Smart Disease Prediction Using Machine Learning

Nivethitha. A
Department of Computer Science Engineering
Sri Ramakrishna Engineering College
Coimbatore, India

Pramoth Krishnan. T
Department of Computer Science Engineering
Sri Ramakrishna Engineering College
Coimbatore, India

Narendran. G
Assistant Professor
Department of Computer Science Engineering
Sri Ramakrishna Engineering College
Coimbatore, India

Abstract:- Disease Prediction exploitation Machine Learning may be a system that predicts the illness-supported data or the symptoms he/she enters into the system and provides the prescription for that illness and conjointly provides the correct results supported that information. It's a system that provides the user the ideas and tricks to keep up the health system of the user and it provides how to seek out the illness exploitation this prediction. From this technique by simply asking the symptoms from the user and getting into within the system and in precisely few seconds they'll tell the precise and up to some extent the correct diseases. This illness Prediction exploitation Machine Learning has totally finished the assistance of Machine Learning and Python artificial language with Tkinter Interface for it and conjointly exploitation the dataset that's obtainable antecedent by the hospital's exploitation that we are going to predict the illness. The info entered by the user holds on within the information.

Keywords:- *Machine Learning, Symptoms based disease prediction, Python, Decision Tree, Random Forest, KNN, Naïve Bayes.*

I. INTRODUCTION

With the increase in the number of patients and diseases per annum medical system is overloaded and with time became overpriced in many countries. Most of the disease involves a consultation with doctors to urge treatment. With sufficient data prediction of disease by an algorithm are often very easy and cheap. Prediction of disease by watching the symptoms is an integral part of treatment. In our project, we've tried to accurately predict a disease by watching the symptoms of the patient. The 4 different algorithms for this purpose and gained an accuracy of 92-95%. Such a system can have a really large potential in medical treatment of the longer term. An intelligently interface to encourage interaction with the framework. We've additionally tried to signify and visualized the results of our study and this project. Currently, a day's doctors square measure adopting several scientific technologies and methodology for each identification and identification not solely common sickness, however additionally several fatal diseases. The prosperous treatment is sometimes attributed to right and correct identification. The project disease prediction using machine learning is developed to beat general disease in earlier stages as we all know in the competitive environment of economic development the mankind has involved such a lot that he/she isn't concerned about health consistent with research there are 40% peoples how ignores about the general disease which results in harmful disease later. the most reason for ignorance is laziness to consult a doctor and time concerns the peoples have involved themselves such a lot that they need no time to require a meeting and consult the doctor which later results in a fatal disease. consistent with research there are 70% peoples in India suffers from general disease and 25% of peoples face death thanks to early ignorance the most motive to develop this project is that a user can sit at their convenient place and have a check-up of their health the UI is meant in such an easy way that everybody can easily operate it and may have a check-up.

II. RELATED WORK

Many researchers have used machine learning techniques like KNN, Naïve Bayes and Decision trees to develop disease Prediction stratergies. Sathyabama Balasubramanian, Balaji Subramani discussed the system to reduces the multiple diseases showing the similar symptoms problem and it will increase the accuracy of such diagnosis. It has received 71.53% accuracy. Aditya Arya, Sudhanshu, Rohan Agarwal, attempted to show and visualized the result of our study and this project. By comparing with other techniques it scores accuracy of 68.5%.Iqra anjum, Mohammed Afreed, Mohammed Kalam has developed a system which predicts the disease based on the information or the symptoms he/she enter into the system and provides the accurate results based on that information.

Raj H. Chauhan, Daksh N. Naik, Rinal A. Halpati, Sagarkumar J. Patel, Mr. A.D.Prajapati developed a system analyzes the symptoms provided by the user as input and gives the probability of the disease as an output Disease Prediction is done by implementing the Decision tree Classifier. Decision tree Classifier calculates the probability of the disease. With big data growth in biomedical and health care communities, accurate analysis of medical data benefits early disease detection, patient care.

