A Methodology for Applying Machine Learning Algorithms in the Medical Industry

Bhavana Shah[1], Prof. Hemali Shah[2]

[1] *Student (M.E), Computer Engineering, Sal Institute of Technology & Engineering Research, Gujarat, India*

[2] *Assistant Professor (M.E.), Computer Engineering, Sal Institute of Technology & Engineering Research, Gujarat, India*

ABSTRACT

In our day-to-day life, we use lots of machine learning (ML) techniques and applications for farming, Medical care, Products Recommendations, stock marketing, Social Media (Facebook, LinkedIn), Traffic flow Alerts (Maps), Transportation, and Commuting (Uber, OLA), etc. Machine learning is a type of learning in which the machine learns by itself without explicitly programmed it. This is the type of application of artificial intelligence that provides the system with the facility so that they can spontaneously learn and develop from their understanding. This research paper discusses about the potential of applying machine learning skills in the medical sector. ML is organized in mainly four learning forms. Supervised learning contains labeled information when unsupervised learning contains unlabeled data. Semi-supervised learning is a combination of supervised and unsupervised learning. Reinforcement learning is a type of learning method that works together with its environment by generating actions and at the same stage determining errors and rewards. Trial & error search and delayed reward are all the most relevant features of reinforcement learning. ML is utilized in the healthcare sector like robotic surgery, Health Imaging Analysis, Sharing Patient Data, Drug Discovery, and Medical Imaging Diagnosis. Here we are studying in brief with several techniques and checking which algorithm is more accurate with less time consumption. This research paper summarizes some machine learning techniques such as K-nearest neighbor, support vector machine, random forest, a decision tree for disease prediction and disease detection. This work supports the dropping research gap between machine learning and the medical sector.

Keyword: - *Machine Learning, Healthcare Prediction, KNN, SVM, RF, DT*

1. INTRODUCTION

Machine learning (ML) applications are applied in everyday actions, for instance searching, advertisements', YouTube, medical care, banking segment [1]. By providing the solution of dipping the increasing cost, ML offers a better doctor-patient relation [2]. For the past decade, numerous healthcare centers are ongoing accepting a patient records system that holds patient files like patient schedules, treatments, and out-patient flow. The big data fetched new prospect-related ML methods, the data has facts related to a medicinal organization like a location, assets, schedule, patient's flow, and patient's data. Patient's schedules cover four units: appointments, difficulties, patients, and resources. Set of appointments that involve the appointment of the patient to reserve is called a schedule. A source may be a doctor, equipment like X-ray, or more [4]. The Machine learning algorithms are obliging in remedial application to distinguish compound patterns in huge data. It is applied in several disease observations and detection. It will build complex judgments about treatment policies for patients by recommendations of effecting beneficial healthcare system [3]. The medical segment deal with storage, restoration, optimum usage of medical

records and deliver awareness for problem-solving plus decision making [1]. We have a unique device Electronic Health Record, it comprises evidence about health seekers medical condition, laboratory test reports, patient history, treatment report such as rays, and scan report which is securely communal among other departments corresponding laboratories, pharmacies, specialists. Medicinal resources comprehending medical records with standardized terms to resolve the conflict problem. Medical records consist of medical conception. It is composed of data and indexed when queries posted by health seekers-related medical records are retrieved. The above-cited approaches are cast-off to give immediate and reliable responses for health seekers which are executed using ML [5]. This introduction of ML in healthcare has greatly saved medical resources and providing a new approach for people to see a doctor and help people's lives. At a similar time, the demand from people also offers inspiration for the research and development of Machine Learning, stimulating its constant enhancement [6]. Local mining, as well as Global learning, can be achieved using a machine learning methodology. Local mining can be responsible for giving solutions to health seekers by keeping personal medical records while global learning learns collaboratively and provides responses to patients [5].

2. ML TECHNIQUES

ML is divided into four types: Supervised Learning deals with learning function from existing training data. This kind of learning is utilized when historical data can be recycled to predict actions in the future [2], Unsupervised Learning: it workings well on transactional data. The methods that embrace self-organizing maps, nearest-neighbor mapping, k-means clustering, and singular value decomposition are most popular. These algorithms are utilized to segment text topics, propose objects [1], Reinforcement learning is accessed by the computer program for accomplished a particular objective in the dynamic environment. Feedback in terms of rewards and penalties is delivered to the program for the reason that it directs its disadvantage [4], Semi-supervised Learning is the combination of Labelled and unlabeled data which are useful for training. In Semi-supervised learning, only a subset of the training data is labelled in the system [7].

3. LITERATURE REVIEW

3.1 Machine Learning Techniques for Several disease: In this portion of review we are study about diverse ML techniques for different diagnosis.

