A Machine Learning Technique for Reducing Hospital Readmissions for Diabetic Diseases

Mohammad Ismail, D.Praveen Kumar Reddy, K.Sai Srikanth

Abstract- The number of readmissions in diabetic diseases keeps increasing from time to time in patients from various hospitals. This brings a dreadful name to the hospital and is also considered as an act of irresponsibility of the doctors. So in order to reduce the readmissions of diabetic patients, we propose an approach which uses a machine learning technique to compare the hospital records of various patients. We have used various diabetic dataset features for our technique to predict the readmission probability rates of patients. We compared our proposed technique with existing Machine Learning algorithms like Random Forest, K-means clustering, Support Vector Machine(SVM) and found the best possible prediction with proposed approach using** receiver operating characteristic(**ROC) curve.

Keywords—Natural Language Handling, Medical Data Frameworks, Decision Emotionally Supportive Networks, Data Mining, Feature Extraction

I. INTRODUCTION

Under an article regarding health information retrieval [1] underlined the selection of wellbeing data innovation through an act passed under health information regarding clinical notes. Two prime commitments that are regarded as crime are(1) Punishing emergency clinics for quiet readmission into various multi-specialty hospitals. (2) Introduction of the Clinical data gathering from previously visited patients in Electronic Health Records through consistence [3]. Right now, it requires an essential execution of guideline based choice emotionally supportive networks where small scale clinical doctors or students who are still in practice use socio economics, meds, sensitivity, and previously gathered patient. There is an Act that states the patient data or clinical notes should be utilized for better treatment of patients. As a major aspect of this demonstration, CMS recognized "clinic readmissions for DIABETIC DISEASES" as an exorbitant issue that should be tended to [4]. This issue is considered serious and the information required is accessible through content management.

Clinical management has begun punishing emergency clinics for over the top 30-day readmissions. Subsequently, there is an expanded measure of weight on emergency clinics to receive the CDSS to recognize the contender for medical clinic readmission and maintain a strategic distance from such readmissions by a progression of endeavors, for example, intently composed change of care.

Lamentably, it is absurd to expect to give such a broad degree of care for each patient because of the measure of assets required, deficiency in therapeutic staff, and the costs associated with such care coordination [4]–[6]. In this way, it is basic to precisely distinguish possibility for emergency clinic readmission and after that dodge such readmission using assets. Further, since persistent hospitalization speaks to such a huge part of social insurance costs, well being plans for various accountable and management service organizations are likewise focusing on clinic readmission so as to improve their gainfulness. In spite of the fact that prescient displaying for some, infections have seen an enormous group of research ,Diabetic diseases prescient demonstrating stays rare.

Tolerant information in medical clinics incorporates a lot of unstructured information. Models incorporate doctor's notes, release synopses, and various radiology reports. As the content is significant as per the patient records using it in the examination or analysis of previously gathered data such as report or a survey of free content is a hectic task. . Along these lines, there is enthusiasm for building up a Natural Language Processing based approach to deal with the data gathered from various hospitals into a dataset.

II. BACKGOUND

Natural language processing is a field of computer science, computational intelligence, and artificial intelligence dealing with the communications between computers and natural languages, in particular and tells us how to **process** and analyze large amounts computer data used in natural language processing . The NLP has the more number of libraries to provide the best suitable outputs without programming.

A. NLP Libraries

Programming libraries are commonly characterized as an accumulation of schedules, capacities, or classes which are intended to extract an intricate issue. They are made considering reusability and intended to empower developers to compose programming without finding duplicate efforts. NLP has various inbuilt libraries accessible, went for various dialects and purposes. This examination utilizes OpenNLP, which is a Java based library in NLP domain.

The openNLP toolkit came into use in 2000 under many Java interfaces. They made a standard application program interface for NLP . The first usage of the following libraries was made by scientists at the University of Edinburgh in a framework for openNLP [7]. In the following years it was mixed with Apache server where they are conversed into a solitary tool box and it is updated to an Apache top level server.

Revised Manuscript Received on December 30, 2019.
* Correspondence Author

Dr. Mohammed ismail.b*, Professor, Department of CSE in KLEF Deemed to be University, received his Bachelor of Engineering Degree from Visvesvaraya Technological University, Belgaum, India.

D.Praveen, Student, Department of Computer Science and Engineering , Koneru Lakshmaiah Educational Foundation situated at Vaddeswaram, Guntur District, Andhra Pradesh, India.