III. EXISTING SYSTEM

Since the arrival of advanced computing, the doctors still needs the technology in numerous possible ways that like surgical illustration method and X-ray photography, however the technology perceptually stayed behind. The strategy still needs the doctor's data and experience because of different factors ranging from medical records to weather, atmosphere, pressure level and various different factors. The massive numbers of variables are granted as entire variables that are needed to grasp the whole operating method itself, still, no model has analyzed with success. To tackle this downside, Medical decision support systems should be used. This technique will assist the doctors to form the right decision. We tend to are applying machine learning to maintained complete hospital knowledge Machine learning technology that permits building models to get quickly analyze knowledge and deliver results quicker, with the utilization of machine learning technology doctors will create an enormous decision for patient diagnoses and treatment selections, that results in improvement of patient care services. When doing the analysis and comparison of all the algorithms and theorems of machine learning we've return to conclusion that everyone those algorithms like decision Tree, KNN, Naïve Bayes, Regression and Random Forest algorithm all are necessary in building a sickness prediction system that predicts the disease of the patients from that he/she is suffering from and to try to do this we've used some performance measures like ROC, KAPPA Statistics, RMSE, MEA and numerous tools. When exploitation numerous techniques like neural networks to form predictions of the diseases and when doing that we tend to return to conclusion that it will predict up to 90% accuracy rate when doing the experimentation and confirmatory the results. Existing system will predict the disease however not the sub kind of the sickness and it fails to predict the condition of the folks, the predictions of disease are indefinite and non-specific.

IV. PROPOSED SYSTEM

The proposed system of disease prediction using machine learning is that we've got used many techniques and algorithms and every one other various tools to make a system which predicts the disease of the patient using the symptoms and by taking those symptoms were comparing with the system's dataset that's previously available and it gives prescription for that disease predicted by those algorithms. By taking those datasets and comparing with the patient's disease we'll predict the accurate percentage disease of the patient. The dataset and symptoms visit the prediction model of the system where the information is pre-processed for the future references so the feature selection is finished by the user where he will enter the assorted symptoms. Then the classification of these data is completed with the help of assorted algorithms and techniques like Decision Tree, KNN, Naïve Bayes, Random Forest etc. Then the information goes within the recommendation model, there it shows the risk analysis that's involved within the system and it also provides the probability estimation of the system such it shows the assorted probability like how the system behaves when there are n number of predictions are done and it also does the recommendations for the patients from their final result and also from their symptoms prefer it can show what to use and what to not use from the given datasets and therefore the final results. Here we've got combined the structure and unstructured form to data for the general risk analysis that's required for doing the prediction of the disease. Using the structured analysis, we will identify the chronic kinds of disease in an exceedingly particular region and particular community. In unstructured analysis we select the features automatically with the help of algorithms and techniques. This technique takes symptoms from the user and predicts the disease accordingly supported the symptoms that it takes and also from the previous datasets, it also helps in continuous evaluation of viral diseases, heart rate, blood pressure, sugar level and far more which is within the system and along with other external symptoms its predicts the acceptable and accurate disease and it gives the prescription details for that disease. And also the data entered by the user are stored within the created database.

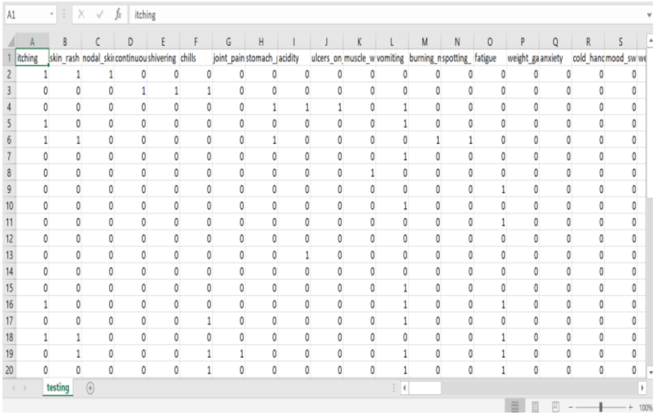
V. MODULE DISCRIPTION

The overall proposed system is classified into five modules.

