CHAPTER 1

INTRODUCTION

Hepatitis is caused by the inflammation of the liver. It may occur with no symptoms which lead to yellow discoloration of the skin and enlargement of the spleen. The presence of jaundice indicates the advanced liver disease. It may also lead to weight loss. Data mining is used to identify uncovering patterns from the stored data and that can be marked as insufficiency of data. Predictive models can be built and developed by using the above information. Besides, this can be used to evaluate the effectiveness of medical treatment by comparing and contrasting the symptoms and courses of treatment. Data obtained from the tested results can be converted into useful information. A machine learning algorithm is used to find accurate and reliable results. The structure of a machine learning algorithm is evaluated in terms of accuracy, sensitivity. An algorithm starts with two major divisions, one as a training set in which the data is stored and the other one is a testing set, where accuracy is obtained. Primary liver cancer (PLC) is a common disease in our country and particularly prevalent in the southeast of China. HBV reactivation is a common complication in patients with primary liver cancer (PLC) after the precise radiotherapy. It affects the prognosis of patients and endangers the patient lives and has pointed out the HBV DNA level, tumor stage TNM, Child-Pugh classification of liver function is the risk factors of HBV reactivation after three-dimensional conformal radiation therapy (3D-CRT) in patients with primary hepatocellular carcinoma. The HBV DNA level, radiation dose is also pointed out. By using t test and logistic regression it is discovered that in tumor stage TNM, outer margin of radiotherapy and HBV DNA level are the risk factors of HBV reactivation.

1.1 BACKGROUND

The healthcare industry generates a tremendous amount of data but struggles to convert that data into useful insights to improve patient outcomes. Data analytics in healthcare is intended to be applied to every aspect of patient care and operations management. It is used to investigate methods of improving patient care, predicting disease outbreaks, reducing the cost of treatment, and so much more. With technological advancement, the impact that analytics can have in the healthcare industry is tremendous. AI and machine learning techniques can use data to diagnose disease, determine the best treatment for each patients use case, and much more.

Predictive analytics in healthcare can predict which patients are at a higher risk and start early innervations so deeper problems can be avoided. For example, it can identify patients with cardiovascular disease with the highest probability of hospitalization based on age coexisting chronic illnesses, and medication adherence. Predictions on the likelihood of disease and chronic illness can help doctors and healthcare organizations proactively provide care rather than waiting for at-risk patients to come in for a regular check-up. Besides the chronically ill patients, there are other at-risk groups, including elderly people and patients that have been recently discharged from the hospital after invasive manipulations.

Hepatitis B is a vaccine-preventable liver infection caused by the hepatitis B virus (HBV). Hepatitis B is spread when blood, semen, or other body fluids from a person infected with the virus enters the body of someone who is not infected. This can happen through sexual contact; sharing needles, syringes, or other drug-injection equipment; or from mother to baby at birth. Not all people newly infected with HBV have symptoms, but for those that do, symptoms can include fatigue, poor appetite, stomach pain, nausea, and jaundice. For many people, hepatitis B is a short-term illness. For others, it can become a long-term, chronic infection that can lead to serious, even life-threatening health issues like cirrhosis or liver cancer. Risk for chronic infection is related to age at infection: about 90% of infants with hepatitis B go on to develop chronic infection, whereas only 2%–6% of people who get hepatitis B as adults become chronically infected. The best way to prevent hepatitis B is to get vaccinated.

The next step is to build a predictive analytics system that can help predict if a person has the hepatitis B virus and needs instant treatment or if a person is safe. The first step of predictive analysis is data gathering and cleaning. In this project the dataset in collected and gathered from UC Irvine Machine Learning Repository. This dataset has 20 attributes and 155 instances. For the data cleaning we follow multiple steps like removing irrelevant data, removing duplicate data, dealing with missing data & filter out data outliers.

The next step is data analysis, it is to be able to understand how the data is behaving and the relationships between variables. If you can't do that, you won't be able to build a good model. If you can, however, you'll learn a lot. By creating a simple chart of your data to study it, you may get a good idea of the answer to the problem you are trying to solve based on the overall trend. The next is building a predictive model, Sometimes the data lends itself to a specific algorithm or model. Other times the best approach is not so clear-cut. As you analyse the data, run as many algorithms as you can and compare their outputs. Identify test data and apply classification rules to check the efficiency of the classification model against test data.

1.2. LITERATURE SURVEY

PAPER	REVIEW	ADVANTAGES	DISADVANTAGES
"Hepatitis B Diagnosis Using Logical Inference and Generalized Regression Neural Networks" Ghumbre Shashikant Uttreshwar, 2009	Artificial Neural Network The proposed system used generalized regression neural network	Efficient with a shorter training dataset	Scalability for high dimensionality data.
"Artificial Neural Networks for Diagnosis of Hepatitis Disease" Lale Ozyilmaz & Tulay Yildirim," Nov 2009	Multilayer Perceptron Radial Basis Function Conic Section Function Neural Network	Best classification accuracy Efficient in statistical methods	Accuracy decreases with increase in number of classes.
"Application of CART algorithm in hepatitis disease diagnosis" G.Sathyadevi, June 2011	C4.5 algorithm & ID3 algorithm CART decision tree algorithm. CART outperforms C4.5 & ID3	CART performs with better accuracy and time complexity. Transparent and easy to understand.	High variance & Unstable Over-fitting
"Diagnosis of Hepatitis using Decision tree algorithm" V.Shankar Sowmien, June 2016	Real Time Application C4.5 algorithm preferred over C5.0, J48, Fuzzy Rule	Attributes are less Complexity of decision tree is reduced	Does not work well with small training set. Small variation in data can lead to different decision trees.
"Random forest and Bayesian prediction for Hepatitis B virus reactivation" Huina Wang, Yihui Liu, 2017	The proposed system used Random Forest under 200 decision trees and Bayesian prediction using 10-fold cross validation.	The classification accuracy of random forest can be reached to 85.15%. The accuracy of Bayesian classifier reached to 84.57%	Can take a long time to train with a large number of trees. They're not easily interpretable.

"Rapid Detection System for Hepatitis B Surface Antigen (HBsAg) Based on Immunomagnetic Separation, Multi-Angle Dynamic Light Scattering and Support Vector Machine" MUBASHIR HUSSAIN, 2020	The proposed HBsAg detection method can differentiate the sample that contains HBsAg enriched IM beads and blank IM beads	The rapid detection of HBsAg using immunomagnetic separation Dynamic light scattering Support vector machine	The rapid identification of HBsAg is challenging in a resource-limited setting
"The research of SARIMA model for prediction of hepatitis B in mainland China" Daren Zhao, Huiwu Zhang June 2022	Constructed a SARIMA model for prediction, and provided corresponding preventive measures.	The SARIMA model includes seasonal characteristics, describes the goodness-of-fit between the predicted and observed values.	It can only extract linear relationships within the time series data.
"Prediction Model of HBV Reactivation in Primary Liver Cancer - Based on NCA Feature Selection and SVM Classifier with Bayesian and Grid Optimization" Yongwang Zhao, Yihui Liu 2018	Bayes and Grid optimization are respectively used to optimize the previous SVM model	Under 10- fold cross validation the prediction accuracy reached to 85.56%.	We can miss good hyperparameter values not set in the beginning.
"Predicting Life Expectancy of Hepatitis B Patients using Machine Learning" Nabeel Ali April 2022	Area Under Curve analysis was used to assess the estimation of various models.	Multiple machine learning models and algorithms testing performed.	All proposed models have similar areas under the ROC curve.

Table 1.1: Literature Survey

From the study, there are different models which play an important role in medical diagnostic field and were discovered to identify the hepatitis disease by using different architectures.

1. Lale ozyilmazet al, proposed the Multilayer Perceptron, Radial Basis Function and Conic Section Function Neural Network are the three neural network algorithms used to detect hepatitis disease in this paper. Results show that Multilayer Perceptron does not have less classification accuracy however Radial Basis Function gives good results. Conic Section Function Neural Network which combines both Multilayer Perceptron and Radial Basis Function are more useful for detection.
2. Ghumbre et al, the application of artificial intelligence for Hepatitis B diagnosis has been introduced. The expert system uses a logical interface along with neural network architecture is used to identify hepatitis. The detection of hepatitis B is done by using different data samples from different patients and the results have shown that artificial neural networks are equivalently good as the logical methods in the diagnosis of hepatitis B.
3. G.Sathyadevi etal, The use of decision tree C4.5 algorithm, ID3 algorithm and CART algorithm are proposed to classify the diseases and compare the effectiveness, correction rate among them . A CART decision tree algorithm is proposed where accuracy and time complexity are required CART algorithm performs better than the other two algorithms.
4. Changjiang Long, Huan et al, A mathematical model is introduced by including HBV, hepatocytes and the immune system. In this study, interaction of virus and immune system is described. This also discusses on the other side of hepatitis B virus, which can suppress the immune system by increasing the death rate of CTLs. So, it is impossible for the immune system to remove the virus, and persistent infection takes place to cure chronic Hepatitis B, HBV viral load reduced firstly and develop immunity to the virus.
5. Na Chu, Lizhuang Ma et al, in this study, it has been concluded that the PETs like NBTree, CITree and CLLTree have an inherent barrier to achieve both high time complexity and equal weight attribute input, this issue can be overcome by combining with an attribute selection method. The model provides good performance in both classification and ranking. It is concluded that WPETs model fits for liver cirrhosis and hepatitis TCM diagnosis.

6. tal, A hybrid intelligent syndrome diagnosis (HISD) model is proposed which helps to compare the traditional Chinese medicine (TCM). Instead of a single technique, the proposed method uses multiple methods, the model is more objective. The method increases the accuracy and helps better for clinical purposes.
7. CheLijuan, Zhou Qiang et al, Hepatitis B has become a serious health problem which causes liver infection due to hepatitis B virus. The detection of disease for a new patient can be performed on the basis of primary phase.
8. C.Mahesh, K.Kiruthika, M.Dhilsathfathima et al, The details from the four diagnostic methods and information of patients are gathered, which are confirmed with Hepatitis B within a certain period of time span from three hospitals. A clinical data base is constructed, which can be used as the dataset for machine learning. Then an association rule mining algorithm is applied to analyses the data automatically. Testing the obtained data with the test data, the results shows that the accuracy rate is increased to 83.5%.

In these literature surveys they used most of the techniques for the detection of the disease for getting accurate results in the diagnosis they not only use the algorithms but also some mathematical and logistic methods for the detection of the disease by observing these literature surveys we are developing the project based on the machine learning concept those faced some complexity problems and the combined some attributes get good and accurate results.

Although the training is faster in RBF network but classification is slow in due to fact that every node in hidden layer have to compute the RBF function for the input sample vector during classification. While training the model sometimes Neural Network tends to over fit the model when there is a smaller dataset When a model gets trained with so much data, it starts learning from the noise and inaccurate data entries in our data set. Then the model does not categorize the data correctly, because of too many details and noise.
Multilayer Perceptron include too many parameters because it is fully connected.

