Disease Prediction from Symptoms

A major project

submitted in partial fulfillment of the

requirements for the award of the degree

of

Bachelor of Technology

in

Eletronics and Communication Engineering

Under the guidance of

Mr. Gagnesh Kumar by

Yugansh Jain (184506)

Peeyush Thakur (184510)

Rohit Bhatia (184518)

Tejesh Reddy (184528)

Department of Electronics and Communication Engineering

National Institute of Technology

Hamirpur – 177005 (India)

May, 2022

CERTIFICATE

I, hereby, certify that the work which is being presented in the B.Tech. major project report entitled **Disease Prediction from Symptoms**, in partial fulfillment of the requirements for the award of the Bachelor of Technology in Electronics and Communication **Engineering and submitted to the Department of Electronics and Comm. Engineering** of National Institute of Technology Hamirpur is an authentic record of my own work carried out during a period from Jan,2022 to May 2022 under the supervision of Mr. Gagnesh Kumar, Assistant Professor, Department of Electronics and Communication Engineering, National Institute of Technology Hamirpur.

The matter presented in this report has not been submitted by me for the award of any other degree elsewhere.

Yugansh Jain(184506)
Peeyush Thakur (184510)
Rohit Bhatia (184518)
Tejesh Reddy(184528)

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Date:

Mr. Gagnesh Kumar
Assistant Professor

Signature of Supervisor

Signature of Co-cooardinator

Acknowledgments

We take this opportunity to express our deep sense of gratitude and sincere thanks to all who helped us to complete the work successfully. Our first and foremost thanks goes to God Almighty who showered in immense blessings on our effort.

We wish to express our sincere thanks to Mr. Gagnesh Kumar, Assistant Professor, **Deptt. of Electronics & Comm. Engineering, NIT Hamirpur for providing us with** all the necessary facilities and support.

We wish to express our sincere gratitude towards all the teaching and non teaching staff members of our Department.

Finally we thank our parents, all our friends, near and dear ones who directly and indirectly contribute to the success of this work.

Yugansh Jain(184506)

Peeyush Thakur (184510)

Rohit Bhatia(184518)

Tejesh Reddy(184528)

Contents

Abstract

Disease Prediction plays an important role in health care information. It is important to diagnose the disease early. This project uses selective traits and classification strategies to diagnose and predict diseases. Adequate selection of features plays an important role in improving the accuracy of classification systems. The reduction in size helps to improve the overall performance of the machine learning algorithm. The use of differentiating algorithms in the disease database produces promising results by developing flexible, automated and intelligent diagnostic programs for diseases. Classification systems can be used to speed up the process and improve the efficiency of result statistics. This work introduces a comprehensive overview of the options for selecting various features and their pros and cons. We then analyzed the flexible classification systems and classification systems for the same stages of disease forecasting.;

Chapter 1

INTRODUCTION

Diagnosis of diseases at early stage is very crucial in the treatment of the disease. Diagnosis of diseases helps to take preventive measures and effective first-line treatment has been found to be helpful to patients.

Currently, the maintenance of health information has become an important part of the medical field. Patient data incorporating a variety of features and diagnostics related to the disease should be entered with great care in order to provide quality services. Since the data stored on a medical website may contain missing amounts and non-essential data, the extraction of medical data becomes more difficult. As it may affect mining results, it is important to have good data processing and data reduction before using data mining algorithms. Disease prognosis becomes simpler and easier if the data is accurate and consistent and free of noise.

Feature Selection is an effective way to process data in data mining to reduce data size. In a medical diagnosis, it is very important to identify the most dangerous factors associated with the disease. Identifying the right factor helps to remove unnecessary, unnecessary attributes from the disease database, which, in turn, provides faster and better results.

Planning and prediction is a data mining method that first uses training data to improve the model and then the resulting model is used in data analysis to obtain predictive results. Classification algorithms have been used in the disease database to diagnose diseases and the results are very promising. There is a great need for a novel novel diagnostic process that can speed up and simplify the process of diagnosing chronic diseases.