Table 3.1 ML Classification Technique for Several disease

Author	Title	Disease	Data set	Technique	Accuracy	Reference
Bala Brahmeswara Kadaru*, B. Raja Srinivasa Reddy	A novel ensemble decision tree classifier using hybrid feature selection measures for Parkinson's disease prediction	Parkinson's disease and Alzheimer's disease	UCI	Ensemble decision tree model • andom tree + Naïve Bayes tree+ Improved NN based random forest	96.7% (0.967)	8
A. K. M. Sazzadur Rahman, Md. Mehedi Hasan, Md. Asaduzzaman, Syed Akhter Hossain	An analysis of computational intelligence techniques for diabetes prediction	Diabetes disease	National Institute of Diabetes and Digestive and Kidney Diseases 768 (80%-20%)	Artificial Neural Network (ANN) + Random Forest (RF) + Naive Bayes (NB) + Support Vector Machine (SVM)	76%	9

Filippo Amato , Alberto López, Eladia María Peña-Méndez , Petr Vaňhara, Aleš Hampl , Josef Havel	Artificial neural networks in medical diagnosis	Diabetes (420 patients) (320trainig -100 testing) Cancer Cardiovascular diseases	laboratory and instrumental data	ANN	Vary according to diseases		10
Hongxun Wu, Zhaohong Deng, Bingjie Zhang, Qianyun Liu, Junyong Chen	Classifier Model Based on Machine Learning Algorithms: Application to Differential Diagnosis of Suspicious Thyroid Nodules via Sonography	thyroid	970	Radial basis function–neural network	88.66%		11
Ji-Won Baek & Kyungyong Chung	CNN-based health model using knowledge mining of influencing factors	Chronic diseases • besity • iabetes (70%-30%) • igh blood pressure	Korea National Health and Nutrition Examination Survey	convolutional neural network (RMSE)	Laye r	Accuracy	12
					1	67.8	
					2	89.1	
					3	88.4	
					4	55.3	
C. VENKATESAN, P. KARTHIGAIKUMAR, ANAND PAUL, S. SATHEESKUMARA	ECG Signal Preprocessing and SVM Classifier-Based Abnormality Detection in Remote Healthcare	Heart related diseases	MIT-BIH database	ECG Signal Preprocessing and SVM Classifier	96 %		13

N, AND R. KUMAR	Applications					
Sellappan Palaniappan, Rafiah Awang	Intelligent Heart Disease Prediction System Using Data Mining Techniques	Heart Disease	909	Decision Tree Naïve Bayes Neural Network	NB:86.12 %NN:85.6 8%DT:80. 4%	14
Evangelia I. Zacharaki · Vasileios G. Kanas · Christos Davatzikos	Investigating machine learning techniques for MRI-based classification of brain neoplasms	Brain tumor	University of Pennsylvania (152)	Best First search KNN classifier WEKA software	96.9%	15
S M Hasan Mahmud, Md Altab Hossin, Md. Razu Ahmed, Sheak Rashed Haider Noori, Md Nazirul Islam Sarkar	Machine Learning Based Unified Framework for Diabetes Prediction	Diabetes	UCI [originally data collected from the National Institute of Diabetes and Digestive and Kidney Diseases] (768)	Artificial Neural Network(ANN) Support Vector Machine(SVM) Logistics Regression (LR) Decision Tree (DT) Random Forest (RF) Naive Bayes (NB)	74%	16
I. Huertas-Fernández , F. J. García-Gómez , D. García-Solís, S. Benítez-Rivero , V. A. Marín-Oyaga, S. Jesús , M. T. Cáceres-Redondo , J. A. Lojo , J. F. Martín-Rodríguez, F. Carrillo, P. Mir	Machine learning models for the differential diagnosis of vascular parkinsonism and Parkinson's disease using [123I]FP-CIT SPECT	Parkinson' s disease	Data from ROI and SPM analyses	ROI analysis and SPM LR, LDA, SVM BRASS software	90 % LR+ROI 90.3% SVM+ SPM 90.4 %	17
Trang Pham, Truyen Tran, Dinh Phung, Svetha Venkatesh	Predicting healthcare trajectories from medical records: A deep learning approach	diabetes (7191)and Mental health	Australian hospital.	Deep Care, an end-to-end deep dynamic memory neural network		18
Daniel Vieira, Jaakko Hollmen.	Resource Frequency Prediction in Healthcare: machine learning approach	resource frequency prediction	Oulu University Hospital in Finland	Nearest Neighbours and Random Forest		19