Parameter number = width x depth x height.

Each node is connected to another in a very dense web resulting in redundancy and inefficiency.

1.3. MOTIVATION

The prediction of hepatitis virus is a significant and tedious task in medicine. The healthcare environment is generally perceived as being 'information rich' yet 'knowledge poor'. There is a wealth of data available within the healthcare system. However, there is a lack of effective analysis tools to discover hidden relationships and trends in data. Some studies have found that about half of patients still cannot accurately comprehend a diagnosis when leaving the doctor's office, making online patient education sites a major source of information for many patients. It can be seen that more and more people are relying on the Internet for health counselling and services. They use the Internet not only to look for health-related information, but also to read more professional medical literature or relevant materials, paying the most attention to measures that can improve life quality. Early and effective prediction of the incidence trend of Hepatitis B can provide a scientific basis for the prevention and treatment of Hepatitis B, as well as for the rational allocation of health resources, thereby reducing unnecessary waste.

1.4. PROBLEM STATEMENT

HBV reactivation is a common complication in patients with primary liver cancer (PLC) after the precise radiotherapy. It affects the life of patients.

The patients are feared because of the process of reactivation during the diagnosis stage, so without medical therapy how the detection of the disease in a patient can be done. So, we use ML algorithms for the prediction of the disease. We use 19 attributes for the prediction of the disease such as bilirubin, spiders, fatigue etc...

1.5. AIM AND OBJECTIVE

The objectives of this proposed project include:

- Exploring and collecting relevant dataset required for the problem.
- Pre-processing and cleaning the dataset for further training processes.
- Choosing a relevant algorithm and training the data.
- Presenting the output in a meaningful form as per requirements.

1.6. SCOPE

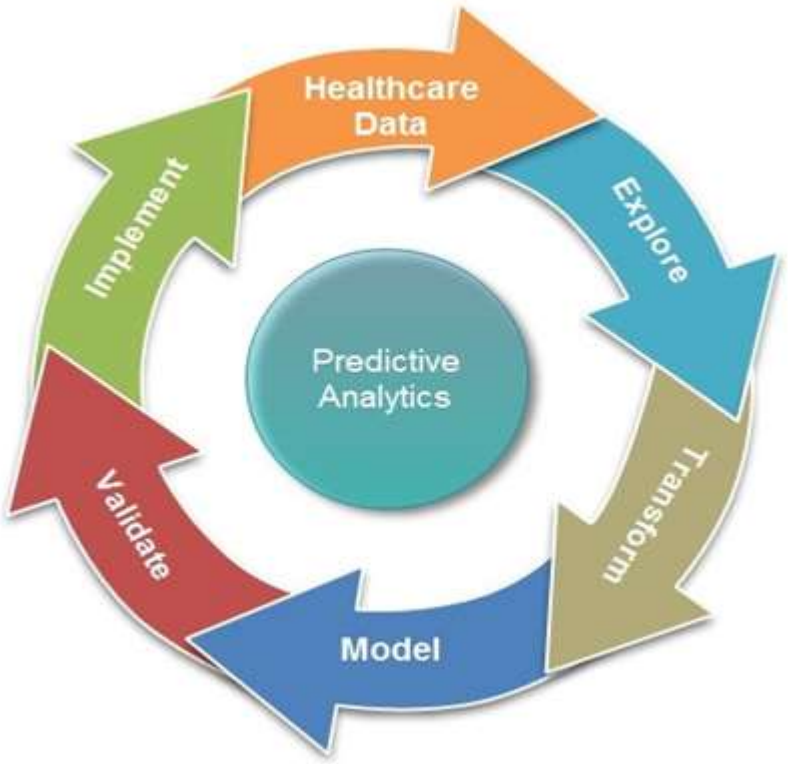

Figure 1.1: Predictive Modelling Scope

Predictive analytics is an advancing method of improving patient outcomes. By looking at data and outcomes of past patients, machine learning algorithms can be programmed to provide insight into methods of treatment that will work best for the current patients. Predictive analytics can be used to identify warning signs before conditions become severe. With the hepatitis disease being at the forefront of healthcare, researchers are putting resources into developing predictive analytic methods for combating the cause. Predictive analytics lightens the load for healthcare works by assisting in the diagnosis process.

It is a discipline that utilizes various techniques including modelling, data mining, and statistics, as well as artificial intelligence (AI) (such as machine learning) to evaluate historical and real-time data and make predictions about the future. These predictions offer a unique opportunity to see into the future and identify future trends in patient care both at an individual level and at a cohort scale.

Some of the key milestones include the digitization of health records, access to big data and storage in the cloud, advanced software, and mobile applications technology. All of these milestones have presented various advantages in the healthcare sector, including an ease of workflow, faster access to information, lower health care costs, improved public health, and the overall improvement of quality of life.

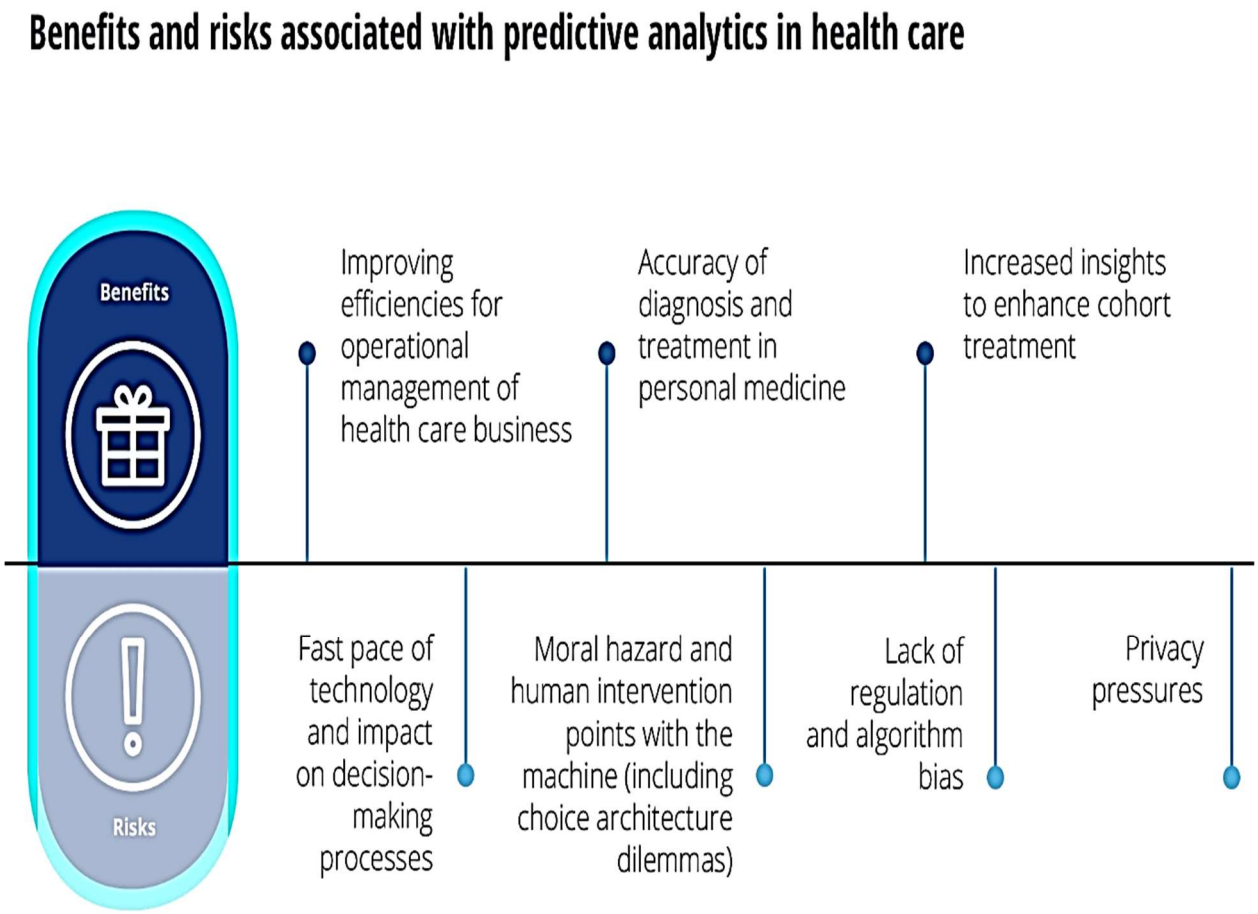

Figure 1.2: Benefits and risks in Health Care

These include operational management such as the overall improvement of business operations; personal medicine to assist and enhance accuracy of diagnosis and treatment; and cohort treatment and epidemiology to assess potential risk factors for public health. Predictive analytics in the health care sector also allows for a more definitive diagnosis of patients, followed by the appropriate treatment of the identified ailment(s). For hospitals this can mean a significant optimization in operations and a reduction in readmissions. Predictive tools such as remote patient monitoring and machine learning can work hand in hand to support decisions made in hospitals through risk scoring as well as threshold alerts.

Predictive analysis on the hepatitis B disease is performed to help the users predict if they have the virus based on the data, they provide which they get from testing. Now a days people are afraid to visit hospitals and clinics for checkups, but it is important to check if a person has a disease if they show symptom. Since Hepatitis B is a common disease which is caused by the inflammation of the liver. It may occur with no symptoms which lead to yellow discoloration of the skin and enlargement of the spleen. The presence of jaundice indicates the advanced liver disease. we built a predictive analytics system that can help predict if a person has the hepatitis B virus and needs instant treatment or if a person is safe.

The first step of predictive analysis is data gathering and cleaning. In this project the dataset in collected and gathered from UC Irvine Machine Learning Repository. This dataset has 20 attributes and 155 instances. For the data cleaning we follow multiple steps like removing irrelevant data, removing duplicate data, dealing with missing data & filter out data outliers. The next step is data analysis, it is to be able to understand how the data is behaving and the relationships between variables. The next is building a predictive model, Sometimes the data lends itself to a specific algorithm or model. As you analyze the data, run as many algorithms as you can and compare their outputs. Identify test data and apply classification rules to check the efficiency of the classification model against test data.

1.7. CHALLENGES

- **Adoption of Technology**

The more difficult a new technology is to use; the less likely end users are to adopt it—and predictive analytics solutions are notoriously difficult in meeting this challenge. This is because they typically live as standalone tools, which means users have to switch from their primary business application over to the predictive analytics solution in order to use it. What's more, traditional predictive tools are hard to scale and deploy, which makes updating them a painful process.

- **Empowering End Users**

No information is valuable in a vacuum. And that's one of the reasons predictive analytics has fallen short in empowering end users. The problem is that predictive analytics tools deliver information and insights, but they fail to let users take action. As we discussed above, if users want to act on the data, they have to jump to yet another application—ultimately wasting time and interrupting their workflow.

- **Business Understanding, to make sense of data**

The process ends with electronic closing, which saves borrowers the difficulty of having to meet a closing agent in-person, enables them to review the closing documentation at their own time and discuss any concerns. This zeroes down the chances of delays caused by a last-minute glitch in the documentation.