This project is organized as follows. Firstly, a short description of feature selection for disease prediction is presented. Secondly, various feature selection methods and related work on various feature selection approaches is presented. Thirdly, a survey on

traditional classification systems and adaptive classification systems for chronic disease prediction is shown.

Chapter 2

FEATURE SELECTION FOR DISEASE PREDICTION

Feature selection, also known as Variable Selection, is an extensively used data preprocessing technique in data mining which is basically used for reduction of data by eliminating insignificant and superfluous attributes from any dataset. Moreover, this technique enhances the comprehensibility of data, facilitates better visualization of data, reduces training time of learning algorithm and improves the performance of prediction.

There exist numerous applications of relevant feature identification techniques in healthcare sector. Filter methods, wrapper methods, ensemble methods and embedded methods are some of the popularly used techniques used for variable selection. In recent years, most of the authors are focusing on hybrid approaches used for feature selection.

Before any model is applied to the data, it is always better to remove noisy and inconsistent data to get more accurate results in less time. Reducing the dimensionality of a dataset is of paramount importance in real-world applications. Moreover, if most important features are selected, the complexity decreases exponentially.

In recent years, various feature identification approaches have been applied on healthcare datasets to get more valuable information. The utilization of feature selection methods is done on clinical databases for the prediction of numerous diseases like diabetes, heart disease, strokes, hypertension, thalassemia etc. Various learning algorithms work efficiently and give more accurate results if the data contains more significant and nonredundant attributes. As the medical datasets contains large number of redundant irrelevant features, an efficient feature selection technique is needed to extract interesting