- Collection of Clinical Data
- Data Pre-processing
- Model Building
- Model Building using Prescription
- Database Creation

A. *Collection of Clinical Data*

This dataset could be a information database of disease-symptom associations generated by an automatic methodology supported data in textual discharge summaries of patients at new york presbyterian Hospital shown in Fig.1. The 1st column shows the disease, the second the amount of discharge summaries containing a positive and current mention of the sickness, and therefore the associated symptom. Associations for the a hundred and fifty most frequent diseases supported these notes were computed and therefore the symptoms are shown hierarchal supported the strength of association.

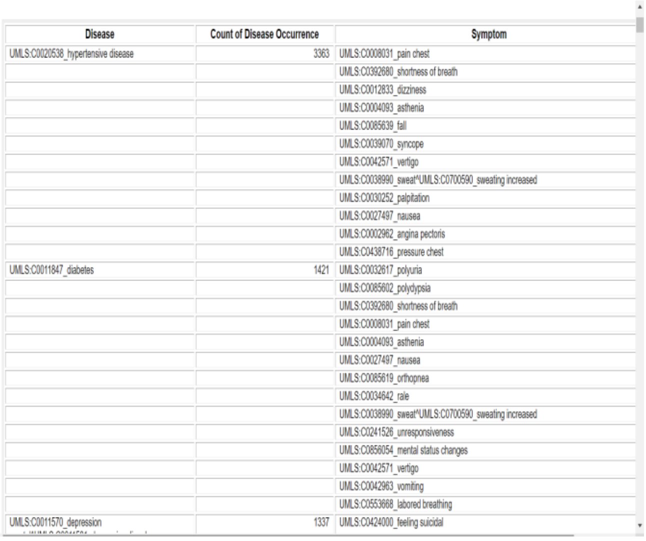


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	
1	itching	skin_rash	nodal_skir	continuou	shivering	chills	joint_pain	stomach_	acidity	ulcers_on	muscle_w	vomiting	burning_n	spotting_	fatigue	weight_ga	anxiety	cold_hanc	mood_sw	we
2	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	0	0	0	0	0	0	0	1	1	1	0	1	0	0	0	0	0	0	0	
5	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	
6	1	1	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	
7	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	
8	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	
10	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
13	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
15	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	
16	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	
17	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	
18	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	
19	0	1	0	0	0	1	1	0	0	0	0	1	0	0	1	0	0	0	0	
20	0	0	0	0	0	1	0	0	0	0	0	1	0	0	1	0	0	0	0	

Fig.1 Clinical Data

B. Data- Preprocessing

As, the information pre-processing is a vital step in machine learning; we have a tendency to, thus removed all those variables that contained over 50% missing price. The strategy used the MedLEE natural language processing system to get UMLS codes for diseases and symptoms from the notes Fig.2.Then applied mathematics ways supported frequencies and co-occurrences were used to acquire the associations.



Disease	Count of Disease Occurrence	Symptom
UMLS:C0020538_hypertensive disease	3363	UMLS:C0008031_pain chest
		UMLS:C0392680_shortness of breath
		UMLS:C0012833_dizziness
		UMLS:C0004093_asthenia
		UMLS:C0085639_fall
		UMLS:C0039070_syncope
		UMLS:C0042571_vertigo
		UMLS:C0038990_sweat^UMLS:C0700590_sweating increased
		UMLS:C0030252_palpitation
		UMLS:C0027497_nausea
		UMLS:C0002962_angina pectoris
		UMLS:C0438716_pressure chest
UMLS:C0011847_diabetes	1421	UMLS:C0032617_polyuria
		UMLS:C0085602_polydypsia
		UMLS:C0392680_shortness of breath
		UMLS:C0008031_pain chest
		UMLS:C0004093_asthenia
		UMLS:C0027497_nausea
		UMLS:C0085619_orthopnea
		UMLS:C0034642_rale
		UMLS:C0038990_sweat^UMLS:C0700590_sweating increased
		UMLS:C0241526_unresponsiveness
		UMLS:C0856054_mental status changes
		UMLS:C0042571_vertigo
		UMLS:C0042963_vomiting
		UMLS:C0553668_labored breathing
UMLS:C0011570_depression	1337	UMLS:C0424000_feeling suicidal