- **Model Overfitting or Underfitting**

This situation arises when a given model is performing well on the training data, but the performance dwindles significantly over the test set - known as overfitting model. On the other hand, if the model is functioning poorly over the test and the train set, then it is an underfitting model.

- **Model Evaluation**

Discordance between the data used and the model built. Model Evaluation is an indispensable part of the model development process. It accommodates to find the most suitable model that describes our data and how considerably the chosen model will work in the future.

- **Interoperability**

One of the biggest challenges facing the healthcare industry is interoperability. As part of federal healthcare mandates, your organization must be able to keep patient data secure while also sharing it quickly with care teams. And not just your internal teams.

- **Big Data**

Doctors expect it interoperability to include patient health data from wearable devices and consumer apps. That's a lot of data flowing into your system, adding to the already massive amount of information you're currently dealing with. It's a challenge not only for accessing and storing all that patient information, but also in converting it into meaningful insights.

- **Data Security**

One of the most controlled and, at the same time, attacked sector, is the healthcare industry. That's why more and more regulations are initiated to control access to personal data and prevent any violations of patient confidentiality or other breaches. But the process should also start from its core which is healthcare institutions. From communication between staff members to a standard for exchanging information between medical applications – it's all in there.

- **Real Quality of your Data, and is it adequate**

Understanding the real worth of data is quintessential. Too many or fewer data sets; this is one problem that every data scientist faces. The initial hurdle may be how do we use, extract, clean or interpret the data, to acquire significant insights and build models from them. Dig deep into your data. Always use curated and clean data.

CHAPTER 2

OVERVIEW

Predictive analysis on the hepatitis B disease is performed to help the users predict if they have the virus based on the data, they provide which they get from testing. Now a days people are afraid to visit hospitals and clinics for checkups, but it is important to check if a person has a disease if they show symptom. Hepatitis is a deadly disease which is caused by the inflammation of the liver.
It may occur with no symptoms which lead to yellow discoloration of the skin and enlargement of the spleen. The presence of jaundice indicates the advanced liver disease.

Predictive models can be built and developed by using the above information. Besides, this can be used to evaluate the effectiveness of medical treatment by comparing and contrasting the symptoms and courses of treatment. Data obtained from the tested results can be converted into useful information. A machine learning algorithm is used to find accurate and reliable results. The structure of a machine learning algorithm is evaluated in terms of accuracy, sensitivity.

Predictive analytics in healthcare can predict which patients are at a higher risk and start early innervations so deeper problems can be avoided. We built a predictive analytics system that can help predict if a person has the hepatitis B virus and needs instant treatment or if a person is safe.
The first step of predictive analysis is data gathering and cleaning. In this project the dataset in collected and gathered from UC Irvine Machine Learning Repository. This dataset has 20 attributes and 155 instances. For the data cleaning we follow multiple steps like removing irrelevant data, removing duplicate data, dealing with missing data & filter out data outliers.

The next step is data analysis, it is to be able to understand how the data is behaving and the relationships between variables. The next is building a predictive model, Sometimes the data lends itself to a specific algorithm or model. As you analyze the data, run as many algorithms as you can and compare their outputs. Identify test data and apply classification rules to check the efficiency of the classification model against test data.

CHAPTER 3

REQUIREMENT SPECIFICATION.

HARDWARE REQUIREMENTS

- Processor: I3 or above
- Main memory: 4 GB RAM (Min.)
- Hard Disk: Built-in is sufficient
- Keyboard: QWERTY Keyboard

SOFTWARE REQUIREMENTS

- Programming language – Python 3
- Operating system – Windows 8 or higher
- IDE – PyCharm, Spyder, VSCode

CHAPTER 4

DETAILED DESIGN

4.1 The overall process of project.

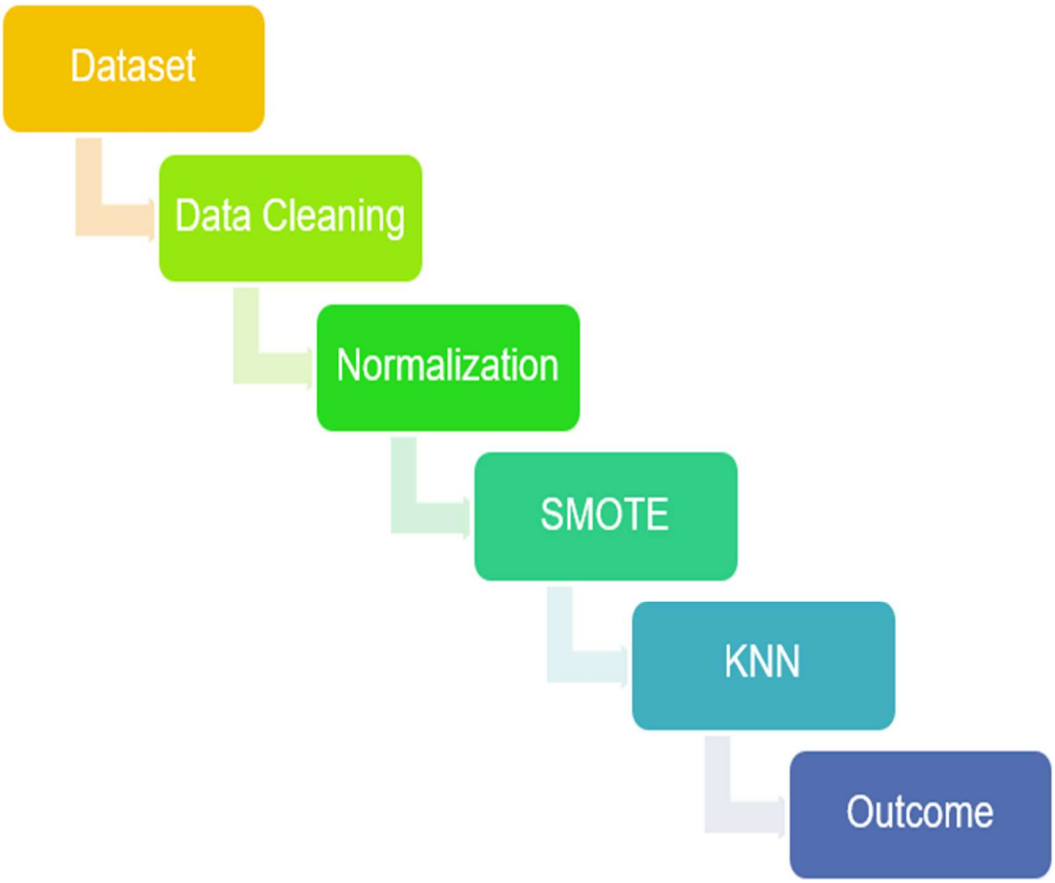

Figure 4.1: System flow diagram.

- **Data Cleaning** - It is the process of identifying the incorrect, incomplete, inaccurate, irrelevant or missing part of the data and then modifying, replacing or deleting them according to the necessity. Data cleaning is considered a foundational element of the basic data science.

- **Normalization** - Normalization is a scaling technique in Machine Learning applied during data preparation to change the values of numeric columns in the dataset to use a common scale. It is not necessary for all datasets in a model. It is required only when features of machine learning models have different ranges.
- **SMOTE** - Synthetic Minority Oversampling Technique is a statistical technique for increasing the number of cases in your dataset in a balanced way. The component works by generating new instances from existing minority cases that you supply as input.
- **KNN Processing** - "K-Nearest Neighbour". It is a supervised machine learning algorithm. The algorithm can be used to solve both classification and regression problem statements. The number of nearest neighbours to a new unknown variable that has to be predicted or classified is denoted by the symbol 'K'.
- **Trained Weights** - It is a relatively simple technique to predict the class of an item based on two or more numeric predictor variables. In weighted KNN, the nearest k points are given a weight using a function called as the kernel function. The intuition behind weighted KNN, is to give more weight to the points which are nearby and less weight to the points which are farther away.

4.2 Architectural design DFD Level 0

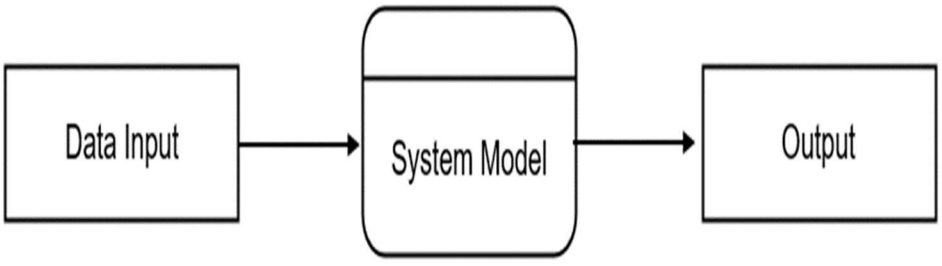

Figure 4.2: Data Flow Diagram Level 0

DFD Level 0 is also called a Context Diagram. It's a basic overview of the whole system or process being analyzed or modeled. It's designed to be an at-a-glance view, showing the system as a single high-level process, with its relationship to external entities.

Data Flow Diagram (DFD) of a system represents how input data is converted to output data graphically. Level 0 also called context level represents most fundamental and abstract view of the system. Subsequently other lower levels can be decomposed from it.

In this project the DFD level 0 represents the process taking place in the project where data is taken from the user and processed in the system model which contains the entire process of prediction analysis and provides the appropriate output.