Prof. Dhomse Kanchan B. , Mr. Mahale Kishor M.	Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis	diabetes	1865	WEKA data mining tool Naive Bayes Decision Tree SVM		7
		heart disease	Cleveland Clinic Foundation			
Rohan Bhardwaj, Ankita R. Nambiar, Debojyoti Dutta	A Study of Machine Learning in Healthcare	Several disease	According to disease	ML Bigdata		2
Binh P. Nguyena, Hung N. Phamb, Hop Trana, Nhung Nghiem, Quang H. Nguyen , Trang T.T. Do, Cao Truong Tran , Colin R. Simpson	Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records	diabetes	EHR dataset	deep learning neural network	84.13%	20
Faizan Zafar, Saad Raza, Muhammad Umair Khalid, Muhammad Ali Tahir	Predictive Analytics in Healthcare for Diabetes Prediction	Diabetes	PIMA dataset	boosting algorithm	89.94%	21
Joshua I. Glaser, Ari S. Benjamin, Roozbeh Farhoodi, Konrad P. Kording	The roles of supervised machine learning in systems neuroscience	Neuro Alzheimer's disease	motor cortex, somatosensory cortex, and hippocampus	deep learning		22
Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh	Analysing Feature Importances for Diabetes Prediction using Machine Learning	Diabetes	National Institute of Diabetes and Digestive and Kidney(262)	Random Forest	84%	23
U Srinivasulu Reddy, Aditya Vivek Thota, A Dharun.	Machine Learning Techniques for Stress Prediction in Working Employees	Stress Prediction	OSMI Mental Health	Boosting random forest	75.13%	24
Dhiraj Dahiwade, Prof. Gajanan Patle, Prof. Ektaa Meshram	Designing Disease Prediction Model Using Machine Learning Approach	general disease prediction	UCI	KNN CNN	84.5%	25

3.2 Machine learning techniques for Diabetes disease: In this part of survey we investigation the specific disease which is diabetes. There are some papers are analyzed. Each and every paper use different datasets, attribute and ML techniques. Then we accomplish them in tabular form.

Table 3.1Machine learning techniques for Diabetes disease

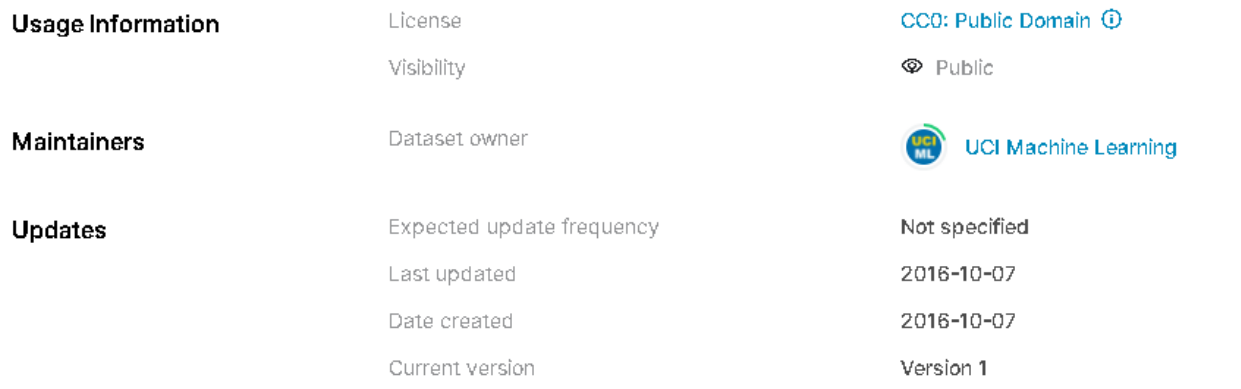
Research paper	Artificial Neural Network (ANN)	Random Forest (RF)	Naive Bayes (NB)	Support Vector Machine (SVM)	Logistics Regression (LR)	Decision Tree (DT)	Boosting algorithm	deep learning neural network
Paper 9	✓	✓	✓	✓				
Accuracy	72%	73%	74%	76%				
Paper 10	✓							
Accuracy	60.5%							
Paper 16	✓	✓	✓	✓	✓	✓		
Accuracy	68%	71%	74%	73%	70%	71%		
Paper 20								✓
Accuracy								84.13%
Paper 21							✓	
Accuracy							89.94%	
Paper 23		✓						
Accuracy		84%						

4. DATA FILE

Context: This dataset is initially from the National Institute of Diabetes and Digestive and Kidney Diseases. The target of the dataset is to indicatively foresee whether a patient has diabetes, in light of certain demonstrative estimations remembered for the dataset. A few requirements were put on the determination of these occurrences from a bigger data set. Specifically, all patients here are females in any event 21 years of age of Pima Indian legacy.

Content: The datasets comprise a few clinical indicator factors and one objective variable, the result. Indicator factors incorporates the number of pregnancies the patient has had, their BMI, insulin level, age, etc.

Acknowledgements: Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Symposium on Computer Applications and Medical Care (pp. 261--265). IEEE Computer Society Press [26].