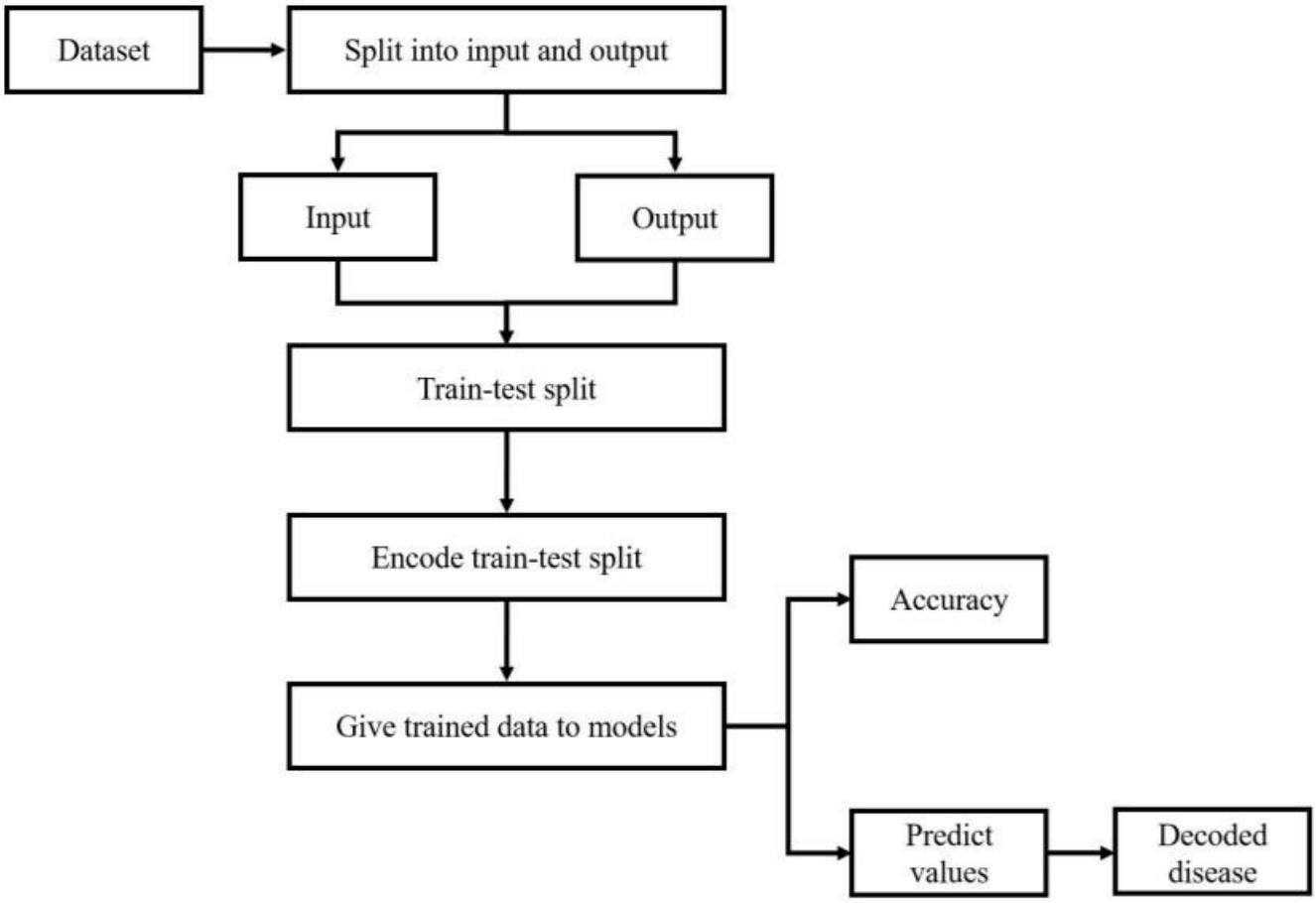

Figure 2.1: Diagrammatic representation of data flow features relevant to the disease.

A basic taxonomy of feature selection and various gene selection methods were reviewed under three divisions – supervised, semi-supervised and unsupervised feature selection. Various challenges and obstacles in extracting knowledge from gene expression data were also addressed. Some of the basic issues raised were (1) reducing dimensionality of data with hundreds of thousands of features (2) how to handle mislabeled, imprecise data (3) how to deal with extremely imbalanced data (4) determining the gene relevancy/redundancy and extracting relevant biological information from the gene expressions. It was revealed through comparative study on gene selection that the classification accuracy of semi-supervised and unsupervised approaches was as promising as supervised feature selection.