4.3 Architectural design DFD Level 1

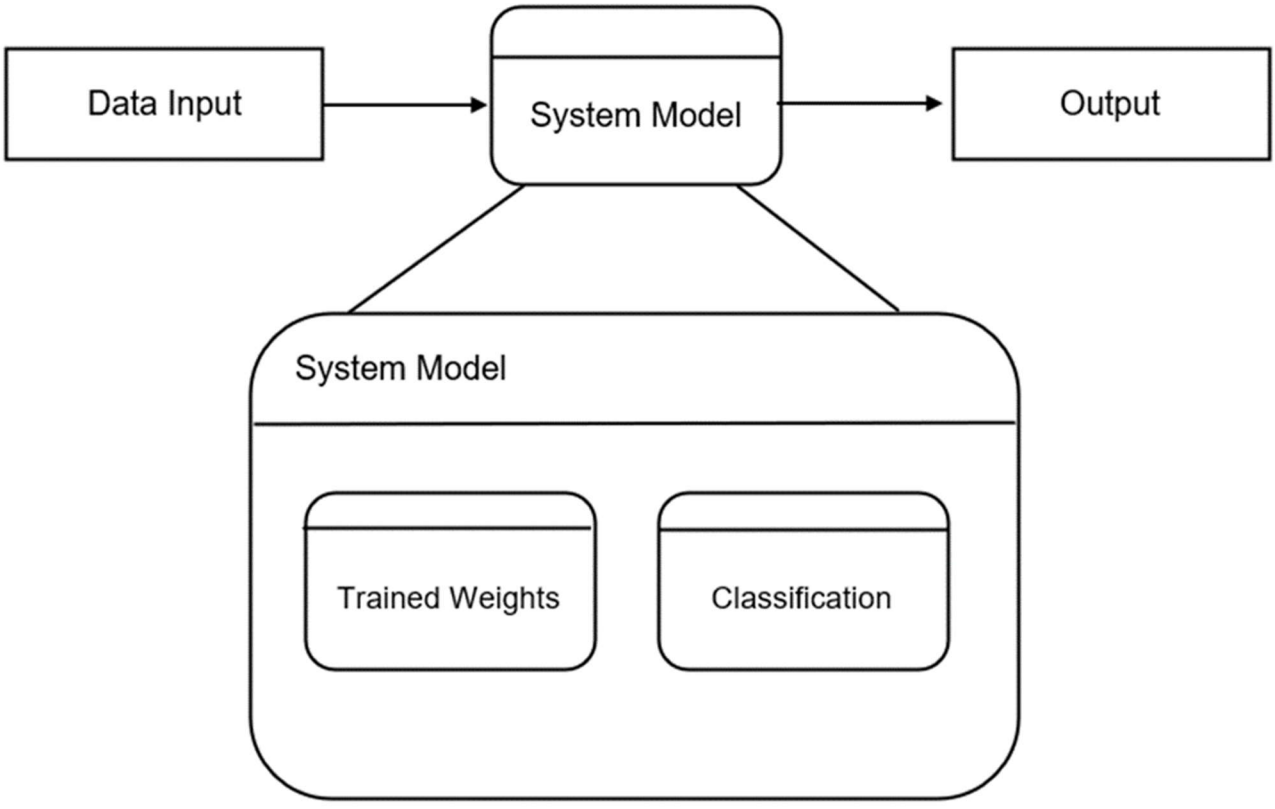

Figure 4.3: Data Flow Diagram Level 1

DFD Level 1 provides a more detailed breakout of pieces of the Context Level Diagram. We will highlight the main functions carried out by the system, as you break down the high-level process of the Context Diagram into its subprocesses. A level 1 DFD notates each of the main sub-processes that together form the complete system.

The system model of this project contains trained weights and classification. Trained Weights is a relatively simple technique to predict the class of an item based on two or more numeric predictor variables. In weighted KNN, the nearest k points are given a weight using a function called as the kernel function. The intuition behind weighted KNN, is to give more weight to the points which are nearby and less weight to the points which are farther away.

Classification is used the compare the new data provided by user and classify it with the trained data.

4.4 Architectural design DFD Level 2

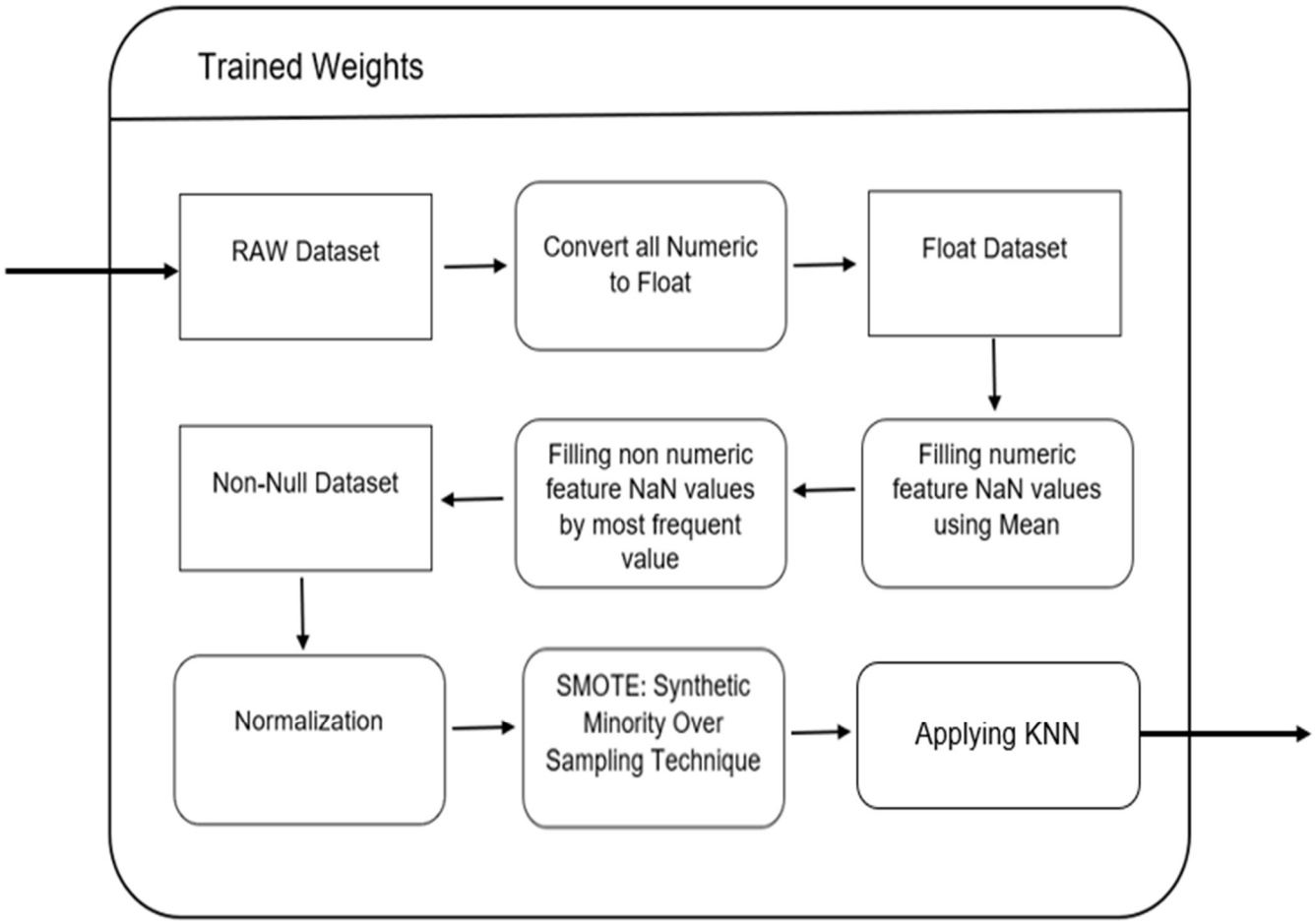

Figure 4.4: Data Flow Diagram Level 2

DFD Level 2 then goes one step deeper into parts of Level 1. It may require more text to reach the necessary level of detail about the system's functioning. It can be used to plan or record the specific/necessary detail about the system's functioning.

In this project the DFD level 2 diagrams shows us the internals process of the trained weights. The trained weights consist of the process of converting raw data into trained data which contains the steps of converting numeric values to float, filling null values using mean and most frequent value, followed by normalization using standard scalar, then followed by applying SMOTE to the dataset and completed by applying KNN to the data to make it trained.

4.5 Behavioral Design

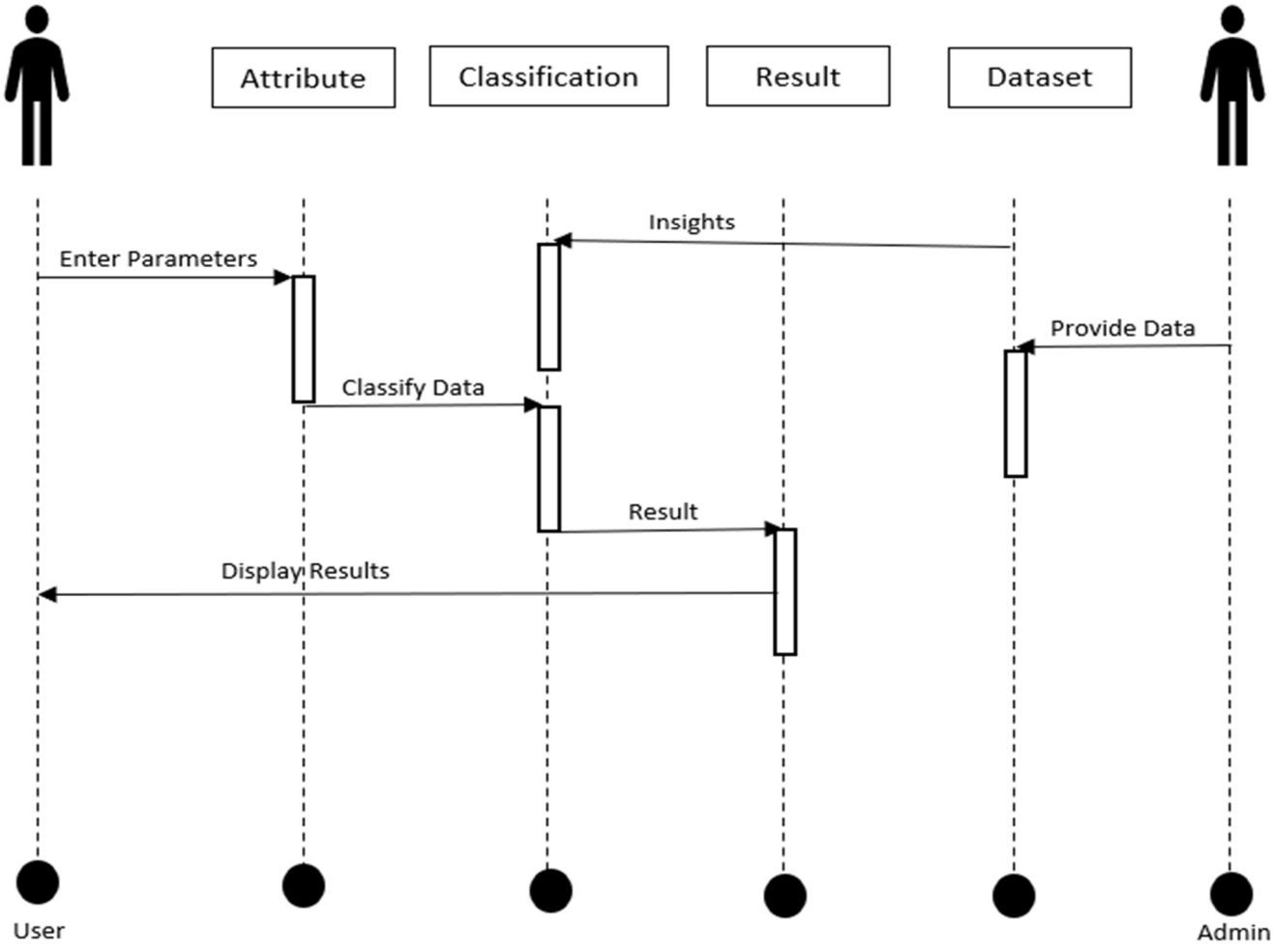

Figure 4.5: Behavioral design using sequence diagram.

