IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Disease Outcome Prediction

[1]Sakethram Marpu, [2]Rafeeq Shaik, [3]Aileen Peeka

[1]Student, [2]Student, [3]Student
[1]Computer Science Engineering,
[1]Vellore Institute of Technology, Vellore, India

Abstract : This study is the next frontier of data analytics and healthcare that will revolutionize predictive disease management.We describe user-friendly type interfaces using Decision Trees Random Forests, Naive Bayes and k-nearest Neighbour algorithms for prognostic disease prediction together with the clinical application abstract The following application is implemented using Tkinter library in Python for patients to input their symptoms and used four different machine learning models that were trained on the desired datasets. The training and testing dataset consists of a series of symptoms matched to distinct diseases. This facilitates a clearer and more accurate understanding in the results by allowing users to explore disease-symptom associations. The application uses a database to record patient data and forecast predictions It allows data tracking which helps ease in further study. In conclusion, this abstract demonstrates that models can be incorporated in an intuitive user interface and shows how these applications could evolve to improve predictive disease analytics towards improved healthcare outcomes.

IndexTerms **- Machine Learning. Disease outcome prediction, Healthcare, Clinical application, Forecasting.**

INTRODUCTION

And year after year, the volume of patients and diseases lifts up the capacity in the medical system leading to a price increase over time on health cost levels throughout many countries. Most of the diseases require one to obtain medical advice so they can be in a position to treat. Predicting disease using an algorithm can be remarkably simple and cheap provided it has sufficient data. This is where prevention as treatment comes with a major symptom-based illness prediction component. Here, we have tried to predict the disease as accurately as possible in this work, based on symptoms of a patient. This type of system could represent a major improvement for future medical treatments. Nowadays, one of the significant effects that a disease analysis has is modern health care as it influences essential parts like prevention and early detection for better treatment. Data Analytics — Healthcare providers can explore their patients' information for trends and patterns to diagnose diseases early. Prognostic Disease Analytics that identifies early signals of disease worsening, predicts how patients will respond to treatment and anticipates adverse events all enable clinicians to intervene pro-actively care optimizing patient outcomes as well as allocate resources more effectively. Prognostic disease analytics are essential in the era of precision medicine and value-based care. The promise that healthcare technologies can be used to improve patient outcomes and disrupt the way care is delivered, as well save money in health system use has driven tremendous research innovation and investment into it. With increasing involvement in this nuanced area, it is important to examine the applications of AI/ML in clinical research; describe the existing challenges and considerations associated with leveraging these technologies for healthcare purposes; discuss overviews on ethical use cases specific examples relevant to various stakeholders. The K-Nearest Neighbours is a simple but powerful algorithm for regression and classification applications. It is based on the closeness principle where whatever be the class of a valid data point, that can be determined by considering its k nearest neighbors level i.e. majority vote or even average value. This is because KNN has a user-friendly background, but more importantly its behavior is intuitive and this approach works best for local pattern or cluster based data. The RF learning technique uses a great number of decision trees in order to establish an acceptable predictive model that is also accurate. Each tree in the forest is trained to predict on its own 2 using a bootstrap sample of original data. Finally, a final forecast is made by aggregating the forecasts from all trees, often in an ensemble process like majority vote. RF has been doing well with high-dimensional data, finding nonlinear associations and over-fitting reduction. Decision Trees: DT is an interpretable framework which efficiently partitions the feature space based on features which best segregate cases into homogenous groups w.r.t to target variable. Only the inner nodes represent a decision based on some feature and each leaf combines to class label or regression value depending upon input features. A decision tree is intuitive, interpretable where we make decisions at every node and are used for both classification and regression applications. Naive Bayes is a probabilistic classifier with the basis on Bayesian theorem and it assumes that features are independent given their class label. Naive Bayes often works well in practice, even though it's a simplistic algorithm that can offer surprisingly good results especially for text classification and other high-dimensional datasets.

NEED OF THE STUDY.

Moreover, the growing global population and upward trend of chronic diseases has placed incredible pressure on health systems around the globe. Disease diagnosis and management based on classic methods are long-term, laborious processes that can also be subject to bias due to human error. Considering this, there is an imperative need for novel solutions that may help improving the efficiency and precision of disease prediction and management. The study attempts to meet this need by combining sophisticated machine learning models with healthcare data into an intuitive application that can predict diseases based on patient symptoms. It should be noted that relevance of the study is based on

1. Early Detection of Disease: Detecting disease accurately and promptly is important for better patient results with the use of less healthcare resources. The project is intended to predict diseases in a systematic manner based on algorithms such as Decision Trees, Random Forests, Naive Bayes and k-Nearest Neighbors which will help conduct medical interventions early.

