Disease Prediction System

Sidrathul Munthaha PV

Date: 06-09-2023

Abstract:

The dependency on computer-based technology has resulted in storage of lot of electronic data in the health care industry. As a result of which, health professionals and doctors are dealing with demanding situations to research signs and symptoms correctly and perceive illnesses at an early stage. However, Machine Learning technology have been proven beneficial in giving an immeasurable platform in the medical field so that health care issues can be resolved effortlessly and expeditiously. Disease Prediction is a Machine Learning based system which primarily works according to the symptoms given by a user. The disease is predicted using algorithms and comparison of the datasets with the symptoms provided by the user.

1.0 Introduction

Machine Learning is the domain that uses past data for predicting. Machine Learning is the understanding of computer system under which the Machine Learning model learn from data and experience. The machinelearning algorithm has two phases: 1) Training & 2) Testing. To predict the disease from a patient's symptoms and from the history of the patient, machine learning technology is struggling from past decades. Healthcare issues can be solved efficiently by using Machine Learning Technology. We are applying complete machine learning concepts to keep the track of patient's health. ML model allows us to build models to get quickly cleaned and processed data and deliver results faster. By using this system, good treatment will be given to the patient, which increases improvement in patient healthcare services. To introduce machine learning in the medical field, healthcare is the prime example. To improve the accuracy of large data, the existing work will be done on unstructured or textual data. For the prediction of diseases, the existing will be done on linear, KNN, Decision Tree algorithm

1.1 Problem Statement

Health information needs are also changing the information seeking behavior and can be observed around the globe. Challenges faced by many people are looking online for health information regarding diseases, diagnosis and different treatments. If a recommendation system can be made for doctors and medicine while using review mining will save a lot of time. In this type of system, the user face problem in understanding the heterogeneous medical vocabulary as the users are laymen. User is confused because a large amount of medical information on different mediums are available.The idea behind recommender system is to adapt to cope with the special requirements of the health domain related with user

2.0 Customer Need Assessment

There has been a huge demand of services provided by the health care system in the past few years. The pandemic period clearly stated the necessity of online treatment . The increased demand of health care facilities asks a solution for big scale implementation of treatment system.

It can be solved by automating the process of disease diagnosis to an extend. With the growth of machine learning and artificial intelligence we can simply implement a machine learning model to automate the task of diagnosis thus introducing a big revolution in the field of health care. It will not only benefit the health care ventures but also the customers who uses this system as they can simply diagnose their disease by just entering their symptoms to the user interface of this system.The system will help the users to identify regular diseases as well as chronic illness according to the trained model.