Behavioural design patterns are design patterns that identify common communication patterns among objects. By doing so, these patterns increase flexibility in carrying out communication. A sequence diagram is a Unified Modelling Language (UML) diagram that illustrates the sequence of messages between objects in an interaction. A sequence diagram consists of a group of objects that are represented by lifelines, and the messages that they exchange over time during the interaction.

A sequence diagram or system sequence diagram (SSD) shows process interactions arranged in time sequence in the field of software engineering. It depicts the processes involved and the sequence of messages exchanged between the processes needed to carry out the functionality. Sequence diagrams are typically associated with use case realizations in the 4+1 architectural view model of the system under development. Sequence diagrams are sometimes called event diagrams or event scenarios.

The above sequence diagram contains 4 stages i.e., Attribute, Classification, Result, and dataset. The user and admin control the behaviour of the sequence diagram, the admin provides the system with the trained dataset with the model. This is useful the provide insights to the system. The user provides the new test data parameters which is classified to the model. The classification uses the insights of the trained data and provides the accurate result to the system which is then shared back to the user.

CHAPTER 5

IMPLEMENTATION

The Hepatitis prediction system can be implemented by creating a predictive analysis system for the prediction of the presence of the virus. The prediction can be performed by training a model using train and test dataset. In our project the dataset used has 20 attributes and 155 instances. This is the initial stage of the implementation. After we gather the data, this data needs to be analyzed and cleaned before being added to the model.

In the analysis and cleaning process some of the processes that needs to be performed on the data is to remove unwanted data, fill in missing data, analyze the pattern in data. This phase is also called EDA. Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

After the data is cleaned the data is split into training data and testing data, usually split in the ratio of 7:3. Training data is used to teach the model to learn the patterns, whereas the test data is used to verify the accuracy of the data. To get the best results we have trained the data across multiple models to test which algorithm can provide the most accurate values.

A machine learning model is defined as a mathematical representation of the output of the training process. Machine learning is the study of different algorithms that can improve automatically through experience & old data and build the model. A machine learning model is similar to computer software designed to recognize patterns or behaviours based on previous experience or data. The learning algorithm discovers patterns within the training data, and it outputs an ML model which captures these patterns and makes predictions on new data.

Based on different business goals and data sets, there are three learning models for algorithms. Each machine learning algorithm settles into one of the three models:

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

Supervised Learning is further divided into two categories:

- Classification
- Regression

Unsupervised Learning is also divided into below categories:

- Clustering
- Association Rule
- Dimensionality Reduction

In this system we will be using Supervised Learning. So, we tried building multiple Supervised Learning Algorithm. We learnt that regression does not work well with our dataset so we started building classification models. Classification models works well with small datasets. Classification algorithm is a Supervised Learning technique that is used to identify the category of new observations on the basis of training data. In Classification, a program learns from the given dataset or observations and then classifies new observation into a number of classes or groups.

The 3 classification problems which we will work with are

- Naïve Bayes
- SVM
- KNN

The algorithm which provides best accuracy will be used to build the model and predict whether a person has Hepatitis B or not. A front-end system will be provided that the user can interact with so that they can easily enter the parameters for the prediction.

Naïve Bayes

Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is mainly used in text classification that includes a high-dimensional training dataset. Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object. Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles. Bayes' theorem is also known as Bayes' Rule or Bayes' law, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability.

The formula for Bayes' theorem is given as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where,

P(A|B) is Posterior probability: Probability of hypothesis A on the observed event B.

P(B|A) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.

P(A) is Prior Probability: Probability of hypothesis before observing the evidence.

P(B) is Marginal Probability: Probability of Evidence.

Advantages of Naïve Bayes Classifier:

- Naïve Bayes is one of the fast and easy ML algorithms to predict a class of datasets.
- It can be used for Binary as well as Multi-class Classifications.
- It performs well in multi-class predictions as compared to the other Algorithms.
- It is the most popular choice for text classification problems.

Disadvantages of Naïve Bayes Classifier:

- Naive Bayes assumes that all features are independent or unrelated, so it cannot learn the relationship between features.

Applications of Naïve Bayes Classifier:

- It is used for Credit Scoring.
- It is used in medical data classification.
- It can be used in real-time predictions because Naïve Bayes Classifier is an eager learner.
- It is used in Text classification such as Spam filtering and Sentiment analysis.

SVM

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

Hyperplane: There can be multiple lines/decision boundaries to segregate the classes in n-dimensional space, but we need to find out the best decision boundary that helps to classify the data points. This best boundary is known as the hyperplane of SVM. The dimensions of the hyperplane depend on the features present in the dataset, which means if there are 2 features (as shown in image), then hyperplane will be a straight line. And if there are 3 features, then hyperplane will be a 2-dimension plane. We always create a hyperplane that has a maximum margin, which means the maximum distance between the data points.

Support Vectors: The data points or vectors that are the closest to the hyperplane and which affect the position of the hyperplane are termed as Support Vector. Since these vectors support the hyperplane, hence called a Support vector.

The SVM algorithm helps to find the best line or decision boundary; this best boundary or region is called as a hyperplane. SVM algorithm finds the closest point of the lines from both the classes. These points are called support vectors. The distance between the vectors and the hyperplane is called as margin. And the goal of SVM is to maximize this margin. The hyperplane with maximum margin is called the optimal hyperplane.

Advantages of SVM:

- SVM works relatively well when there is a clear margin of separation between classes.
- SVM is more effective in high dimensional spaces.
- SVM is effective in cases where the number of dimensions is greater than the number of samples.
- SVM is relatively memory efficient

Disadvantages of SVM:

- SVM algorithm is not suitable for large data sets.
- SVM does not perform very well when the data set has more noise i.e. target classes are overlapping.
- In cases where the number of features for each data point exceeds the number of training data samples, the SVM will underperform.
- As the support vector classifier works by putting data points, above and below the classifying hyperplane there is no probabilistic explanation for the classification.

Applications of SVM in Real World

- Face detection – SVMc classify parts of the image as a face and non-face and create a square boundary around the face.
- Text and hypertext categorization – SVMs allow Text and hypertext categorization for both inductive and transudative models. They use training data to classify documents into different categories. It categorizes on the basis of the score generated and then compares with the threshold value.
- Classification of images – Use of SVMs provides better search accuracy for image classification. It provides better accuracy in comparison to the traditional query-based searching techniques.

- Bioinformatics – It includes protein classification and cancer classification. We use SVM for identifying the classification of genes, patients on the basis of genes and other biological problems.
- Protein fold and remote homology detection – Apply SVM algorithms for protein remote homology detection.
- Handwriting recognition – We use SVMs to recognize handwritten characters used widely.
- Generalized predictive control (GPC) – Use SVM based GPC to control chaotic dynamics with useful parameters.

KNN

K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

The K-NN working can be explained on the basis of the below algorithm:

Step-1: Select the number K of the neighbors

Step-2: Calculate the Euclidean distance of K number of neighbors

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Step-3: Take the K nearest neighbors as per the calculated Euclidean distance.

Step-4: Among these k neighbors, count the number of the data points in each category.

Step-5: Assign the new data points to that category for which the number of the neighbor is maximum.

Step-6: Our model is ready.

Advantages of KNN Algorithm:

- It is simple to implement.
- It is robust to the noisy training data
- It can be more effective if the training data is large.

Disadvantages of KNN Algorithm:

- Always needs to determine the value of K which may be complex some time.
- The computation cost is high because of calculating the distance between the data points for all the training samples.

Applications of KNN

- Banking System - KNN can be used in banking system to predict weather an individual is fit for loan approval.
- Calculating Credit Ratings - KNN algorithms can be used to find an individual's credit rating by comparing with the persons having similar traits.
- Politics - With the help of KNN algorithms, we can classify a potential voter into various classes like "Will Vote", "Will not Vote", "Will Vote to Party 'Congress', "Will Vote to Party 'BJP'.
- Speech Recognition,
- Handwriting Detection
- Image Recognition
- Video Recognition

5.1 Technologies Used

- **Python**

Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built-in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together. Python's simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms, and can be freely distributed.

- **Flask**

Flask is a small and lightweight Python web framework that provides useful tools and features that make creating web applications in Python easier. It gives developers flexibility and is a more accessible framework for new developers since you can build a web application quickly using only a single Python file. Flask is also extensible and doesn’t force a particular directory structure or require complicated boilerplate code before getting started. It provided us with libraries, modules and tools to help build Web-Applications such as a blog or wiki. Flask, unlike Django does not depend on other libraries and is hence termed as a micro framework. But this also means that the more complex your web app you wish to develop is, the more dependencies you will have to install by yourself.

Syntax:

```
import flask
app = flask.Flask(_name_)              #-- Creates an application

@app.route('/')
def run_app():
   return flask.render_template('base.html')
if _name_ == '_main_':
   app.debug=True
   app.run()
#-- base.html is a basic HTML file
```

IMPLEMENTATION OF FLASK IN PROJECT

```
from flask import Flask,request, url_for, redirect, render_template
import pickle
import numpy as np

app = Flask(__name__)

model=pickle.load(open('model.pkl','rb'))

@app.route('/')
def hello_world():
    return render_template('hep_fe.html')

@app.route('/predict',methods=['POST','GET'])
def predict():
    int_features=[x for x in request.form.values()]
    final=[np.array(int_features)]
    final[0][0] = (final[0][0].astype(float)-4.120000e+01)/12.565878
    .. ..
    final = final[0].astype(np.float)
    #final = final.reshape(1,17)
    prediction=model.predict([final])
    print(prediction)
    #output='{0:.{1}f}'.format(prediction[0][1], 2)

    if prediction[0] == 2:
        return render_template('hep_fe.html',pred='Chances of Hepatitis')
    else:
        return render_template('forest_fire.html',pred='No chances of
Hepatitis')


if __name__ == '__main__':
    app.run(debug=True)
```

In the above code Flask is imported and the app is initialized using app = Flask(__name__), then the front file made in html is rendered to the app. This code collects the data from the front end and normalizes the data. After the prediction is done the if statement helps render the correct output. This app is then run using the last two lines of the code.

- **Sklearn**

Scikit-learn is a key library for the Python programming language that is typically used in machine learning projects. Scikit-learn is focused on machine learning tools including mathematical, statistical and general-purpose algorithms that form the basis for many machine learning technologies. As a free tool, Scikit-learn is tremendously important in many different types of algorithm development for machine learning and related technologies.

- **Pandas**

Pandas is an open-source library that is made mainly for working with relational or labeled data both easily and intuitively. It provides various data structures and operations for manipulating numerical data and time series. This library is built on top of the NumPy library. Pandas is fast and it has high performance & productivity for users.
Pandas Data Frame is two-dimensional size-mutable, potentially heterogeneous tabular data structure with labeled axes (rows and columns). A Data frame is a two-dimensional data structure, i.e., data is aligned in a tabular fashion in rows and columns. Pandas Data Frame consists of three principal components, the data, rows, and columns.
We will get a brief insight on all these basic operations which can be performed on Pandas Data Frame:

- Creating a DataFrame
- Dealing with Rows and Columns
- Indexing and Selecting Data
- Working with Missing Data
- Iterating over rows and columns

DataFrame can be created using a single list or a list of lists.

```
import pandas as pd
# List of strings
lst = ['Geeks', 'For', 'Geeks', 'is',
        'portal', 'for', 'Geeks']
# Calling DataFrame constructor on list
df = pd.DataFrame(lst)
print(df)
```

- **Standard Scalar**

It standardizes features by removing the mean and scaling to unit variance. Centering and scaling happen independently on each feature by computing the relevant statistics on the samples in the training set. Mean and standard deviation are then stored to be used on later data using transform. Standardization of a dataset is a common requirement for many machine learning estimators: they might behave badly if the individual features do not more or less look like standard normally distributed data.