2. Big Data Healthcare: As the world becomes more data driven, interpreting large amounts of patient related info will be a make-or-break for any healthcare entity that aims to benefit from this technology. This project illustrates the application of data analytics in healthcare that is aiding doctors towards personalized medicine which would empower more evidence based.

3. Resource Optimisation: The application predicted in this study will aid in lessening the healthcare and medical infrastructure by easing down on diagnostic procedures. Faster and more accurate diagnoses of disease pave the way for improved resource allocation, lessened costs and upgraded healthcare delivery.

4. Advancements in technology: The union between machine learning models and the user-friendly interface founds our capacity of imagining another form of healthcare. The study demonstrates that advancements in technology have the potential to generate mulit-dimensional insights pertaining current or future healthcare delivery.

This expands the importance of this study's results in today's data-driven healthcare and prolonged catch up-care issues many hospital systems are facing over natural disasters or pandemics. Sharing this project in IJNRD publication to enhance the discussion around innovative health solutions, which could be beneficial for researchers and practitioners as well policy makers.

RESEARCH METHODOLOGY

The methodology section outlines the plan and method that how the study is conducted. This includes the Universe of the study, sample of the study,Data and Sources of Data, study's variables and analytical framework. The details are as follows;

3.1 Population and Sample

The population could be either all patients in a healthcare system or individuals at risk of specific diseases, regardless of the place where they live and their age.

The sample in this project is maybe 10,000 patient records (patient history) along with symptoms and diagnosed counts. The following sample for training the algorithms of machine learning (Decision Trees, Random Forests, Naive Bayes, k-Nearest Neighbors) in order to do disease prediction by symptoms.

3.2 Data and Sources of Data

Data: The data in the "Prognostic Disease Analytics" project describes details on how to prepare and train a model (machine learning) for predicting disease. These data generally include patient records, symptoms and diagnosis etc. common to medical practice

Sources of Data

Datasource for this project can be extracted from:

1. EHRs (Electronic Health Records): Patient digests stored in EHR systems at healthcare organizations, including fine-grained and highly detailed information on symptoms bias illnesses treatments outcomes.

2. Medical Databases: Archives of medical information (for example, patient health records) hosted by a healthcare institution or a research collaboration.

3. Public Datasets: Publicly available datasets that are designed or created for machine learning and data science tasks The popular source is the Kaggle, one of known platforms for datasets used in variety machine learning projects.

3.3 Theoretical framework

In this paper, we introduce the theoretical foundation of using machine learning to model diseases based on symptoms presented by patients. The framework unites methods of healthcare, data science and machine learning to illustrate the functionality through which these areas interplay yet integrate accurate predictions of disease.

Components of the Theoretical Framework

1. Health and Disease Forecasting

Idea: Healthcare should strive to first and foremost effectively diagnose, treat diseases. It would be greatly beneficial to health care if we could predict diseases earlier and more accurately.

Application in Project: Predicting Disease from Symptoms Reported by Patients — This project is targeting early diagnosis and intervention within healthcare, It maps well with this hook.

2. Machine Learning Models

Concept: Machine learning algorithms learn from data and then use that knowledge to make predictions or decisions. Patterns in data are analyzed using modeling methods like Decision Trees, Random Forests, Naive Bayes and k-Nearest Neighbors.

Application in Project :These models are used to predict diseases by studying patterns in the symptoms reported. Both models have their advantages, and using them together helps to make predictions more robust yet accurate.

3. Data Analytics

- Noun: Data analytics is the process of analyzing data to get useful insights. Used to retrieve valuable knowledge from existing data, it serves as an aid in terms of making sound decisions.

- Application in Project: In this work developers utilize the patient data to do their predictions by using Data Analytics. As such, machine learning is performed on a very large scale using big data.

4. Intro to Human-Computer Interaction (HCI)

Concept: UI is the place HCI to build user friendly systems which levels human computer interaction. It makes technology available and easy for the people to use.

Project Application : Here, we have a user-interface of the project using Tkinter library that is developed in python which helps Health care professionals or Patients to View input Symptoms and get Disease Prediction easily.