5. Proposed work

5.1Text Mining Algorithm

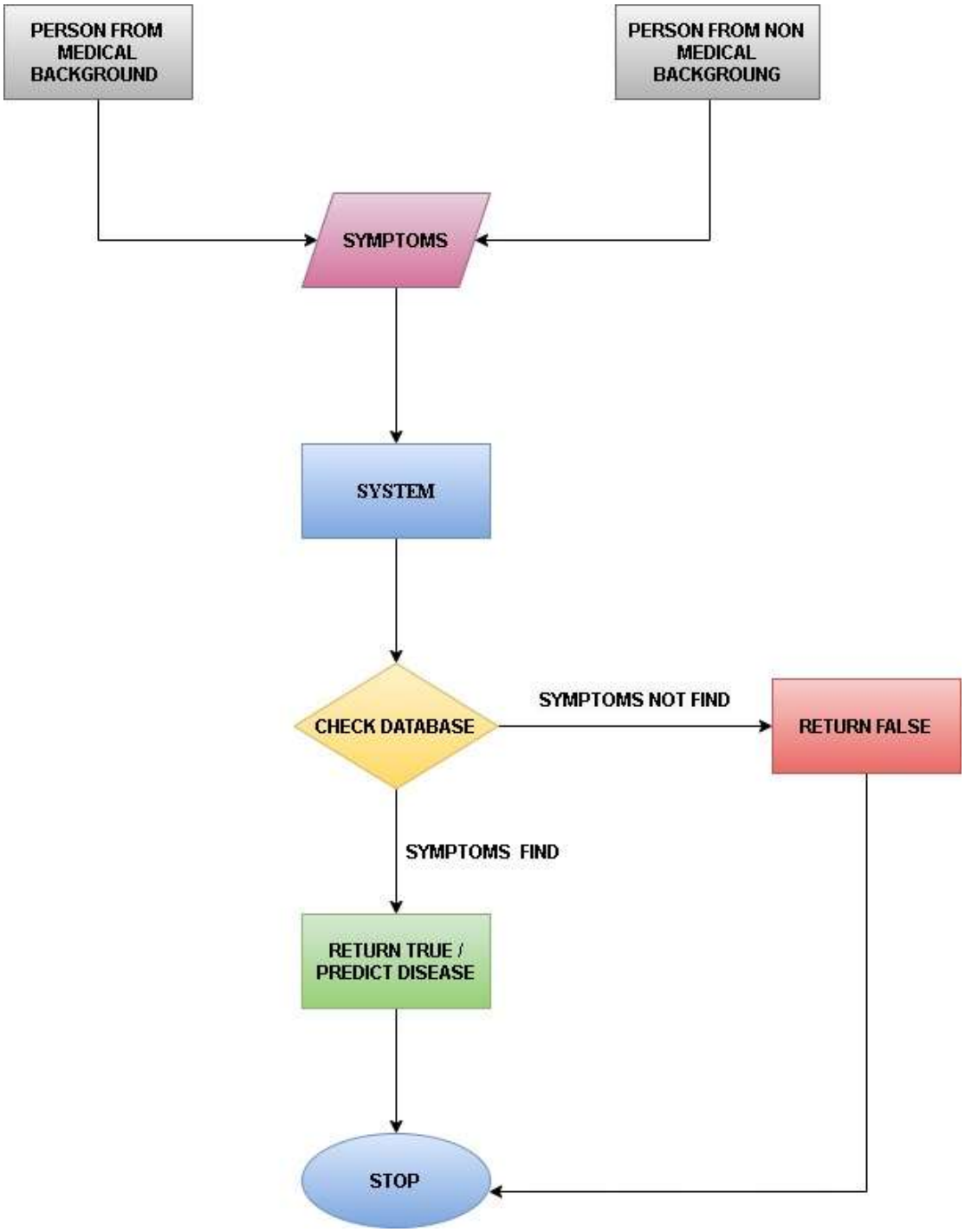


Fig-1 Text Mining Algorithm

5.2 ML Algorithm

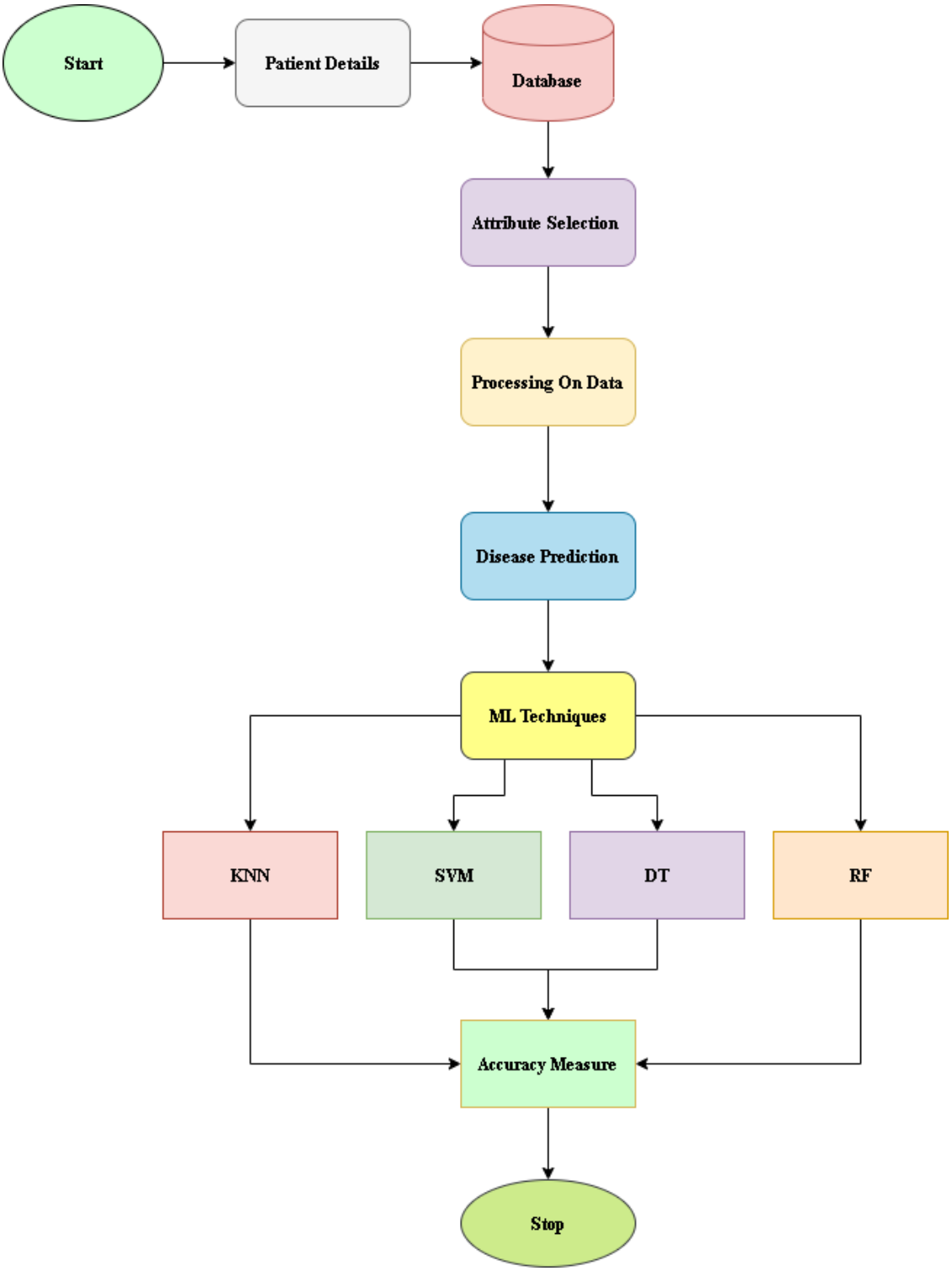


Fig-2 ML Algorithm

6. ALGORITHAM & PROCESS

Text Mining

Input:

Dataset D

Enter Symptoms

Output:

1. If symptom is match then outcome is true (disease predicted)

2. If symptom is not match then outcome is False (disease not predicted)

Process:

Step 1: Import libraries

Step 2: describe () and info () methods for preview the data.

Step 3: Enter symptoms

Step 4: Machine predict the person is suffering from the disease or not.

Step 5: Apply ML techniques.

Step 6: Find the accuracy.

Step 7: Stop

7. TOOL USED