3.0 Target Specification and characterization

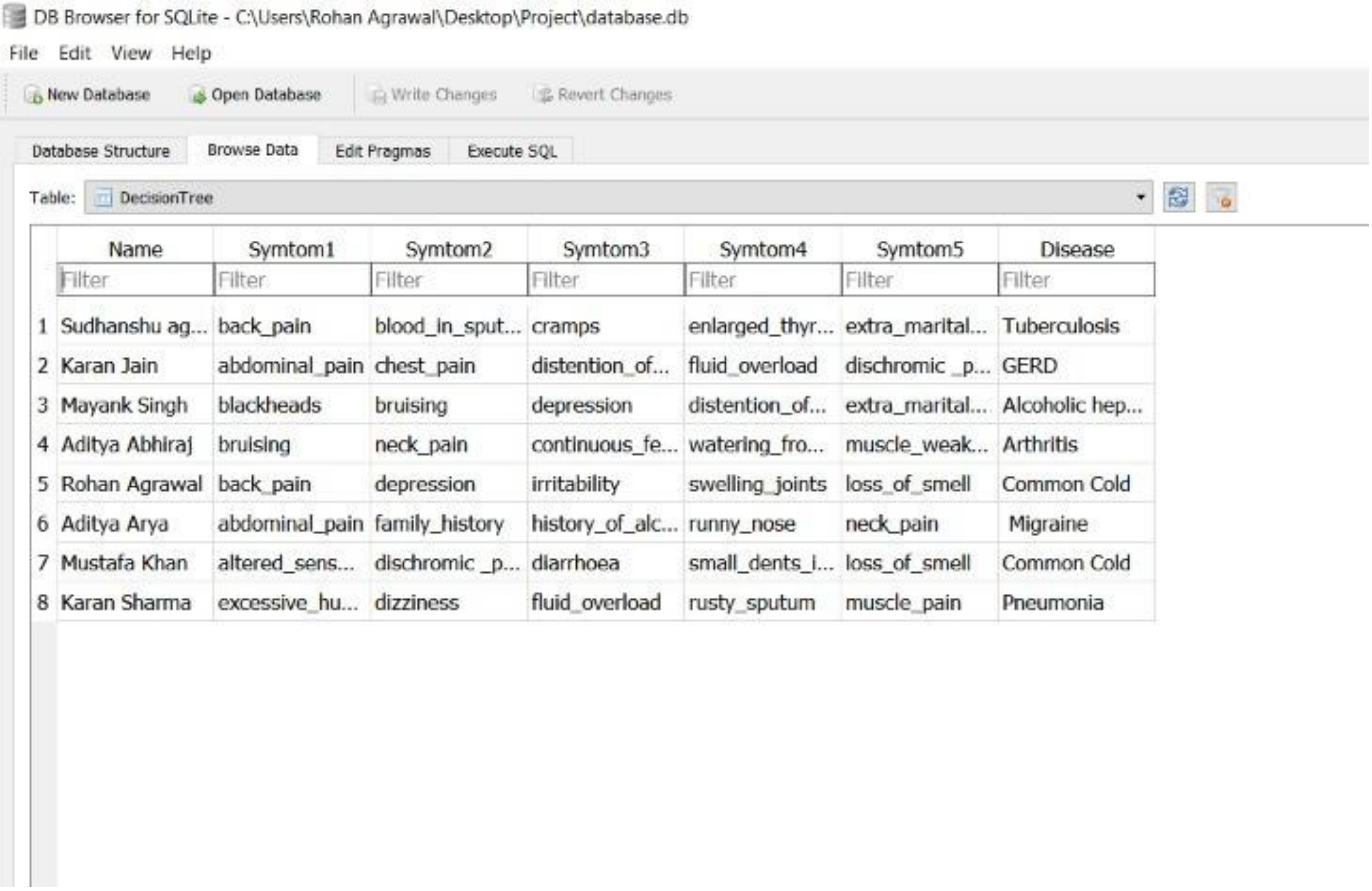
Predicting the possible disease according to the potential symptoms specified by the user.
Targeted customers for this system are common people who are in need of diagnosing a disease. So this is a system which will have a greater audience acceptance.

the system will take the user inputs and machine learning model will analyse the symptoms and predict the disease possible to have.

4.0 External Search(information sources)

The database used in this project is "sqlite" which consist of four tables in which we have shown the results of four different algorithms. We are saving the results of users with their names for future preferences..

:



	Name	Symtom1	Symtom2	Symtom3	Symtom4	Symtom5	Disease
1	Sudhanshu ag...	back_pain	blood_in_sput...	cramps	enlarged_thyr...	extra_marital...	Tuberculosis
2	Karan Jain	abdominal_pain	chest_pain	distention_of...	fluid_overload	dischromic _p...	GERD
3	Mayank Singh	blackheads	bruising	depression	distention_of...	extra_marital...	Alcoholic hep...
4	Aditya Abhiraj	bruising	neck_pain	continuous_fe...	watering_fro...	muscle_weak...	Arthritis
5	Rohan Agrawal	back_pain	depression	irritability	swelling_joints	loss_of_smell	Common Cold
6	Aditya Arya	abdominal_pain	family_history	history_of_alc...	runny_nose	neck_pain	Migraine
7	Mustafa Khan	altered_sens...	dischromic _p...	diarrhoea	small_dents_i...	loss_of_smell	Common Cold
8	Karan Sharma	excessive_hu...	dizziness	fluid_overload	rusty_sputum	muscle_pain	Pneumonia

5.0 Algorithms used

- Decision Tree
- Random Forest
- KNearestNeighbour
- Naive Bayes

5.1 Decision Tree

A decision tree is a structure that can be used to divide up a large collection of records into successfully smaller sets of records by applying a sequence of simple decision tree. With each successive division, the members of the resulting sets become more and more similar to each other. A decision tree model consists of a set of rules for dividing a large heterogeneous population into smaller, more homogeneous (mutually exclusive) groups with respect to a particular target. The target variable is usually categorical and the decision tree is used either to:

- Calculate the probability that a given record belong to each of the category and
- To classify the record by assigning it to the most likely class (or category). In this disease prediction system, decision tree divides the symptoms as per its category and reduces the dataset difficulty