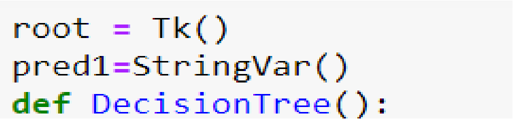
Fig.2 UMLS Codes

C. Model Building

The predictive classifier models were developed for accurately identify Disease given by the user. The classification model for predict the Disease is Random Forest (RF), Decision Tree, K Nearest Neighbour, Naïve Bayes Algorithm.

1) Decision Tree

Decision tree induction is the learning of decision trees from class-labelled training tuples. A decision tree is a flowchart-like tree structure, Decision tree algorithms are quite strong to the presence of noise, particularly once strategies for avoiding overfitting.



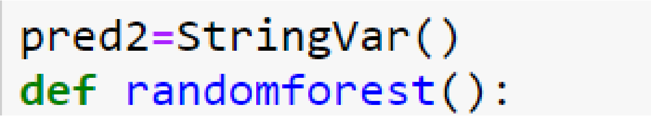
```
root = Tk()
pred1=StringVar()
def DecisionTree():
```

Fig.3 Decision Tree Funtion

DecisionTreeClassifier() is used to train the model and predict the disease on testing dataset according to symptoms entered by the user. Final disease for decision tree is stored in a variable named “pred1” shown in Fig.3. Accuracy of predicting the disease is printed using accuracy score and confusion matrix is created using confusion matrix which are imported from sklearn metrices.

2) Random Forest

Random forest (RF) comes under the ensemble classification algorithm which is composed of a large number of decision trees. The algorithm can handle thousands of input attributes without variable deletion. Accuracy and variable importance data will be supplied with the results. A random forest is that the classifier consisting of a group of tree structured classifiers k, wherever the k is severally, identically distributed random trees and every random tree incorporates the unit of vote for classification of input.



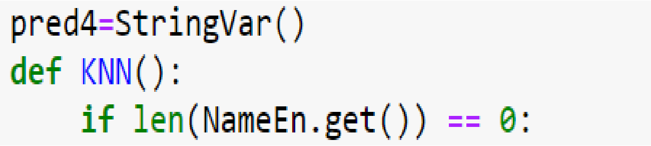
```
pred2=StringVar()
def randomforest():
```

Fig.4 Random Forest Function

Definition of randomforest() function. “pred2” is used to store the predicted disease using random forest algorithm shown in Fig.4. RandomForestClassifier() is used to train the model and predict the disease on testing dataset according to symptoms entered by the user. Final disease for random forest is stored in a variable named “pred2”. Accuracy of predicting the disease is calculated using accuracy_score and confusion matrix is created using confusion_matrix which are imported from sklearn.metrices.

3) K-nearest Neighbour

K-Nearest Neighbour is one amongst the best Machine Learning algorithms supported supervised Learning technique. K-NN rule stores all the available knowledge and classifies a replacement information supported the similarity. this suggests once new knowledge seems then it may be simply classified into a well suite category by exploitation K-NN algorithmic rule. K-NN rule may be used for Regression also as for Classification however principally it's used for the Classification issues. It works by finding a pattern in knowledge that links knowledge to results and it improves upon the pattern recognition with each iteration.



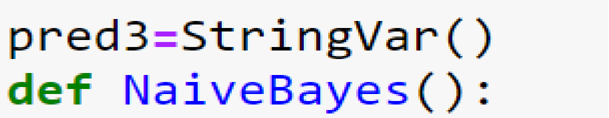
```
pred4=StringVar()
def KNN():
    if len(NameEn.get()) == 0:
```

Fig.5 KNN funtion

Definition of KNN() function. “pred4” is used to store the predicted disease using kNearestNeighbour algorithm shown in Fig.4.