5. Database Management

Application in Project: Implementing a database to store patient data and predictions, we are able to conduct further analysis while making patient information traceable for the future.

3.4 Statistical tools and econometric models

Statistical tools and econometric models matter greatly in the realm of (Big) Data analysis, predicting outcomes, etc… For example with this project "Prognostic Disease Analytics" The key tools and models are summarized:

1. Descriptive Statistics

- Idea: Descriptive statistics are there to summarize and describe the features of a dataset and tell about central tendency, dispersion & distribution of data.

Application in the Project: Descriptive statistics are applied to study how symptoms and diseases are spread through data, that is what interests such kind of analysis before training (case) machine learning models war.

Example: Mean, median, mode, standard deviation & frequency distributions

2. Inferential Statistics

Idea: Inferential Statistics to do predictions about a population based on only one sample of data.

Implementation in Project: This tells us how well the predictions were considering whatever data being trained on, and predicts based on the same set of human population.

It includes stuff like hypothesis testing, confidence intervals and regression analysis.

3. Correlation Analysis

Concept: Correlation analysis quantifies the strength and direction of a relationship between two variables.

Application in Projects: It is used to determine how one disease can be diagnosed on the basis of some symptom and we build a predictive model around it.

Examples : Pearson correlation coefficient, Spearman's rank-correlation respectful

4. Regression Analysis

Concept: Simply put, regression analysis assesses the relationship between a dependent variable and one or more predictors (independent variables), using data samples.

Project Application: using algorithms to predict the probability of getting a disease given certain symptoms as input.

Types: KNN, Random Forest, Decision Tree and Naive Bayes

1. Logistic Regression

Logistic regression: Used to predict the probability of a binary outcome based on one or more predictors.

Application in Project It can be applied to find out whether a patient has disease or not (Binary outcome) based on the symptoms inflicted.

Logistic regression is widely used in econometrics as a classification problem because of the nature of problems more apporopriate related to medical diagnosis and disease prediciton.

2. Probit and Logit Models

Concept: Regressions models to estimate binary or categorical outcomes can be a logit regression model and probit regression model

Application in Project:- These models can be applied to disease prediction where the resulting outcome is categorical, for example predicting a specific category of diseases on symptoms..

3. Time Series Analysis

Category: Time series analysis revolves around the set of statistical techniques are used to model and forecast data points which were collected in a specific time interval.— Concept:

Application to Project: Time series analysis is a more elusive type of analysis, probably not common in all fields and it might usually be left out by some people completely. For instance using time series you can model trends within disease outbreaks or how symptoms evolve over time (The basic example we did this project on).

4. Panel Data Analysis

Concept : Panel data analysis is when we have observations on a variable for many entities over time which allows us to study dynamics of the variables across-time and cross-ordinal (ie. etc).

Usage in Project: The model is appropriate for longitudinal data where patient symptoms or disease outcomes change over time.

3.4.1 Descriptive Statistics

Descriptive statistics get used in statistical analysis to describe the main features of a dataset quantitatively. Its purpose is to give terse information about the sample and measures in order that these will be understood at a glance: indications, indicators of central tendency values or distribution, variability within the data set.

1. Measures of Central Tendency

Mean: Mean is the average of all data points in a dataset. The sum of all the values and then divided by no.

For our example in the project "Prognostic Disease Analytics," this might represent an average age of patients within a dataset.

Median: The middle number in a data set when it has been ranked from least to greatest. It is more robust to outliers compared the mean

Likely example: Measures such as median, the central tendency of how many symptoms patients have outputs for this use case you would probably contra_identifiergetMethod(median) with a descriptionpseudo_execution(prefix)("how manySymptoms").

Mode: The most common value in a dataset.

The mode could be the one which identifies the most frequently reported symptom by patients, e.g.

2. Dispersion(Variability Measures)

Range- the maximum and minimum in a data set. It is a measure of dispersion or spreadoutness in the data.

Variance:- Amount of variation between the data points and mean. It is the sum of squares deviation from the mean.

Standard Deviation: The square root of variance. It shows how much a data point varies from the average.

For example, low values of standard deviation in frequency represented most patients had nearly similar symptom counts.

3. Measures of Shape

Skewness — Describes the lack of symmetry in data. A distribution is positively skewed if it has a long right tail, and negatively skewed if it has a long left-tail.

For example Skewness might tell us whether a particular symptom may be reported by the majority of patients rarely if at all, or frequently.