5.2 K Nearest Neighbour (KNN)

KNN could be terribly easy, simple to grasp, versatile and one amongst the uppermost machine learning algorithms. In the Healthcare System, the user will predict the disease. In this system, the user can predict whether the disease will detect or not. In the proposed system, classifying disease in various classes that shows which disease will happen on the basis of symptoms. KNN rule used for each classification and regression issue. KNN algorithm is based on feature similarity approach. It is the best choice for addressing some of the classification related tasks. K-nearest neighbour classifier algorithm is to predict the target label of a new instance by defining the nearest neighbour class. The closest class will be identified using distance measures like Euclidean distance. If K = 1, then the case is just assigned to the category of its nearest neighbour. The value of ‘k’ has to be specified by the user and the best choice depends on the data. The larger value of ‘k’ reduces the noise on the classification. If the new feature i.e in our case symptom has to classify, then the distance is calculated and then the class of feature is selected which is nearest to the newer instance. In the instance of categorical variables, the Hamming distance must be used. It conjointly brings up the difficulty of standardization of the numerical variables between zero and one once there's a combination of numerical and categorical variables within the dataset

5.3. Naive Bayes

Naive Bayes is an easy however amazingly powerful rule for prognosticative modelling. The independence assumption that allows decomposing joint likelihood into a product of marginal likelihoods is called as 'naive'. This simplified Bayesian classifier is called as naive Bayes. The Naive Bayes classifier assumes the presence of a particular feature in a class is unrelated to the presence of any other feature. It is very easy to build and useful for large datasets. Naive Bayes is a supervised learning model. Bayes theorem provides some way of calculative posterior chance P(b|a) from P(b), P(a) and P(a|b). Look at the equation below:

P(b v a)= P(a v b)P(b)/P(a)

Above,

- P(b|a) is that the posterior chance of class (b,target) given predictor (a, attributes). P(b) is the prioriprobability of class.
- P(a|c)iis that chance that is that the chance of predictor given class.
- P(a) is the prioriprobability of predictor. In our system, Naïve Bayes decides which symptom is to put in classifier and which is not. Logistic regression could be a supervised learning classification algorithm accustomed to predict the chance of a target variable that is disease.