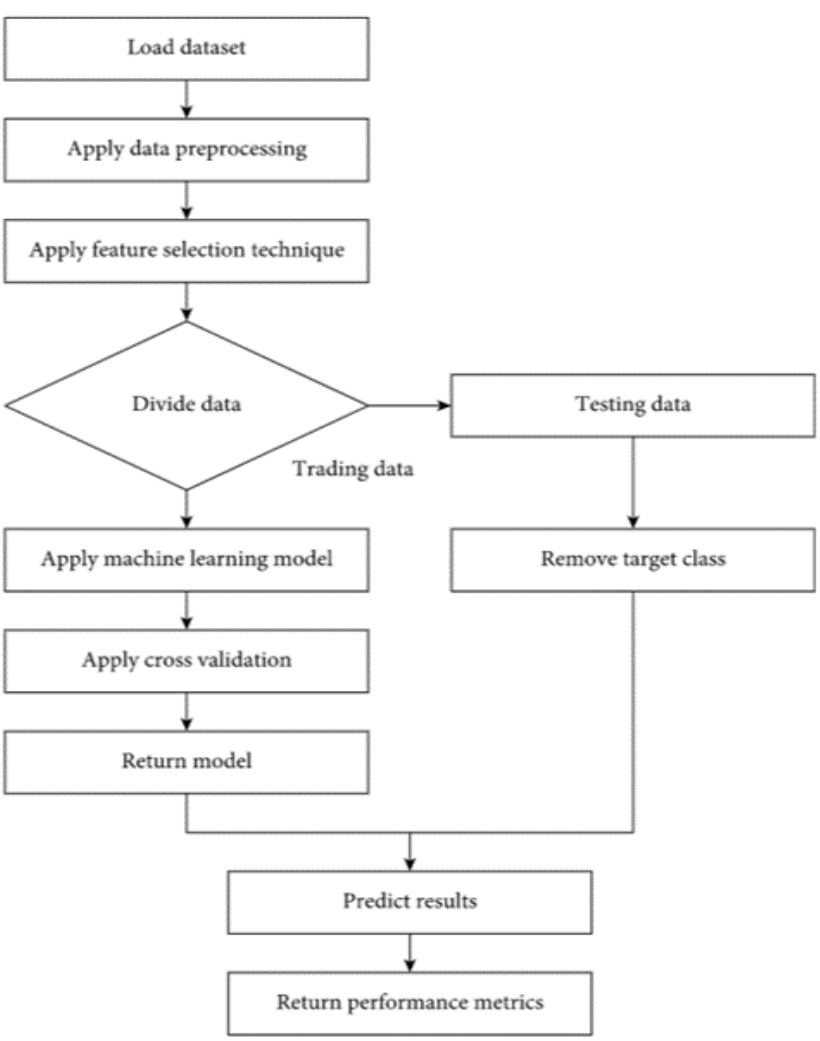

Figure 221: Diagrammatic representation of the workflow

Chapter 3

DECISION TREE ALGORITHM

3.1 How does Decision Tree work?

The methodology used in the Decision tree is a commonly used data mining method for establishing classification and prediction systems based on multiple explanatory parameters for developing prediction models for a target instance. This path classifies a population into branch-like segments in a tree that construct an inverted tree with a root node, internal nodes, and leaf nodes. A decision tree is a non-parametric algorithm which can efficiently deal with huge, complicated data sets without involving multiple parametric structures. If the sample size is large enough, study data can be divided into training and validation data sets. Using the training data set to build a decision tree model and a validation data set decide on the appropriate tree size to achieve the optimal final model.

3.2 Execution

The Decision tree works with the underlying symptoms and predicts a disease. Initially, we get the user's top five symptoms and put it in an array with the value assigned as 1 across these values. This is passed as an input to the model for predicting the disease. This array matches the disease data collection and ends at a common leaf node with the highest degree of trust.

3.3 Recursive Part

In the recursive part, we repeat the above-mentioned approach with increasing tree-level in order to construct the tree. We set the current node as a leaf node when there is no question to ask if the output is published for the symptoms given. We also use electronic

health records to expand the dataset with more diseasesymptom pairs for better prediction of the disease based on the symptoms.

3.4 General rules for Building Decision Tree

- Choose the best attribute / feature.
- The best attribute is the one that best separates or divides the data into subsets.
- The recursive process ends when all elements belong to the same attribute, no more attributes and instances are left.

Chapter 4

SUPPORT VECTOR MACHINE (SVM)

4.1 Definition

"Support Vector Machine" (SVM) is a supervised machine learning algorithm that can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is a number of features you have) with the value of each feature being the value of a particular coordinate. The dimension of the hyperplane depends upon the number of features. If the number of input features is two, then the hyperplane is just a line. If the number of input features is three, then the hyperplane becomes a 2-D plane. It becomes difficult to imagine when the number of features exceeds three.

4.2 How does SVM works?

The working of the SVM algorithm can be understood by using an example. Suppose we have a dataset that has two tags (green and blue), and the dataset has two features x1 and x2. We want a classifier that can classify the pair(x1, x2) of coordinates in either green or blue. So as it is 2-d space so by just using a straight line, we can easily separate these two classes. But there can be multiple lines that can separate these classes.Hence, 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.

SVM Kernel:

The SVM kernel is a function that takes low dimensional input space and transforms it into higher-dimensional space, ie it converts not separable problem to separable problem. It is mostly useful in non-linear separation problems. Simply put the kernel, it does some extremely complex data transformations then finds out the process to separate the data based on the labels or outputs defined.