Anaconda mainly used for distribution of the Python and R programming languages for logical dispensation, which means to work on manage and send a package. The distribution integrates data science packages suitable for Windows, Linux, as well as mac OS.

Jupyter is free of cost, open-source, intelligent web tool known as a computational journal, which professionals can utilize intersection with programming code, computational revenue, logical content, and diversified media assets in an introverted record.

Fig-3 Jupyter Notebook Environment

8. IMPLEMENTATION

8.1 Disease prediction via symptoms:

If disease predicted by symptoms output will be true

```
Slow-healing sores
Blurred vision
Irritability
True
```

Fig-4 disease predicted output is true

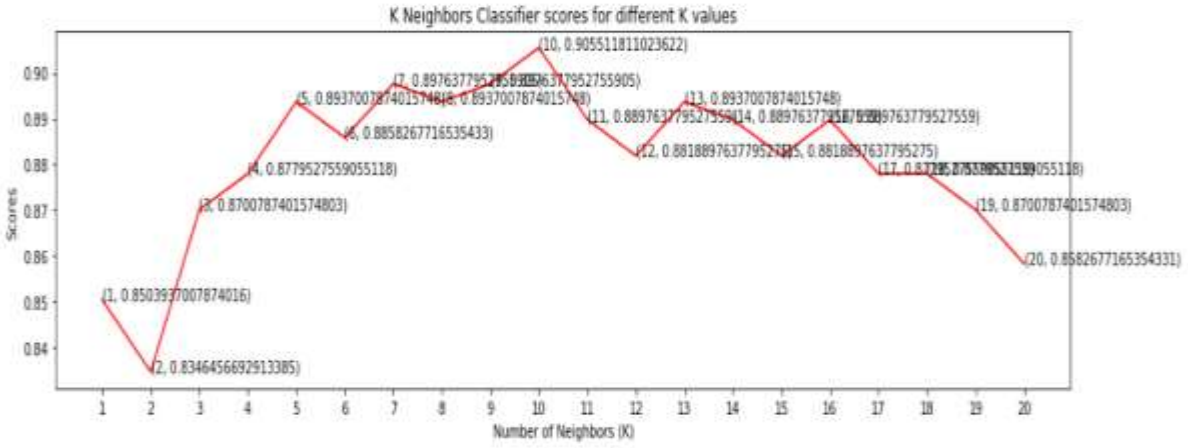
If disease predicted by symptoms output will be false.

```
bp
headache
leg pain
False
```