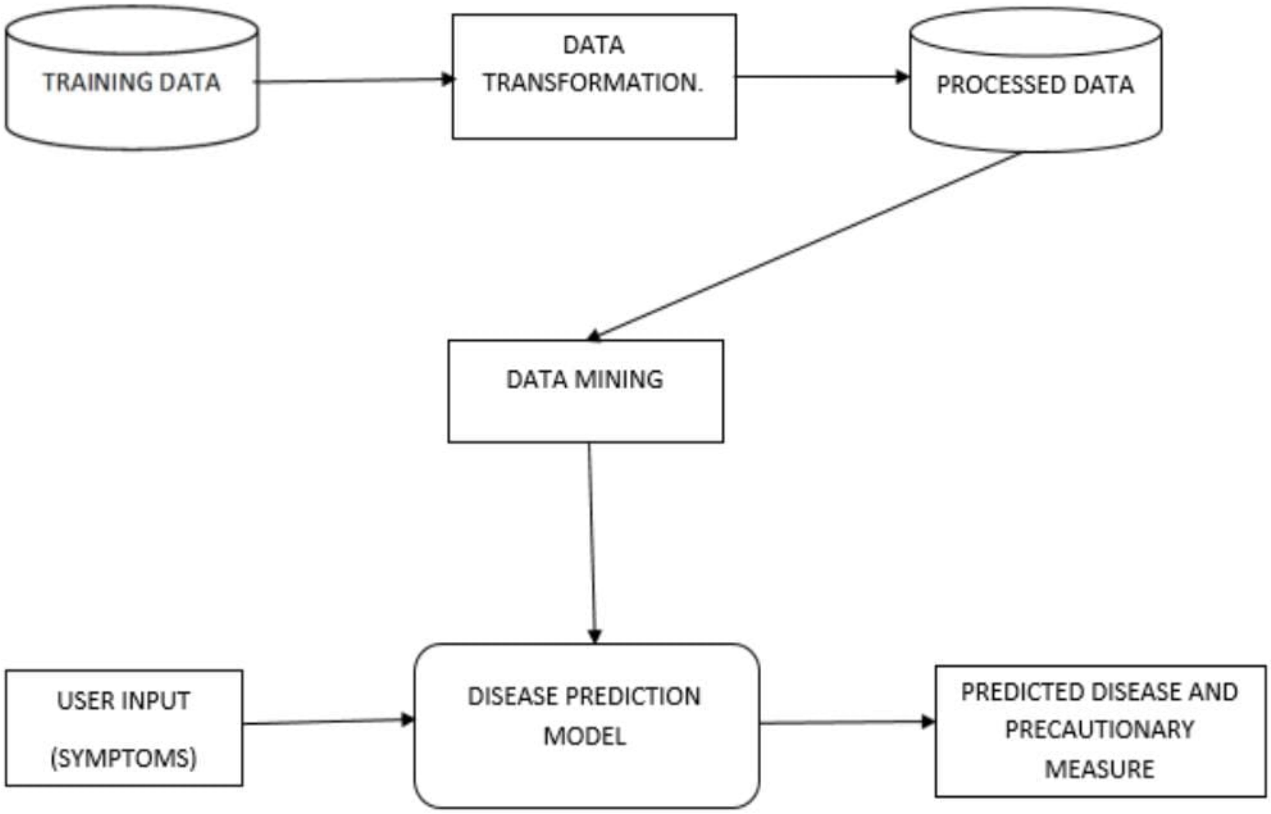
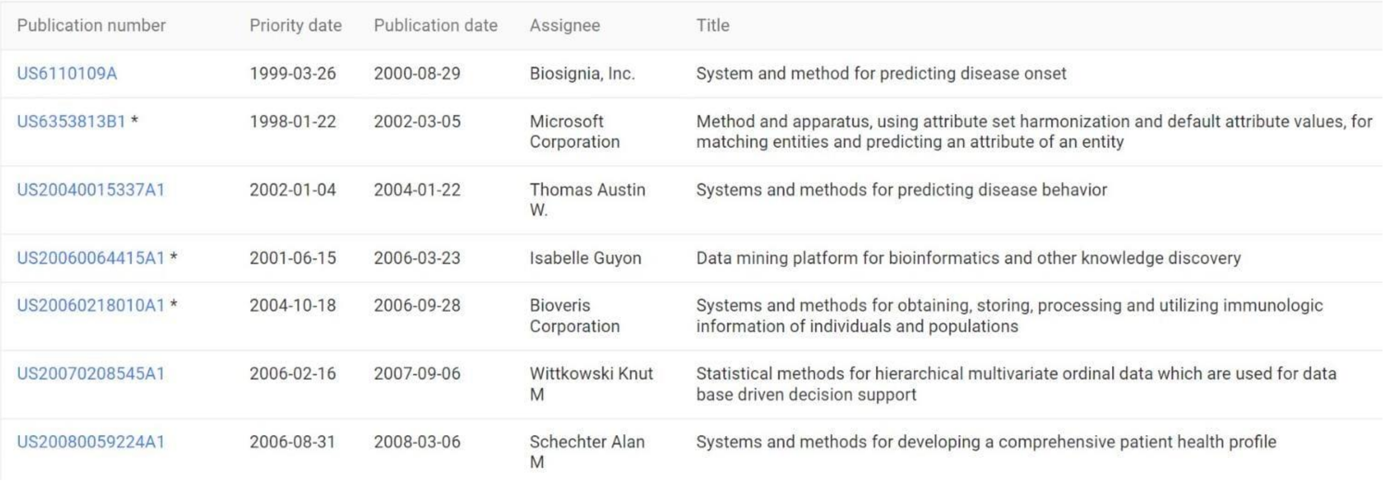


Fig: System architecture of disease prediction system

6.0 Applicable Patents



Publication number	Priority date	Publication date	Assignee	Title
US6110109A	1999-03-26	2000-08-29	Biosignia, Inc.	System and method for predicting disease onset
US6353813B1 *	1998-01-22	2002-03-05	Microsoft Corporation	Method and apparatus, using attribute set harmonization and default attribute values, for matching entities and predicting an attribute of an entity
US20040015337A1	2002-01-04	2004-01-22	Thomas Austin W.	Systems and methods for predicting disease behavior
US20060064415A1 *	2001-06-15	2006-03-23	Isabelle Guyon	Data mining platform for bioinformatics and other knowledge discovery
US20060218010A1 *	2004-10-18	2006-09-28	Bioveris Corporation	Systems and methods for obtaining, storing, processing and utilizing immunologic information of individuals and populations
US20070208545A1	2006-02-16	2007-09-06	Wittkowski Knut M	Statistical methods for hierarchical multivariate ordinal data which are used for data base driven decision support
US20080059224A1	2006-08-31	2008-03-06	Schechter Alan M	Systems and methods for developing a comprehensive patient health profile

7.0 Applicable Constraints:

1. The use of GUI or web app for the user interface.
2. Building database of diseases and symptoms.
3. For Evaluation of the model is done using data visualisation.
4. For modelling KNN, Decision tree, Random Forest and Naïve bayes is applied.