4) Naïve Bayes Algorithm

NaiveBayes is used to predict the categorical class labels. It classifies the category information supported the training set and also the values during a classifying attribute and uses it in classifying new data. It could be a two-step process model Construction and Model Usage. This Bayes theorem is called when Thomas Bayes and it's technique method for classification and supervised learning method. It will solve both categorical and continuous values attributes.



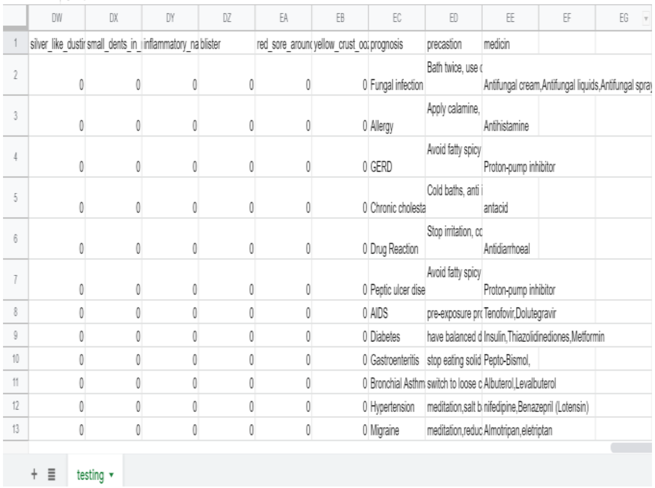
```
pred3=StringVar()
def NaiveBayes():
```

Fig.6 Naïve Bayes Function

Definition of NaiveBayes() function. “pred3” is used to store the predicted disease using Naïve Bayes algorithm shown in Fig.6.

D. Model Building Using Medicine and Prescription

Medicine Prescription is one of the most important thing in everyone’s life. Collecting the medicine and precaution dataset from textual discharge summaries of patients at New York Presbyterian Hospital shown in Fig.7. When disease is predicted medicine and precaution for that disease is displayed in GUI page.

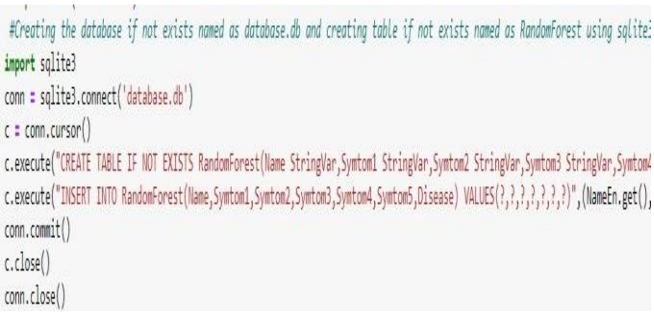


	DW	DX	DY	DZ	EA	EB	EC	ED	EE	EF	EG
1	silver_like_dustir	small_dents_in_	inflammatory_na	blister	red_sore_arounc	yellow_crust_oo	prognosis	precaution	medicin		
2	0	0	0	0	0	0	Fungal infection	Bath twice, use c	Antifungal cream,Antifungal liquids,Antifungal spray		
3	0	0	0	0	0	0	Allergy	Apply calamine,	Antihistamine		
4	0	0	0	0	0	0	GERD	Avoid fatty spicy	Proton-pump inhibitor		
5	0	0	0	0	0	0	Chronic cholesta	Cold baths, anti i	antacid		
6	0	0	0	0	0	0	Drug Reaction	Stop irritation, cc	Antidiarrhoeal		
7	0	0	0	0	0	0	Peptic ulcer dise	Avoid fatty spicy	Proton-pump inhibitor		
8	0	0	0	0	0	0	AIDS	pre-exposure prc	Tenofovir,Dolutegravir		
9	0	0	0	0	0	0	Diabetes	have balanced d	Insulin,Thiazolidinediones,Metformin		
10	0	0	0	0	0	0	Gastroenteritis	stop eating solid	Pepto-Bismol,		
11	0	0	0	0	0	0	Bronchial Asthm	switch to loose c	Albuterol,Levalbuterol		
12	0	0	0	0	0	0	Hypertension	meditation,salt b	nifedipine,Benazepril (Lotensin)		
13	0	0	0	0	0	0	Migraine	meditation,reduc	Almotripan,eletriptan		