K.Sai Srikanth, Student, Department of Computer Science and Engineering , Koneru Lakshmaiah Educational Foundation situated at Vaddeswaram, Guntur District, Andhra Pradesh, India.

B. Medical NLP

The medicinal area has been probably the most punctual utilization of NLP [1]–[6]. Medicinal experts frequently compose the medical data which help us to understand the patient condition, medications, lab results, hereditary problems and everything else considered. Patient records for the most part incorporate organized restorative data notwithstanding unstructured content. In any case, this data is normally implied for checking the patients information to request government announcing laws. It isn't intended to pass on a total image of the patient. As there isn't much understanding to precisely how much data is in unstructured format ,the reports coincide a significant part of the data is kept in unstructured records [2]. The explanation so much therapeutic information isn't organized is extra information called as restorative programmers who decipher the clinical notes to organized structure. This representation is expensive as a rule just the absolute minimum required for preparing is performed. Along these lines, NLP offers a strategy to perhaps extricate a lot of data that isn't caught in organized notes. The Clinical Text analysis and Extraction system (cTAKES) made by experts the Mayo center started in the 19th century .kept up.

cTAKES uses a design mechanism called Apache UIMA [11]. cTAKES makes use of various programming segments to group the pieces of the framework. Apache helps to understand the usefulness to low level NLP errands, for example, sentence discovery, dividing into tokens lumping, grammatical feature identification, and some more basic NLP assignments. cTAKES separates essential NLP data from records and adds it to the data in CAS.

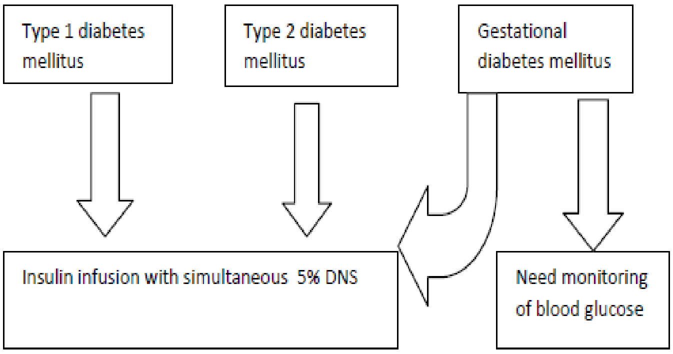
III. RELATED WORKS

With the entry of government enactment punishing unnecessary multi day emergency clinic readmission, readmission hazard demonstrating has been a functioning field of research. An implementation was done by Kansagara in 2011. It analyzes information, approach, and also the outputs [2]. The audit confirmed that high readmission expectation is a troublesome issue and is highly restricted by various organizations. Research which endeavored to take care of the 30-day readmission issue for the most part performed more regrettable than the people attempting to be readmitted after many months . An expected explanation for this is that patients are readmitted after long intervals of time, hence making marginally progressively adjusted class conveyance. Also, numerous frameworks had the option to build execution by taking into account a few number of patients for example, those with severe diabetic patients[5]. The dataset example changed significantly. An investigation utilizing a British health service examined over millions of patients with AUC(0.72) and another examination just utilized 500 patients with AUC=0.70 [4]. Here considering all accounts only dataset model when examine proved that the best execution used for 700 patients gave AUC(0.83). Obviously, nature of information would in general have an enormous effect on model quality while the models utilizing information from CMS proved to give better results.

1) Prediction of hospital readmission in Diabetic patients:

A system was made by specialists at Deakin University to break down numerous ceaseless malady readmissions [2]. Infections are coordinated to formats of highlights and diabetic diseases is an ailment that was considered. This framework speaks to one of only a handful couple of research endeavors to dissect diabetic diseases patients however doesn't concentrate exclusively on diabetic diseases and moreover centers explicitly around the 30 days readmission of patients. This framework works makes diagrams that are utilized while recording the information of a particular patient[9]. On account of diabetic diseases, a explicit layout is utilized. Ailment explicit models are then manufactured[1]. This strategy functions admirably in light of the fact that as recently referenced, constructing a solitary classifier for a differing understanding populace can regularly diminish model execution. 1,816 patients were broke down. The model can anticipate 30 days readmission rate recorded in the diabetic diseases patients showed an AUC of 0.62 and was an improvement upon co-dreariness pattern techniques that are frequently utilized for readmission examination.