The standard score of a sample x is calculated as: $z = (x - u) / s$

where u is the mean of the training samples or zero if with_mean=True, and s is the standard deviation of the training samples or one if with_std=False

- **SMOTE**

SMOTE (synthetic minority oversampling technique) is one of the most commonly used oversampling methods to solve the imbalance problem. It aims to balance class distribution by randomly increasing minority class examples by replicating them.

SMOTE synthesises new minority instances between existing minority instances. It generates the virtual training records by linear interpolation for the minority class. These synthetic training records are generated by randomly selecting one or more of the k-nearest neighbours for each example in the minority class. After the oversampling process, the data is reconstructed and several classification models can be applied for the processed data.

More Deep Insights of how SMOTE Algorithm work

Step 1: Setting the minority class set A, for each $x \epsilon A_1$, the k-nearest neighbors of x are obtained by calculating the Euclidean distance between x and every other sample in set A.

Step 2: The sampling rate N is set according to the imbalanced proportion. For each $x \epsilon$ A, N examples (i.e x1, x2, …xn) are randomly selected from its k-nearest neighbours, and they construct the set A_1

Step 3: For each example x ϵ A_1 (k=1, 2, 3…N), the following formula is used to generate a new example: x' = x + rand (0, 1) * | x – x_k | in which rand (0, 1) represents the random number between 0 and 1.

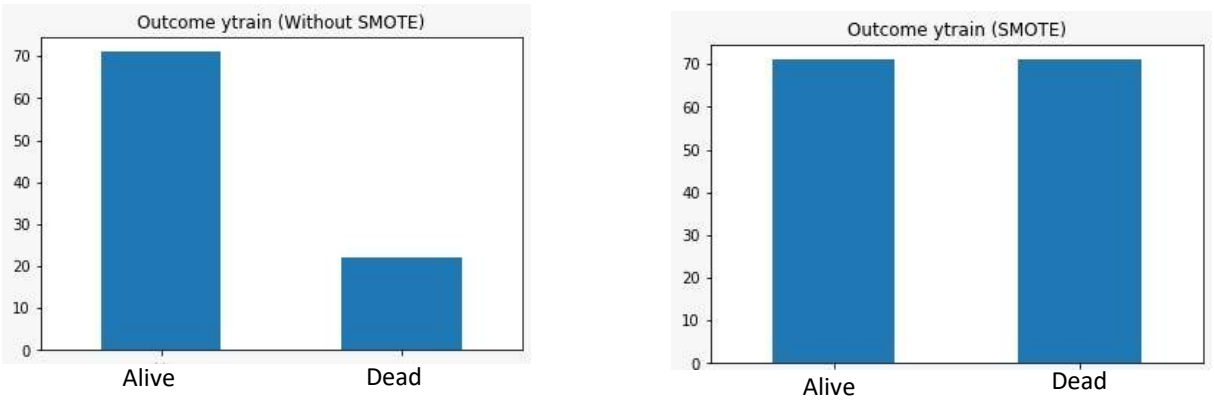

Figure 5.1: Comparison in data before and after SMOTE

5.2 Implementation of the Technologies

```
import pandas as pd
import numpy as np
import pickle
from sklearn.preprocessing import StandardScaler
from imblearn.over_sampling import SMOTE
from sklearn.impute import SimpleImputer
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import train_test_split

columNames = ['Class', 'AGE', 'SEX', 'STEROID', 'ANTIVIRALS', 'FATIGUE',
'MALAISE', 'ANOREXIA', 'LIVER BIG',
              'LIVER FIRM', 'SPLEEN PALPABLE', 'SPIDERS', 'ASCITES',
'VARICES', 'BILIRUBIN', 'ALK PHOSPHATE', 'SGOT',
              'ALBUMIN', 'PROTIME', 'HISTOLOGY']
df = pd.read_csv("hepatitis.data", names=columNames)
df.replace("?", np.nan, inplace=True)
df.isnull().sum()

# Convert the type of numericals
numerical_variables = ['AGE', 'BILIRUBIN', 'PROTIME', 'ALBUMIN', 'ALK
PHOSPHATE', 'SGOT']
df["BILIRUBIN"] = df.BILIRUBIN.astype(float)
df["PROTIME"] = df.PROTIME.astype(float)
df["ALK PHOSPHATE"] = df["ALK PHOSPHATE"].astype(float)
```

```
df["SGOT"] = df.SGOT.astype(float)
df["ALBUMIN"] = df.ALBUMIN.astype(float)

# empty space to mean
df[numerical_variables].fillna(df[numerical_variables].mean()).head(5)

# Categorical variables
categorical_variables = ['SEX', 'STEROID', 'ANTIVIRALS', 'FATIGUE', 'MALAISE',
'ANOREXIA', 'LIVER BIG', 'LIVER FIRM',
                         'SPLEEN PALPABLE',
                         'SPIDERS', 'ASCITES', 'VARICES', 'HISTOLOGY']

# Fmissing data -> most frequent data
imp_mean = SimpleImputer(missing_values=np.nan, strategy='most_frequent')
imp_mean.fit(df)
imputed_train_df = imp_mean.transform(df)
imputedDf = pd.DataFrame(imputed_train_df, columns=columNames)

# normalizing
sc = StandardScaler()
sc.fit(imputedDf.drop(["Class", "PROTIME", "BILIRUBIN"], axis=1))
scaled_features = sc.transform(imputedDf.drop(["Class", "PROTIME",
"BILIRUBIN"], axis=1))
X = scaled_features
y = imputedDf["Class"]

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.4,
random_state=3)
y_train = y_train.astype('int')
y_test = y_test.astype('int')
sm = SMOTE(random_state=33)
X_train_new, y_train_new = sm.fit_resample(X_train, y_train.ravel())

knn = KNeighborsClassifier(n_neighbors=1)
knn.fit(X_train_new, y_train_new)

pickle.dump(knn, open('model.pkl', 'wb'))
model = pickle.load(open('model.pkl', 'rb'))
```

In the above code we can see the utilization of the technologies and processes mentioned earlier. The data is read using pandas and then the data is being cleaned and processed, after the data is normalized using standard scalar. After which SMOTE is applied the data. The data is further split into training and testing data using train test split then this data is fit into the knn model. Pickle is further used to connect the model to the front-end and processed data.

CHAPTER 6

TESTING

For the project we are focusing on 3 algorithms Naïve Bayes, SVM and KNN algorithm. For these algorithms we have built the model using the train data. We are going to use the test data to find the accuracy of these models.

The accuracy of theses algorithms are as follows –

Naïve Bayes – 57%

KNN – 83%

SVM – 85%

As we see the accuracy is under 90%, which is not sufficient in medical field. Also, the accuracy of naïve bayes is very less.

After further testing we find that this caused due to the data that is being used. It is noticed that the class is imbalanced which consist of 26% of the patient die and 74% of the patient alive. Imbalanced data in a classification problem possess a significant challenge in the quality of results obtained through the predictive models.

The dataset consists of imbalance data class, random splitting of training, and a test set produced an unequally distributed class outcome. It is imbalance as alive patient consist of 71 patient and the dead patient is 22. We have applied the SMOTE technique to balance the data using the imblearn library.

Now for the given dataset we have implemented three algorithms, i.e., Naive Bayes, KNN, SVM. We have classified the accuracy of these algorithm with and without the balanced training examples.

The Accuracy of the algorithm using balanced data are as follows –

Naïve Bayes – 82%

KNN – 93%

SVM – 92%

We can see that the accuracy of the model has increased drastically by using balanced data. The imbalanced data created a high volume of data that fall under the same class after the train and test split. This caused the model the learn more about only one class.
SMOTE helped increase the number of values of the other class to make the data balanced. So now the model has learned equally about both the classes. This has helped increase the accuracy.

Algorithm	Balance	Accuracy
Naive Bayes	No	0.5700
Naive Bayes	Yes	0.8185
KNN	No	0.8388
KNN	Yes	0.9304
SVM	No	0.8511
SVM	Yes	0.9238

Table 6.1: Accuracy Table

Test Case ID	Test Case Description	Test Steps	Test Data	Expected Results	Actual Results	Pass/Fail
TU01	Check Customer Valid Data to be tested	1. Go to site http://127.0.0.1:5000 2. Enter the values as required 3. Click on Predict Probability	numerical values = fields provided	User should be able to get the accurate data	As Expected,	Pass
TU02	Check Customer Valid Data to be tested	1. Go to site http://127.0.0.1:5000 2. Enter the values as required 3. Click on Predict Probability	numerical values = fields provided	User should be able to get the accurate data	As Expected,	Pass
TU03	Check Customer Invalid Data to be tested	1. Go to site http://127.0.0.1:5000 2. Enter the values as required 3. Click on Predict Probability	numerical values != fields provided	User should be able to get the in-accurate data	As Expected,	Fail
TU04	Check Customer Invalid Data to be tested	1. Go to site http://127.0.0.1:5000 2. Enter the values as required 3. Click on Predict Probability	numerical values != fields to be filled(*)	User should be able to get the in-accurate data	As Expected,	Fail

Table 6.2: Test Cases

In the above test cases the user should provide valid test data of his/her medical report and based on the data provided the variables are checked for its accuracy and validity. Based on these credentials the output is being predicted.