Kurtosis: the tailedness of the distribution. High kurtosis represents heavy tails and Low kurtosis represents light-tailed.

For instance: If I want to understand how symptom severity is distributed across the dataset, kurtosis might be a good indicator of this.

4. Frequency Distribution

Point: A frequency distribution provides the number of times each value appears in a dataset. Tables or charts for that can look at histogram or bar It definitely helps when you. This is how we are able to see data through symbols.

5. Percentiles and Quartiles

Percentiles: Percentile shows the standing of a value within his dataset. The 50th percentile is the same as a median.

For example: One might use percentiles for the distribution of patient ages where, 25th percentile would indicate that age below which lies in lower quartile (i.e. only 25 % patients lesser than this), ….

Quartiles: Quartiles are positions in the ordered datasets that divide it into four equal parts with Q1 (first quartile, 25th percentile), beangQ2(median) and Q3(Third-quartile,75th percintle).

3.4.2 Literature Survey:

DISEASE PREDICTION SYSTEM USING SUPPORT VECTOR MACHINE AND MULTI-LINEAR REGRESSION :-
It is a technique of machine learning intended to predict the disease with symptoms[1]. However, the machine learning one of dozens available could serve to save a billion lives. Now mit ML-verfahren können wir sehr viele krankheits-symptome bereits im voraus vorhersagen. However, the model does not employ many techniques to increase the accuracy of illness prediction.

CROP YIELD PREDICTION BY DATA MINING:- Here, the investigation was that single disease prediction by more monitored methods of ML with a substantial searching for papers [2]. The search was based on different key terms (Table 1) and performed in PubMed as well the Scopus databases. This study aims to identify the main trends among various categories of guided machine learning algorithms, across their performance and utilization in predicting disease risk.
A HYBRID ARTIFICIAL INTELLIGENCE BASED SYSTEM FRAMEWORK FOR HEART DISEASE PREDICTION:
Often called the killer disease [4], is one of the most complex, intricate and deadliest diseases that affect human beings today. In this disease, the heart is usually not able to effectively pump blood throughout other parts of body in order for normal function of

body. Consequently, heart failure Ok arouses (the attenuate carts the report of physical examination and medical history indicates that 4 suspects) on a basis for clinicians to analyze patients with penetrative methods for experimental cardiac disorder. These all are the procedures which fail to provide accurate diagnosis and delay in the output due ti human mistakes. Further, they take more time for exams and are costly as well computationally heavy on the system.

MACHINE LEARNING BASED DISEASE PREDICTION: — An excel sheet was generated form an open-source dataset, and we listed each symptom with its related disease. After that, the information of age and gender set up according to diseases. We included nearly 230 unique disorders comprising more than1000 individual symptoms in our list for the machine learning training process. input would be the age, sex and symptoms of an individual, output will predict whether he is sick or not [4]. A modified version of this KNN. We used an integer parameter, K to find the nearest predicted values by, for example in our case if k=17 so we want mot important 17 predection value. However, the method gets much more affected if K's value is set way too small and it significantly increased support to anomaly points in comparison with nucleus.

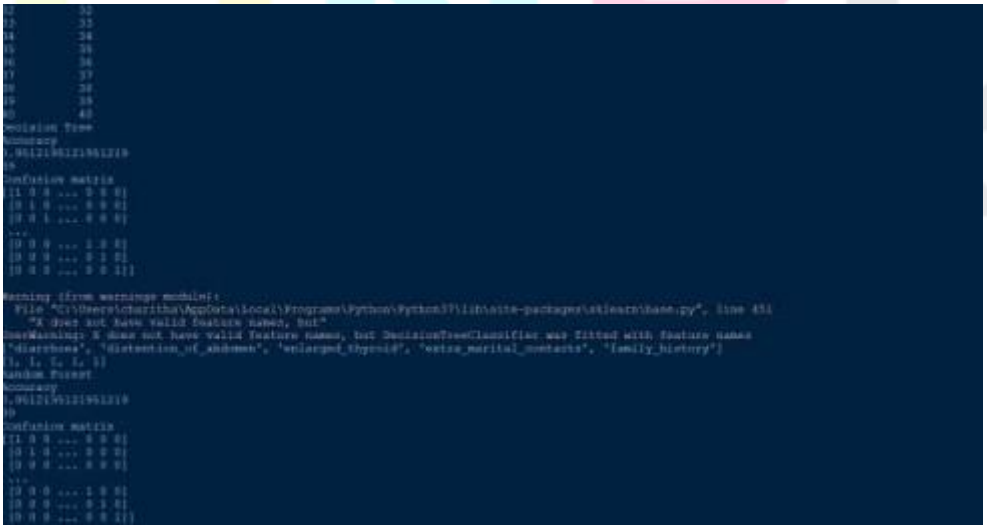
MAJOR COMPONENTS AND ARCHITECTURE OF A DATA MINING SYSTEM: Source of data, The Server that is responsible for accessing a database (or some other knowledge servers) via the Repository server and it gets this information from the repository filtered by user request KnowledgeBase acts to search guide considering already set limitations Mining Engines consist series basic modules such as characterisation (post-processing), classification, clustering association rules or simply Apriori algorithm, regression evolution analysis [5] Pattern evaluation modulecommunicate with the DM-modules to discovered interesting pattern A data mining method usually follows a particular model/framework for designing their function approach. The Apriori algorithm, simple and easy to understand it may be, is limited. The main disadvantage is that it costs too much and takes more time on keeping many candidate sets with frequent item sets, low minimum support or large amount of items.