Fig-5 Disease is not predicted output is false

8.2 Accuracy according several ML techniques:

1. K-nearest neighbour

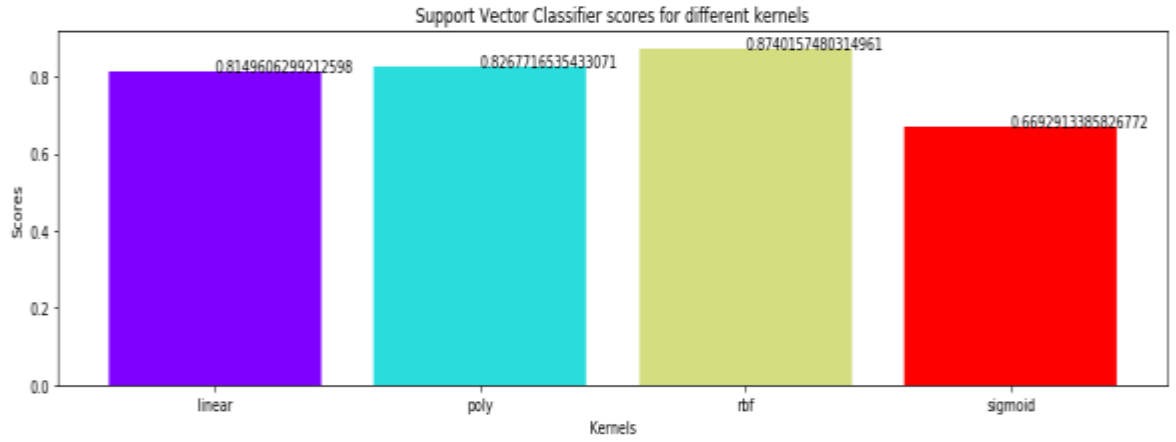



```
print("The score for K Neighbors Classifier is {}% with {} nieghbors.".format(knn_scores[3]*100, 10))
```

```
The score for K Neighbors Classifier is 87.79527559055119% with 10 nieghbors.
```

Fig-6 Accuracy of KNN

2. Support Vector Machine

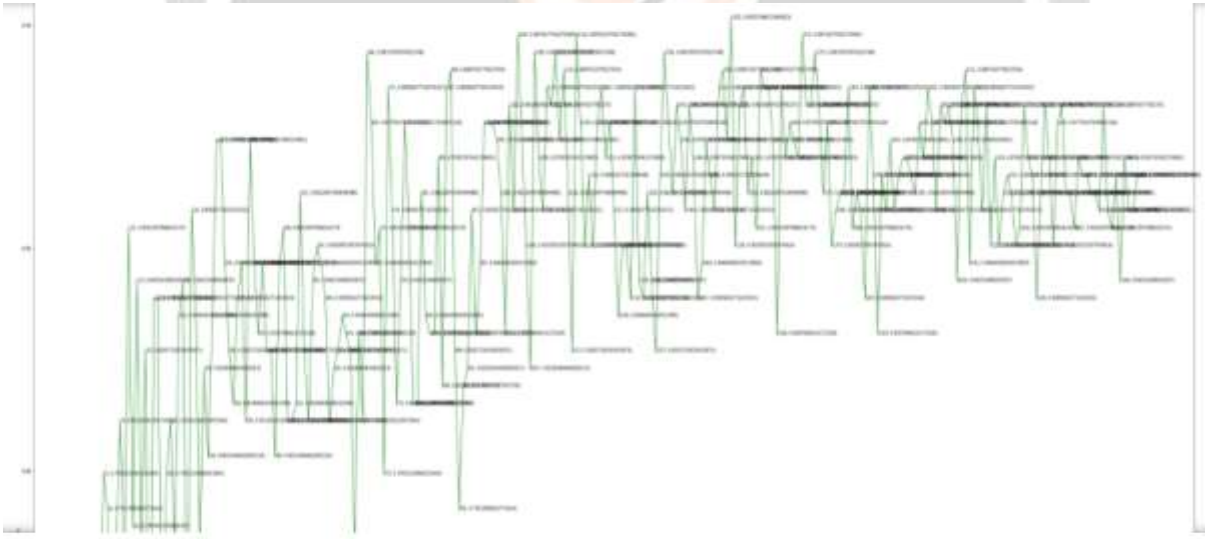



```
print("The score for Support Vector Classifier is {}% with {} kernel.".format(svc_scores[0]*100, 'rbf'))
```

```
The score for Support Vector Classifier is 81.49606299212599% with rbf kernel.
```