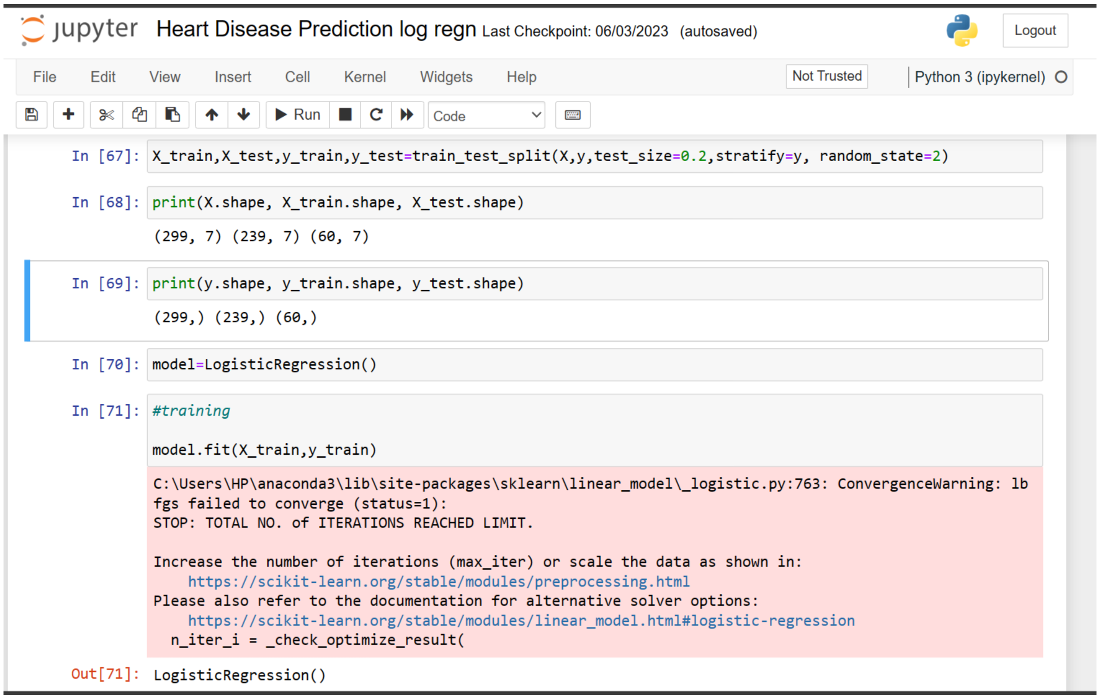
8.0 Business Opportunity

Health care system can utilise the system and make huge profit by automating the tasks which allows the services to implement on a large scale basis in a reduced cost margin. This systematic review aims to determine the performance, limitations, and future use of Software in health care. Findings may help inform future developers of Disease Predictability Software and promote personalized patient care.

.

9.0 Concept Generation

This product needs EDA and machine learning modelling to build a predictive system. For this purpose it uses four different algorithms. It also needs front end development besides machine learning part.



```
In [67]: X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.2,stratify=y, random_state=2)

In [68]: print(X.shape, X_train.shape, X_test.shape)

         (299, 7) (239, 7) (60, 7)

In [69]: print(y.shape, y_train.shape, y_test.shape)

         (299,) (239,) (60,)

In [70]: model=LogisticRegression()

In [71]: #training

         model.fit(X_train,y_train)

         C:\Users\HP\anaconda3\lib\site-packages\sklearn\linear_model\_logistic.py:763: ConvergenceWarning: lb
         fgs failed to converge (status=1):
         STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

         Increase the number of iterations (max_iter) or scale the data as shown in:
             https://scikit-learn.org/stable/modules/preprocessing.html
         Please also refer to the documentation for alternative solver options:
             https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
           n_iter_i = _check_optimize_result(

Out[71]: LogisticRegression()
```

Fig: sample code snippet from heart disease prediction using logistic regression

10.0 Final Report Prototype

The product takes the following functions to perfect and provide a good result.

Back-end

Identification of machine learning algorithms should be done so as to best fit the data into the model and to extract the desired target prediction.

1. Performing EDA to realize the dependent and independent features.
2. Algorithm training and testing is done so as to obtain accurate prediction result.

Front End

1. User interface should take the input from the user and must be able to guide the user through the system.

2.It can either be implemented as a web page or as a mobile app or just a GUI will also work.

11.0 Product details - How does it work?

- An interactive user interface will take inputs from the patient about the symptoms they have.
- Then it will check for the similar inputs in the database which is trained for the model.

- It will analyse the various matching diseases in the database for the particular symptoms.
- Using the machine learning algorithm it will predict the disease .
- The prediction result will then pop up on the user interface.

12. References/Source of Information

https://patents.google.com/patent/US8504343B2/en#title

https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset?select=heart.csv

https://www.geeksforgeeks.org/disease-prediction-using-machine-learning/