4.3 Advantages of SVM:

- Effective in high dimensional cases
- Its memory efficient as it uses a subset of training points in the decision function called support vectors
- Different kernel functions can be specified for the decision functions and its possible to specify custom kernels

Chapter 5

RESULTS AND DISCUSSION

5.1 Result using DECISION TREE CLASSIFIER

Accuracy using decision tree classifier is 97.31

5.2 Result using SVM

Accuracy using SVM is 98.11

Figure 5.1: Decision Tree.

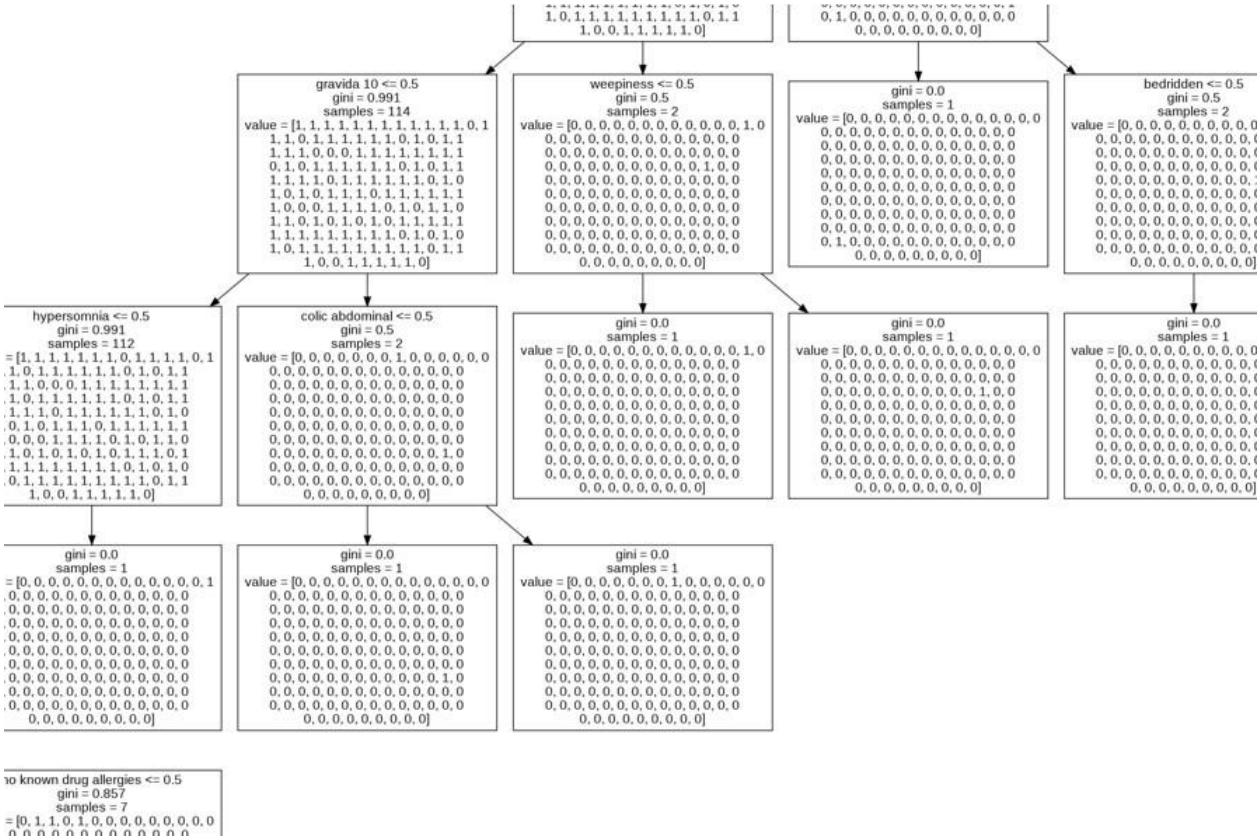

Figure 5.2: zoomed decision tree.

References

1. https://www.hindawi.com/journals/cmmm/2020/9689821/materials-and-methods
2. https://arxiv.org/ftp/arxiv/papers/1801/1801.05412.pdf
3. https://scikit-learn.org/stable/modules/svm.html
4. https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html