When the user is providing the invalid set of data to be tested due to the inaccuracy and insufficient data provided the output may vary and result in failure of the test cases

CHAPTER 7

EXPERIMENTAL RESULTS

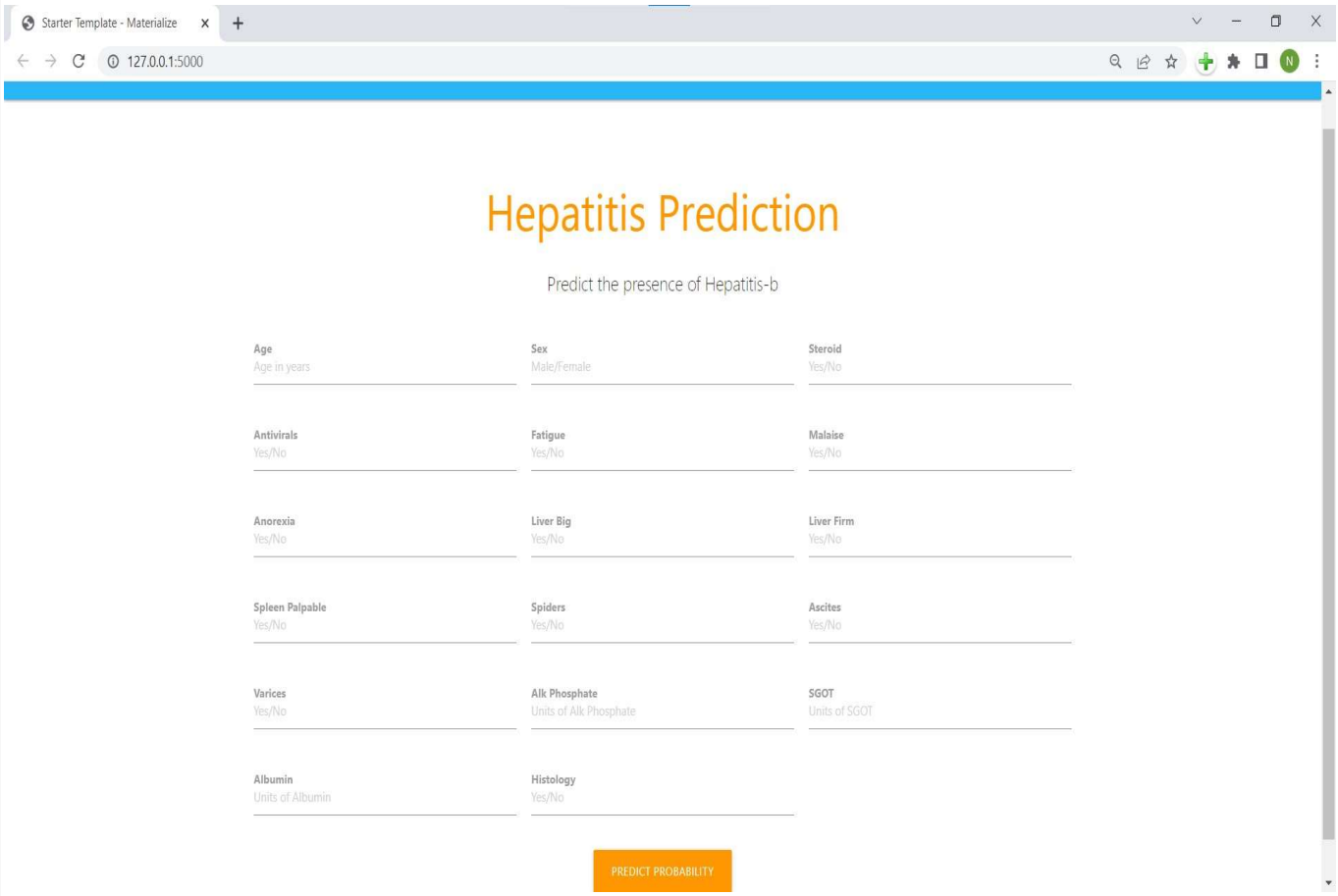

Figure 7.1: Main Page of the Project for user interaction

The main interface of the front-end design has been made for the user to interact with and provide the data to the system. The design is made simple, to ensure that the project is easy to access for any person using it. In the main page of the project the user is asked to enter the personal details related to Hepatitis B virus. The system takes input of 17 attributes that will sent to back-end of the project. In the back-end these values are run on the model and the model provides the result based on predictions. This prediction is then displayed on the main page.

Case 1: When the system predicts No Hepatitis B Virus

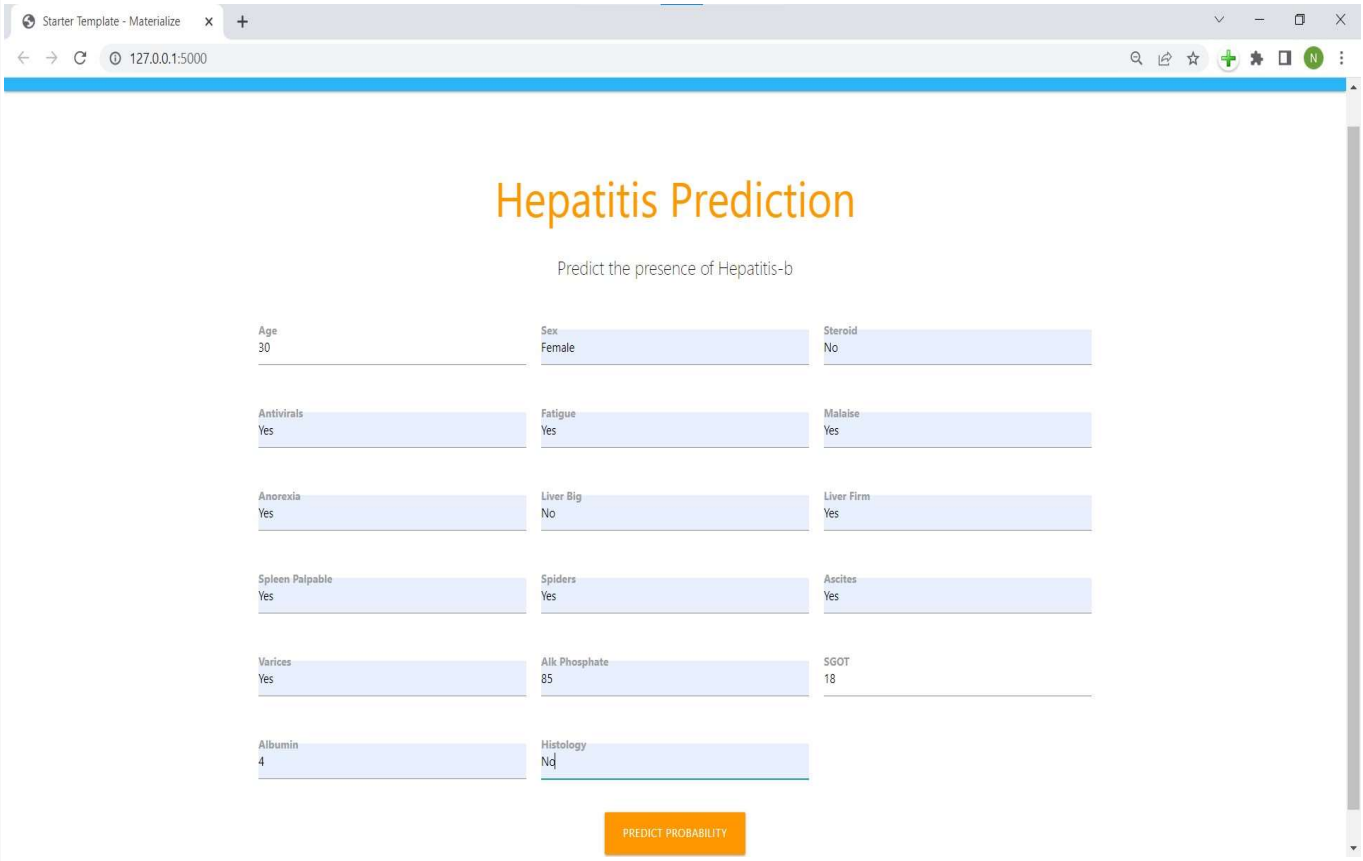

Figure 7.2: Adding data to the project in Case 1

Figure 7.3: The result of low chances of the virus

When the system reads the values provided, the model analyses the data and predicts that there almost no presence of the Hepatitis B virus and person is safe. When the predict button is pressed it process the data in milli seconds and provides the result as shown in the Figure.

Case 2: When the system predicts presence of Hepatitis B Virus

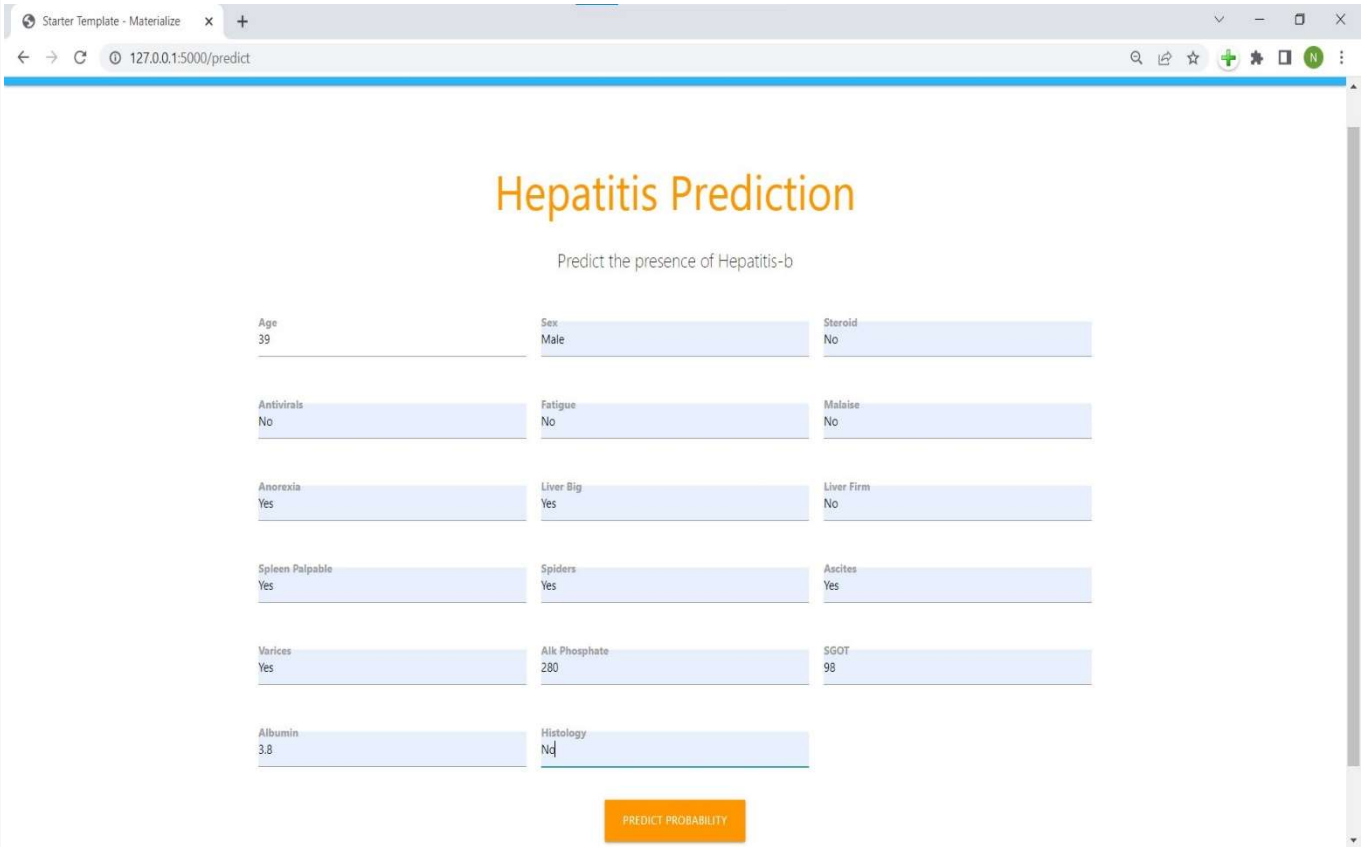

Figure 7.4: Adding data to the project in Case 2

Figure 7.5: The result of High Chance of the virus

When the system reads the values provided, the model analyses the data and predicts that there is presence of the Hepatitis B virus and person is on the verge of death and needs immediate treatment. When the predict button is pressed it process the data in milli seconds and provides the result as shown in the Figure.