Fig.7 Clinical Dataset for Medicine and prescription

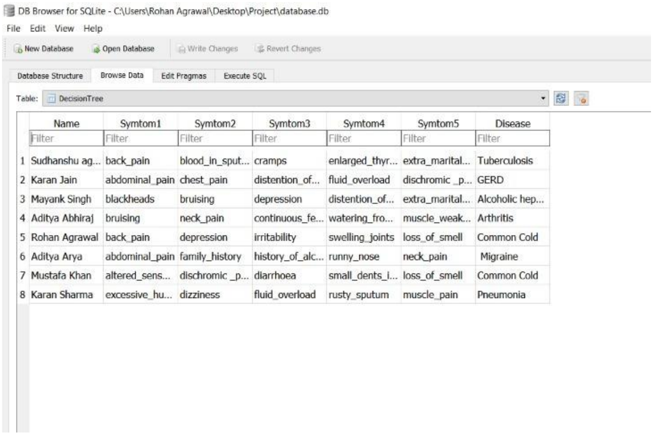
1) Database Creation

The database is created by using SQLite to store the details entered by the user in the GUI page shown in Fig.9.



```
#Creating the database if not exists named as database.db and creating table if not exists named as RandomForest using sqlite3
import sqlite3
conn = sqlite3.connect('database.db')
c = conn.cursor()
c.execute("CREATE TABLE IF NOT EXISTS RandomForest(Name StringVar,Symtom1 StringVar,Symtom2 StringVar,Symtom3 StringVar,Symtom4
c.execute("INSERT INTO RandomForest(Name,Symtom1,Symtom2,Symtom3,Symtom4,Symtom5,Disease) VALUES(?,?,?,?,?,?,?)",(NameEn.get(),
conn.commit()
c.close()
conn.close()
```

Fig.8 Database Code



DB Browser for SQLite - C:\Users\Rohan Agrawal\Desktop\Project\database.db

Table: DecisionTree

	Name	Symtom1	Symtom2	Symtom3	Symtom4	Symtom5	Disease
	Filter	Filter	Filter	Filter	Filter	Filter	Filter
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

Fig.9 Database Storage

VI. RESULTS

The feature extracted data is further evaluated to predict the disease by using the classifiers such as Random forest and Naïve Bayes Classifiers, Decision tree, K-Nearest Neighbour. On comparing the four machine learning algorithms, K-Nearest Neighbour shows 95.6%, Naïve Bayes Classifiers shows 94.5%, Random Forest model has the highest accuracy of 95.7% than the Decision Tree algorithm shows 92.4%. The results were evaluated with accuracy, sensitivity, specificity, positive predictive value and negative predictive value. The final accuracy of each model was shown in Fig. 10.