Fig-7 Accuracy of SVM

3.Decision Tree:

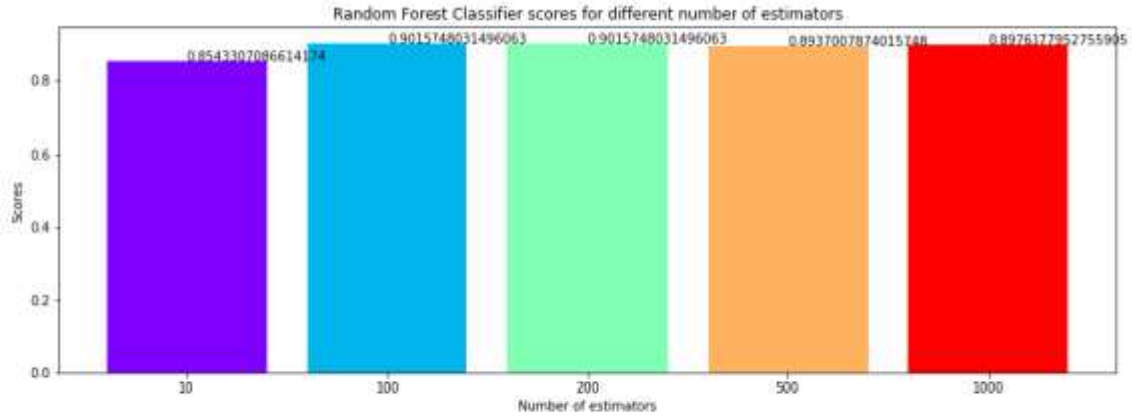


```
print("The score for Decision Tree Classifier is {}% with {} maximum features.".format(dt_scores[17]*100, [155]))
```

```
The score for Decision Tree Classifier is 83.85826771653542% with [155] maximum features.
```

Fig-8 Accuracy of DT

4.Random Forest:




```
print("The score for Random Forest Classifier is {}% with {} estimators.".format(rf_scores[1]*100, [100, 200]))
```

```
The score for Random Forest Classifier is 90.15748031496062% with [100, 200] estimators.
```

Fig-9 Accuracy of RF

9. CONCLUSION

These work displays that ML techniques are significant in medical segment. ML give diverse techniques for getting the accurate result. It also confirms that ml approaches are exclusively utilized in the diagnosis and predictions of diabetes disease. These work is also demonstrations that Jupyter notebook can be measured to be as one of the most popular in addition to suitable ML tool as it supports numerous tasks and text mining algorithm is consume less time to distinguish disease through the symptoms.

10. ACKNOWLEDGEMENT

I am deeply indebted & would like to express my gratitude to my thesis guide Asst. Prof. Hemali Shah. Post-Graduation Department for her great efforts and instructive comments in the dissertation work. She has devoted significant amount of her valuable time to plan and discuss the Literature review work. During the dissertation she provided me excellent guidance & support during discussion about my progress. She gave me a great deal of freedom to choose a research topic, focusing more on my research interests and building skills that will allow me to be a successful in my life, without her experience and insights, it would have been very difficult to do qualitative work. I am grateful for her cooperation during the period of my dissertation phase. I would also like to thank all staff members and my colleagues who were always there at the need of hour and provided with all help and facilities, which I required. I would also like to thanks to all my friends for helping me in the hour of need and providing me all the help and supported me during this thesis work. Last but not the least I am highly grateful to all my family members for their inspiration and ever encouraging moral support, which enables me to pursue my studies.

11. REFERENCES

[1] Nithya, B. and Ilango, V., 2017, June. Predictive analytics in health care using machine learning tools and techniques. In 2017 International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 492-499). IEEE

[2] Bhardwaj, R., Nambiar, A.R. and Dutta, D., 2017, July. A study of machine learning in healthcare. In 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC) (Vol. 2, pp. 236-241). IEEE.

[3] Shailaja, K., Seetharamulu, B. and Jabbar, M.A., 2018, March. Machine Learning in Healthcare: A Review. In 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA) (pp. 910-914). IEEE

[4] Vieira, D. and Hollmén, J., 2016, June. Resource frequency prediction in healthcare: machine learning approach. In 2016 IEEE 29th International Symposium on ComputerBased Medical Systems (CBMS) (pp. 88-93). IEEE

[5] Saranya, M.S., Selvi, M., Ganapathy, S., Muthurajkumar, S., Ramesh, L.S. and Kannan, A., 2017, January. Intelligent medical data storage system using machine learning approach. In 2016 Eighth International Conference on Advanced Computing (ICoAC) (pp. 191-195). IEEE