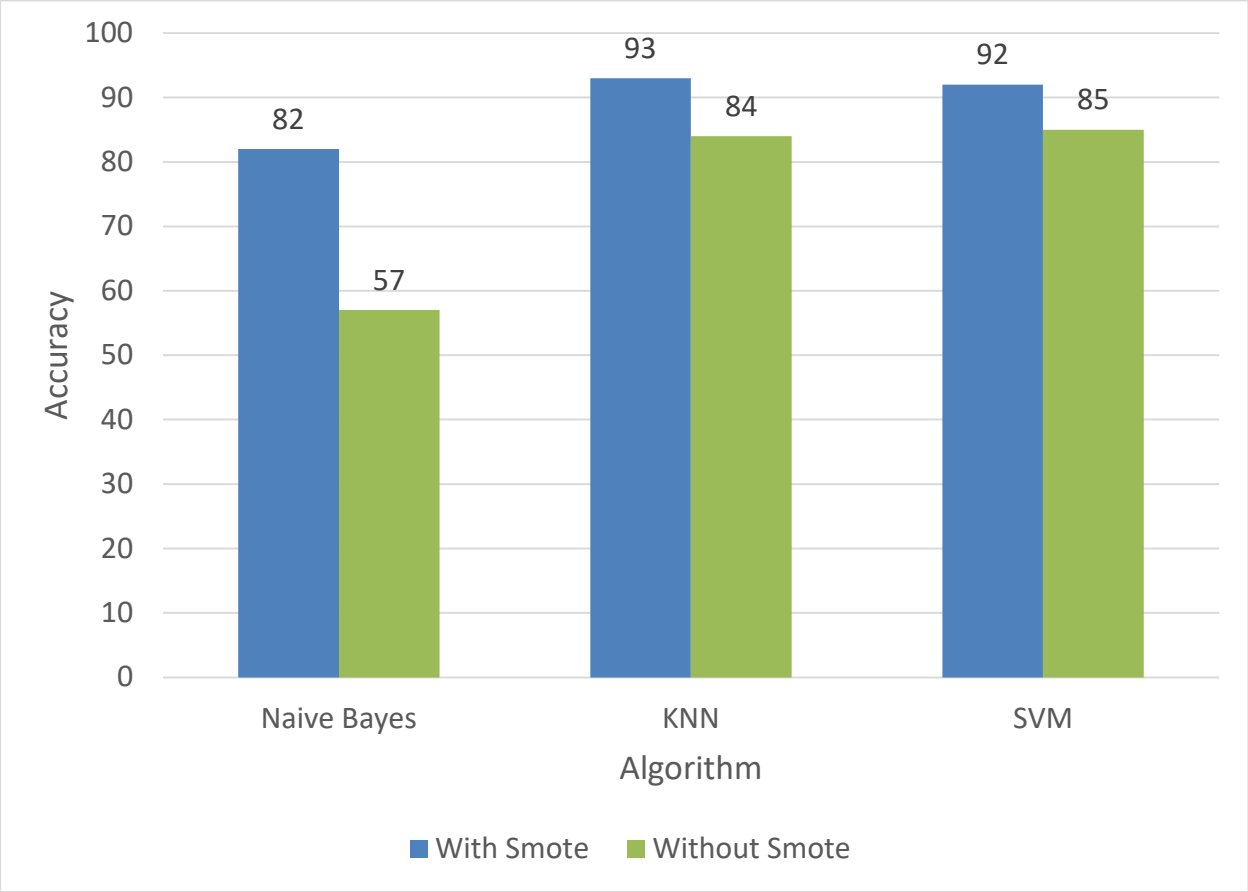

Figure 7.6: Graphical Representation of Algorithms used

In the project, we have tested the prediction system over 3 classification algorithms. These models were tested on the system to find the model with best accuracy. When calculated the accuracy of these models were lower than expected.

When further analysis was performed it was found that this occurred due to the size of the classes in that dataset. To deal with this issue we used SMOTE. It was used to balance the dataset so that the system can learn better and provide better accuracy. As we can see from the Figure the accuracy of the model has increased drastically to provide better accuracy of the system.

CHAPTER 8

CONCLUSION

Hepatitis B is one of the most infectious and dangerous diseases in the world. But because the relevant statistics are voluminous and complex, medical institutions require 1–2 months to release monitoring data, resulting in serious lag between data and the real-time situation. Therefore, it is of great practical significance to predict incidences of Hepatitis B with high levels of accuracy and timeliness. This study applied machine learning algorithms to predict the risk of HBV infection for each patient based on health examination data, and evaluated the predictive effects for the models. Our findings revealed that Borderline-SMOTE sample pre-processing and the KNN algorithm together can be used for disease risk prediction with high classification accuracy, which could better assist clinical decision making and treatment. This risk assessment model can be used to diminish the need for antigen screening among low or non-risk individuals. Through regular follow-up, patients with hepatitis B in the general population can be found earlier. In this we conducted an evaluation and comparison of three well-known machine learning algorithms by regressing status of each patient against laboratory and demographic variables. The results show that machine learning algorithms, especially KNN can predict with an efficient accuracy. This study also showed a potential of machine learning algorithms being used for clinical outcome predictions.

Future Enhancement

- Aims at giving more sophisticated prediction models, risk calculation tools and feature extraction tools for other clinical risks.
- Aid in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications
- Accumulation of larger dataset to enhance and improve the model performance
- Develop a better User Interface to improve the usability for different types of users
- Development of a hardware unit that can perform the analysis in faster and more efficient manner

REFERENCES

[1] Lale Ozyilmaz Tulay Yildirim," Artificial Neural Networks for Diagnosis of Hepatitis Disease", Nov 2009.

[2] Ghumbre Shashikant Uttreshwar, Dr. A.A. Ghatol," Hepatitis B Diagnosis Using Logical Inference and Generalized Regression Neural Networks" March 2009.

[3] G.Sathyadevi, "application of cart algorithm inhepatitis disease diagnosis" June 2011.

[4] V.Shankar sowmien, "Diagnosis of Hepatitis using Decision tree Algorithm" June 2016

[5] Mubashir Hussain, Songsheng Zhu "Rapid Detection System for Hepatitis B Surface Antigen (HBsAg) Based on Immunomagnetic Separation, Multi-Angle Dynamic Light Scattering and Support Vector Machine" June 2020

[6] Huina Wang, Wei Huang "Random Forest and Bayesian Prediction for Hepatitis B Virus Reactivation" 2017

[7] Yongwang Zhao, Yihui Liu "Prediction Model of HBV Reactivation in Primary Liver Cancer - Based on NCA Feature Selection and SVM Classifier with Bayesian and Grid Optimization" 2018

[8] Daren Zhao, Huiwu Zhang "The research of SARIMA model for prediction of hepatitis B in mainland China" June 2022

[9] Nabeel Ali, Dolley Srivastava, Aditya Tiwari "Predicting Life Expectancy of Hepatitis B Patients using Machine Learning" April 2022

[10] Tahira Islam Trishna, Shimoul Uddin Emon, "Detection of Hepatitis (A, B, C and E) Viruses Based on Random Forest, K-nearest and Naïve Bayes Classifier" December 2019

[11] Lailil Muflikhah, Widodo, Wayan Firdaus Mahmudy "DNA Sequence of Hepatitis B Virus Clustering Using Hierarchical k-Means Algorithm" June 2020

[12] Gian Paolo Caviglia, Yulia Troshina "Usefulness of a Hepatitis B Surface Antigen-Based Model for the Prediction of Functional Cure in Patients with Chronic Hepatitis B Virus Infection Treated with Nucleos(t)ide Analogues: A Real-World Study" July 2021

[13] Changjiang Long, Huan, “Modeling the virus and immune system of chronic hepatitis B”, April 2014. [4] Na Chu, Lizhuang Ma, “Attribute Weighting with Probability Estimation Trees for Improving Probability based Ranking in Liver Diagnosis”, March2010.

[14] Na Chu,” An Intelligent Diagnosis Method for Chronis Hepatitis B in TCM” June 2013.

[15] Che Lijuan, Zhou Qiang,” Clinical Study on Data Mining of TCM Zheng in Chronic Hepatitis B “, April 2014.

[16] C.Mahesh, K.Kiruthika, M.Dhilsathfathima, Assistant Professors, Department of Information Technology Veltech Dr. RR& Dr.SR Technical University. ” diagnosing hepatitis b using artificial neural network based expert system”, March 2014.

[17] Wen Shen, Zhihua Wei, Yunyi Li,” Multiple granular analysis of TCM data with applications on hepatitis B”, April 2015.

[18] Na Chu, Zhiying Che, Xiaoyu Chen, Min Zhou, Lizhuang Ma, and Yu Zhao. Hybrid Feature Selection based on Multi-view for Improved Diagnosis of Chronic Hepatitis. Chinese Journal of Integrative Medicine, Article ID 20110274.

[19] Tao Wang, "Model of Life Expectancy of Chronic Hepatitis B Carriers in an Endemic Region. Journal of Epidemiology", 2009

[20] Somaya Hashem, Gamal Esmat, "Comparison of Machine Learning Approaches for Prediction of Advanced Liver Fibrosis in Chronic Hepatitis C Patients. IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM." 2017

[21] Yaming Zhang, "Modeling for the prediction of Hepatitis B incidence based on integrated online search indexes. Informatics in Medicine Unlocked.", 2018

[22] J.Wolfson, "A Naïve Bayes machine learning approach to risk prediction using censored, time-to-event data. US National Library of Medicine National Institutes of Health (NCBI).", 2015

[23] Xiaolu Tian, Yutian Chong, "Using Machine Learning Algorithms to Predict Hepatitis B Surface Antigen Seroclearance. Hindawi Computational and Mathematical Methods in Medicine.", 2019

[24] Hailemichael Desalegn, "Predictors of mortality in patients under treatment for chronic hepatitis B in Ethiopia. BMC Gastroenterology.", 2019

[25] Mingxue Yu, Xiangyong Li, "Development and Validation of a Novel Risk Prediction Model Using Recursive Feature Elimination Algorithm for Acute-on-Chronic Liver Failure in Chronic Hepatitis B Patients with Severe Acute Exacerbation. Frontiers in Medicine.", 2021

[26] http://www.hepb.org/patients/journal_articles.html

[27] Machine Learning Repository, www.archive.ics.uci.edu/ml/datasets.html

[28] UCI Machine Learning Repository http://www.ics.uci.edu/~mlearn/MLRepository.html