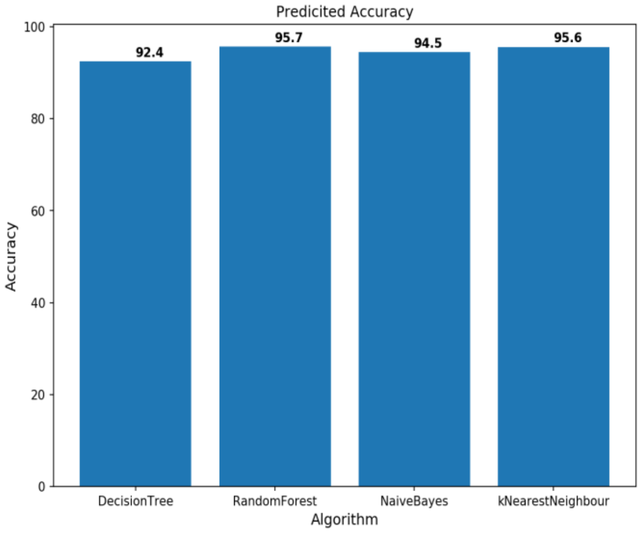


Fig.10 Final Accuracy of each model

VII. CONCLUSION

To conclude, this project disease prediction victimization machine learning is extremely a lot of helpful in everyone’s day to day life and it's in the main additional vital for the health care sector, as a result of they're the one that daily uses these systems to predict the diseases of the patients supported their general data and there symptoms and provides the drugs and precaution for that disease that they're been through. Our system is useful to those folks that are continuously worrying concerning their health and that they got to understand what happens with their body. Our main shibboleth to develop this method is to grasp them for his or her health. Especially, folks that are littered with psychological state like depression, anxiety. they will start up of those issues and might live their daily lives simply. On a median we tend to achieved accuracy of ~95%. Such a system will be mostly reliable to try to the task. making this method we tend to additionally side some way to store the information entered by the user within the information which may be utilized in future to assist in making higher version of such system. Our system additionally has a simple to use interface. It additionally has numerous visual illustration of information collected and results achieved.

REFRENCES

[1]. Pingale, Kedar, et al. "Disease Prediction using Machine Learning." (2019).Mr. Chala Beyene, Prof. Pooja Kamat, "Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques", International Journal of Pure and Applied Mathematics, 2018.

[2]. Pingale, K., Surwase, S., Kulkarni, V., Sarage, S., & Karve, A. (2019). Disease Prediction using Machine Learning.

[3]. Aiyesha Sadiya, Differential Diagnosis of Tuberculosis and Pneumonia using Machine Learning(2019).

[4]. S. Patel and H. Patel, "Survey of data mining techniques used in healthcare domain," Int. J. of Inform. Sci. and Tech., Vol. 6, pp. 53-60, March, 2016.

[5]. Balasubramanian, Satyabhama, and Balaji Subramanian. "Symptom based disease prediction in medical system by using Kmeans algorithm." International Journal of Advances in Computer Science and Technology 3.

[6]. Dhenakaran, K. Rajalakshmi Dr SS. "Analysis of Data mining Prediction Techniques in Healthcare Management System." International Journal of Advanced Research in Computer Science and Software Engineering 5.4 (2015)

[7]. Jin Ma, Sung Chan Park, Jung Hun Shin, Nam Gyu Kim, Jerry H. Seo, Jong Suk Ruth Lee, Jeong Hwan Sa. "AI based intelligent system on the EDISON platform", Proceedings of the 2018 Artificial Intelligence and Cloud Computing Conference on ZZZ - AICCC '18, 2018

[8]. W. Yin and H. Schutze, "Convolutional neural network for paraphrase identification", *Proc. HLT-NAACL*, pp. 901-911, 2015.

[9]. Shadab Adam et al., "Prediction system for Heart Disease using Naïve Bayes", *International Journal of advanced Computer and Mathematical Sciences*, vol. 3, no. 3, pp. 290-294, 2012, ISSN 2230-9624.

[10]. J.R. Qulan, "Induction of Decision Trees", *Mach.Learn*, vol. 1, no. 1, pp. 81-10, Mar. 1986.

[11]. Sayantan Saha, Argha Roy Chowdhuri et al., "Web Based Disease Detection System", *IJERT*, vol. 2, no. 4, April 2013, ISSN 2278-0181.

[12]. Palli Suryachandra and Venkata Subba Reddy, "Comparison of Machine Learning algorithms For Breast Cancer".

[13]. Andrew Alikberov, Stephan Broadly et al., "The Learning Machine", [online] Available: https://www.thelearningmachine.ai.

[14]. M. Chen, Y. Hao, K. Hwang, L. Wang and L. Wang, "Disease prediction by machine learning over big data from healthcare communities", *IEEE Access*, vol. 5, no. 1, pp. 8869-8879, 2017.

[15]. *Disease and symptoms Dataset*, [online] Available: www.github.com.