	7800	4.4	yes	yes	no	good	yes	no	ckd
6	68	70	1.01	0	0		normal	notpresen	notpresen		100	54	24	104	4	12.4	36		no	no	no	good	no	no	ckd	
7	24		1.015	2	4	normal	abnormal	notpresen	notpresen		410	31	1.1		12.4	44	6900	5	no	yes	no	good	yes	no	ckd	
8	52	100	1.015	3	0	normal	abnormal	present	notpresen		138	60	1.9		10.8	33	9600	4	yes	yes	no	good	no	yes	ckd	
9	53	90	1.02	2	0	abnormal	abnormal	present	notpresen		70	107	7.2	114	3.7	9.5	29	12100	3.7	yes	yes	no	poor	no	yes	ckd
10	50	60	1.01	2	4		abnormal	present	notpresen		490	55	4		9.4	28			yes	yes	no	good	no	yes	ckd	
11	63	70	1.01	3	0	abnormal	abnormal	present	notpresen		380	60	2.7	131	4.2	10.8	32	4500	3.8	yes	yes	no	poor	yes	no	ckd
12	68	70	1.015	3	1		normal	notpresen	notpresen		208	72	2.1	138	5.8	9.7	28	12200	3.4	yes	yes	yes	poor	yes	no	ckd
13	68	70						notpresen	notpresen		98	86	4.6	135	3.4	9.8			yes	yes	yes	poor	yes	no	ckd	
14	68	80	1.01	3	2	normal	abnormal	present	present		157	90	4.1	130	6.4	5.6	16	11000	2.6	yes	yes	yes	poor	yes	no	ckd
15	40	80	1.015	3	0		normal	notpresen	notpresen		76	162	9.6	141	4.9	7.6	24	3800	2.8	yes	no	no	good	no	yes	ckd
16	47	70	1.015	2	0		normal	notpresen	notpresen		99	46	2.2	138	4.1	12.6			no	no	no	good	no	no	ckd	
17	47	80						notpresen	notpresen		114	87	5.2	139	3.7	12.1			yes	no	no	poor	no	no	ckd	
18	60	100	1.025	0	3		normal	notpresen	notpresen		263	27	1.3	135	4.3	12.7	37	11400	4.3	yes	yes	yes	good	no	no	ckd
19	62	60	1.015	1	0		abnormal	present	notpresen		100	31	1.6		10.3	30	5300	3.7	yes	no	yes	good	no	no	ckd	
20	61	80	1.015	2	0	abnormal	abnormal	notpresen	notpresen		173	148	3.9	135	5.2	7.7	24	9200	3.2	yes	yes	yes	poor	yes	yes	ckd
21	60	90						notpresen	notpresen		180	76	4.5		10.9	32	6200	3.6	yes	yes	yes	good	no	no	ckd	
22	48	80	1.025	4	0	normal	abnormal	notpresen	notpresen		95	163	7.7	136	3.8	9.8	32	6900	3.4	yes	no	no	good	no	yes	ckd
23	21	70	1.01	0	0		normal	notpresen	notpresen										no	no	no	poor	no	yes	ckd	
24	42	100	1.015	4	0	normal	abnormal	notpresen	present		50	1.4	129	4	11.1	39	8300	4.6	yes	no	no	poor	no	no	ckd	
25	61	60	1.025	0	0		normal	notpresen	notpresen		108	75	1.9	141	5.2	9.9	29	8400	3.7	yes	yes	no	good	no	yes	ckd
26	75	80	1.015	0	0		normal	notpresen	notpresen		156	45	2.4	140	3.4	11.6	35	10300	4	yes	yes	no	poor	no	no	ckd
27	69	70	1.01	3	4	normal	abnormal	notpresen	notpresen		264	87	2.7	130	4	12.5	37	9600	4.1	yes	yes	yes	good	yes	no	ckd
28	75	70		1	3			notpresen	notpresen		123	31	1.4						no	yes	no	good	no	no	ckd	
29	68	70	1.005	1	0	abnormal	abnormal	present	notpresen		28	1.4			12.9	38			no	no	yes	good	no	no	ckd	
30		70						notpresen	notpresen		93	155	7.3	132	4.9				yes	yes	no	good	no	no	ckd	
31	73	90	1.015	3	0		abnormal	present	notpresen		107	33	1.5	141	4.6	10.1	30	7800	4	no	no	no	poor	no	no	ckd
32	61	90	1.01	1	1		normal	notpresen	notpresen		159	39	1.5	133	4.9	11.3	34	9600	4	yes	yes	no	poor	no	no	ckd
33	60	100	1.02	2	0	abnormal	abnormal	notpresen	notpresen		140	55	2.5		10.1	29			yes	no	no	poor	no	no	ckd	
34	70	70	1.01	1	0	normal		present	present		171	153	5.2						no	yes	no	poor	no	no	ckd	
35	65	90	1.02	2	1	abnormal	normal	notpresen	notpresen		270	39	2		12	36	9800	4.9	yes	yes	no	poor	no	yes	ckd	
36	76	70	1.015	1	0	normal	normal	notpresen	notpresen		92	29	1.8	133	3.9	10.3	32		yes	no	no	good	no	no	ckd	
37	72	80						notpresen	notpresen		137	65	3.4	141	4.7	9.7	28	6900	2.5	yes	yes	no	poor	no	yes	ckd
38	69	80	1.02	3	0	abnormal	normal	notpresen	notpresen		103	4.1	132	5.9	12.5				yes	no	no	good	no	no	ckd	
39	82	80	1.01	2	2	normal		notpresen	notpresen		140	70	3.4	136	4.2	13	40	9800	4.2	yes	yes	no	good	no	no	ckd
40	46	90	1.01	2	0	normal	abnormal	notpresen	notpresen		99	80	2.1		11.1	32	9100	4.1	yes	no	no	good	no	no	ckd	
41	45	70	1.01	0	0		normal	notpresen	notpresen		20	0.7							no	no	no	good	yes	no	ckd	
42	47	100	1.01	0	0		normal	notpresen	notpresen		204	29	1	139	4.2	9.7	33	9200	4.5	yes	no	no	good	no	yes	ckd
43	35	80	1.01	1	0	abnormal		notpresen	notpresen		79	202	10.8	134	3.4	7.9	24	7900	3.1	no	yes	no	good	no	no	ckd
44	54	80	1.01	3	0	abnormal	abnormal	notpresen	notpresen		207	77	6.3	134	4.8	9.7	28		yes	yes	no	poor	yes	no	ckd	
45	54	80	1.02	3	0		abnormal	notpresen	notpresen		208	89	5.9	130	4.9	9.3			yes	yes	no	poor	yes	no	ckd	
46	48	70	1.015	0	0		normal	notpresen	notpresen		124	24	1.2	142	4.2	12.4	37	6400	4.7	no	yes	no	good	no	no	ckd
47	11	80	1.01	3	0		normal	notpresen	notpresen		17	0.8			15	45	8600		no	no	no	good	no	no	ckd	