[6] Dou, H., 2019, June. Applications of Machine Learning in the Field of Medical Care. In 2019 34rd Youth Academic Annual Conference of Chinese Association of Automation (YAC) (pp. 176-179). IEEE

[7] Kanchan, B.D. and Kishor, M.M., 2016, December. Study of machine learning algorithms for special disease prediction using principal of component analysis. In 2016 international conference on global trends in signal processing, information computing and communication (ICGTSPICC) (pp. 5-10). IEEE

[8] Kadaru, B.B. and Reddy, B.R.S., 2018. A novel ensemble decision tree classifier using hybrid feature selection measures for Parkinson's disease prediction. International Journal of Data Science, 3(4), pp.289-307

[9] Dwivedi, A.K., 2018. Analysis of computational intelligence techniques for diabetes mellitus prediction. Neural Computing and Applications, 30(12), pp.3837-3845.

[10] Amato, F., López, A., Peña-Méndez, E.M., Vaňhara, P., Hampl, A. and Havel, J., 2013. Artificial neural networks in medical diagnosis.

[11] Wu, H., Deng, Z., Zhang, B., Liu, Q. and Chen, J., 2016. Classifier model based on machine learning algorithms: application to differential diagnosis of suspicious thyroid nodules via sonography. American Journal of Roentgenology, 207(4), pp.859-864.

[12] Baek, J.W. and Chung, K., 2019. CNN-based health model using knowledge mining of influencing factors. Personal and Ubiquitous Computing, pp.1-11.

[13] Venkatesan, C., Karthigaikumar, P., Paul, A., Satheeskumaran, S. and Kumar, R.J.I.A., 2018. ECG signal preprocessing and SVM classifier-based abnormality detection in remote healthcare applications. IEEE Access, 6, pp.9767-9773.

[14] Palaniappan, S. and Awang, R., 2008, March. Intelligent heart disease prediction system using data mining techniques. In 2008 IEEE/ACS international conference on computer systems and applications (pp. 108-115). IEEE

[15] Zacharaki, E.I., Kanas, V.G. and Davatzikos, C., 2011. Investigating machine learning techniques for MRI-based classification of brain neoplasms. International journal of computer assisted radiology and surgery, 6(6), pp.821-828.

[16] Mahmud, S.H., Hossin, M.A., Ahmed, M.R., Noori, S.R.H. and Sarkar, M.N.I., 2018, August. Machine Learning Based Unified Framework for Diabetes Prediction. In Proceedings of the 2018 International Conference on Big Data Engineering and Technology (pp. 46-50).

[17] Huertas-Fernandez, I., Garcia-Gomez, F.J., Garcia-Solis, D., Benitez-Rivero, S., MarinOyaga, V.A., Jesus, S., Caceres-Redondo, M.T., Lojo, J.A., Martin-Rodriguez,] J.F., Carrillo, F. and Mir, P., 2015. Machine learning models for the differential diagnosis of vascular Parkinsonism and Parkinson's disease using [123 I] FP-CIT SPECT. European journal of nuclear medicine and molecular imaging, 42(1), pp.112-119.

[18] Pham, T., Tran, T., Phung, D. and Venkatesh, S., 2017. Predicting healthcare trajectories from medical records: A deep learning approach. Journal of biomedical informatics, 69, pp.218-229.

[19] Vieira, D. and Hollmén, J., 2016, June. Resource frequency prediction in healthcare: machine learning approach. In 2016 IEEE 29th International Symposium on ComputerBased Medical Systems (CBMS) (pp. 88-93). IEEE.

[20] Nguyen, B.P., Pham, H.N., Tran, H., Nghiem, N., Nguyen, Q.H., Do, T.T., Tran, C.T. and Simpson, C.R., 2019. Predicting the onset of type 2 diabetes using wide and deep Learning with electronic health records. Computer methods and programs in biomedicine, 182, p.105055.

[21] Zafar, F., Raza, S., Khalid, M.U. and Tahir, M.A., 2019, March. Predictive Analytics in Healthcare for Diabetes Prediction. In Proceedings of the 2019 9th International Conference on Biomedical Engineering and Technology (pp. 253-259).

[22] Glaser, J.I., Benjamin, A.S., Farhoodi, R. and Kording, K.P., 2019. The roles of supervised machine learning in systems neuroscience. Progress in neurobiology.

[23] Dutta, D., Paul, D. and Ghosh, P., 2018, November. Analysing Feature Importances for Diabetes Prediction using Machine Learning. In 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON) (pp. 924- 928). IEEE.

[24] Reddy, U.S., Thota, A.V. and Dharun, A., 2018, December. Machine Learning Techniques for Stress Prediction in Working Employees. In 2018 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC) (pp. 1-4). IEEE.

[25] Dahiwade, D., Patle, G. and Meshram, E., 2019, March. Designing Disease Prediction Model Using Machine Learning Approach. In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC) (pp. 1211-1215). IEEE

[26] **https://www.kaggle.com/uciml/pima-indians-diabetes-database/metadata**