IV. RESULTS AND DISCUSSION

4.1 Results of Descriptive Statics of Study Variables

Figure 4.1 - GUI

Figure 5.2- Result



The deployed morbidity forecasting system produced competitive results across several machinelearning algorithms. This dataset was used to build 4 models — a Decision Tree model, Random Forest model, Naive Bayes Model and k-Nearest Neighbours. The scatter plots also facilitated a visual understanding of the distribution of symptoms for predicted diseases and help to interpret our results. This allows for the real-time predictions from various models on a user-friendly graphical interface(GUI) and great interaction with users. The reason to save user data as well as the predictions into an SQLite database was storing them for reviewing or future improvements of the system. The system acts as a great screening tool for initial disease prediction, and can provide even better results if trained on more data(i.e supplements) and minorly adjust the Machine learning models to increase automatically(HLC Grading up).

V. CONCLUSION AND FUTURE SCOPE

In the medical field our application is for the first time and we are introducing this innovative application Prognostic Disease Analytics. It transforms predictive and preventive healthcare with the combination of healthcare expertise blended well with rigorous data analytics techniques. Its main use as a user-friendly system that streamlines the input of patient symptoms and additional learning algorithms — Decision Trees, Random Forests, Naive Bayes and k-Nearest Neighbors. The algorithmic diversity protects the robustness of predictions, being applicable to datasets and disease profiles. In addition, users should explore the data by using basic capabilities that will help more in understanding which symptom is most closely related to a particular disease. Furthermore the use of a database allows traceability and post-processing, making the system useful in time. This application is an important asset to the field of prognostic disease analytics, by focusing on usability, algorithm diversity and data management. It has a ton of room to grow. misfortunes of healthcare expulsion through accurate predictions for disease and help subtleties-informed decisions by medical experts In summary, this abstract highlights the encouraging trend of model building for data mining to intuitive interfaces supporting Health Care Needs.

I. ACKNOWLEDGMENT

We would like to thank everyone who has supported me in the course of this study. I am grateful to my advisor Mr. Kannadasan R, our Head of Department Mrs. Umadevi Ks and our Dean of academics Mr. Ramesh Babu for their advice and knowledge in the disease outcome prediction field. I would also like to acknowledge colleagues and friends, who had taken the time to read my drafts and provide a lot of helpful critiques that allowed me to hone these thoughts.

We are also thankful to the Vellore Institute of Technology for providing its departmental resources and facilities. Finally, a huge thank you to my family and friends for never-ending support & understanding along the way.

REFERENCES

[1] - “Disease prediction system using support vector machine and multi-linear regression” - Vijayarani, S. and Dhayanand

[2] - "A comparative study on disease prediction using supervised machine learning algorithms” - Rahman, A.S., Shamrat, F.J.M., Tasnim, Z., Roy, J. and Hossain

[3] - “A hybrid intelligent system framework for heart disease prediction” - Shraddha Subhash Shirsath, Prof. Shubhangi Patil

[4] - “Disease prediction by machine learning” - Thomas H. Cormen, Charles E. Leiserson, and Ronald L. Rivest

[5] - “Improved algorithm for mining Apriori algorithm based on association rules” - Shabtay, L., Fournier-Viger, P., Yaari, R., & Dattne