2)A precise survey was performed in 2011 by Kansagara et al. which thinks about information, strategy, and results [10]. The survey affirmed that readmission expectation is a troublesome issue and late models don't really perform superior to look into 10 years earlier. A system was made by analysts at Deakin University to break down numerous incessant malady readmissions [4]. The framework works by making constructions to be utilized when catching information for a patient. On account of disease, a explicit format is utilized. Illness explicit models are then constructed. 1,816 patients were broke down. The model can anticipate 30-day readmission rate in patients with an AUC=0.67 and was an improvement upon co-dreariness standard techniques that are frequently utilized for readmission investigation. Another framework which is explicit to patients was made by Fan et al. [3]. This framework was not anyway utilized in the examination of medical clinic readmissions. Rather, patients were examined for intensifications inside the time of one year. Standard strategies for examination utilized a model comprising of fundamental highlights, for example, socioeconomics and survey data[12]. An improved model Which was introduced which additionally incorporated the highlights Kidney Damage, Alzheimer’s Disease earlier intensifications and co-dreariness[6]. The AUC for this model was 0.68.


Type 1 diabetes mellitus
Type 2 diabetes mellitus
Gestational diabetes mellitus
Insulin infusion with simultaneous 5% DNS
Need monitoring of blood glucose

Figure(1):Types of Diabetes and how they are monitored.

IV. METHODOLOGY

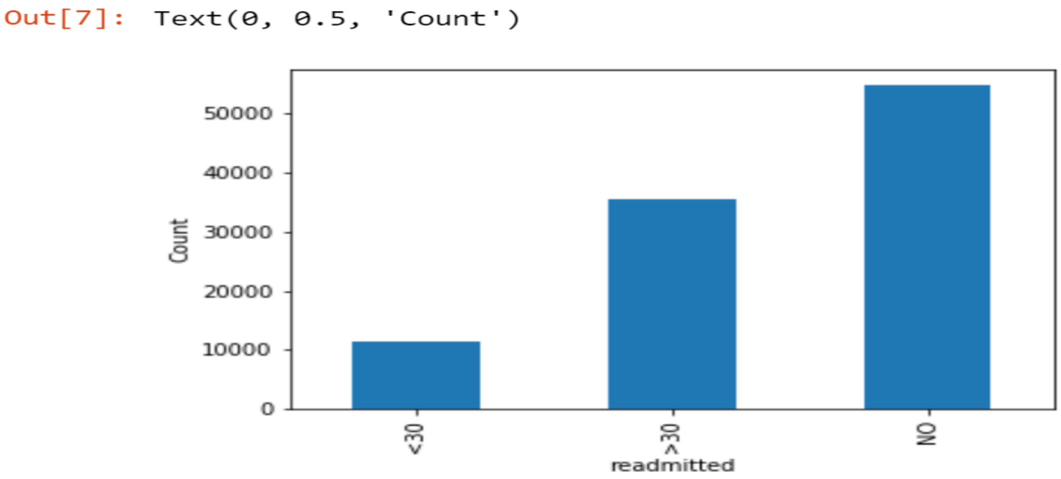
The structure can be separated into four subsystems: (1) Feature extraction (2) Feature determination (3) Classification (4) Performance assessment.

A .Data: The diabetes Dataset we have utilized comprises of 10,000 records and 52 highlights. Underneath given is the point by point depiction of a portion of the significant highlights in the dataset.

Table(1):Class Distribution for Features

S no	Feature name	Class Distribution
1	Repaglinide	("No": 9870, "Steady": 112, "Up":13, "Down":5)
2	Nateglinide	("No": 9949, "Steady": 49, "Up":1, "Down":1)
3	Chlorpropamide	("No": 9987, "Steady": 12, "Up":1, "Down":0)
4	Tolbutamide	("No": 9997, "Steady": 3, "Up":0, "Down":0)
5	Acarbose	("No": 9968, "Steady": 31, "Up":1, "Down":0)

1. For the principle task shown in fig 1 (Readmitted or not), the connection investigation brought about 21 were not
We conducted the correlation analysis for the main task (readmitted or not), and the two other subtasks on the imputed dataset to identify the significant features. For categorical data, we used, Chi square test, and for numerical data, we used Pearson's coefficient. We used the metric p-value to determine
the significance of a feature for a task. If the p value is below a threshold, which is 0.05, then the feature is considered significant.



```
Out[7]:  Text(0, 0.5, 'Count')
```

Fig 1:Results of readmitted rates compared with age groups

2. For task 1 (time in medical clinic), the connection investigation brought about 22 critical highlights as shown in fig 2.

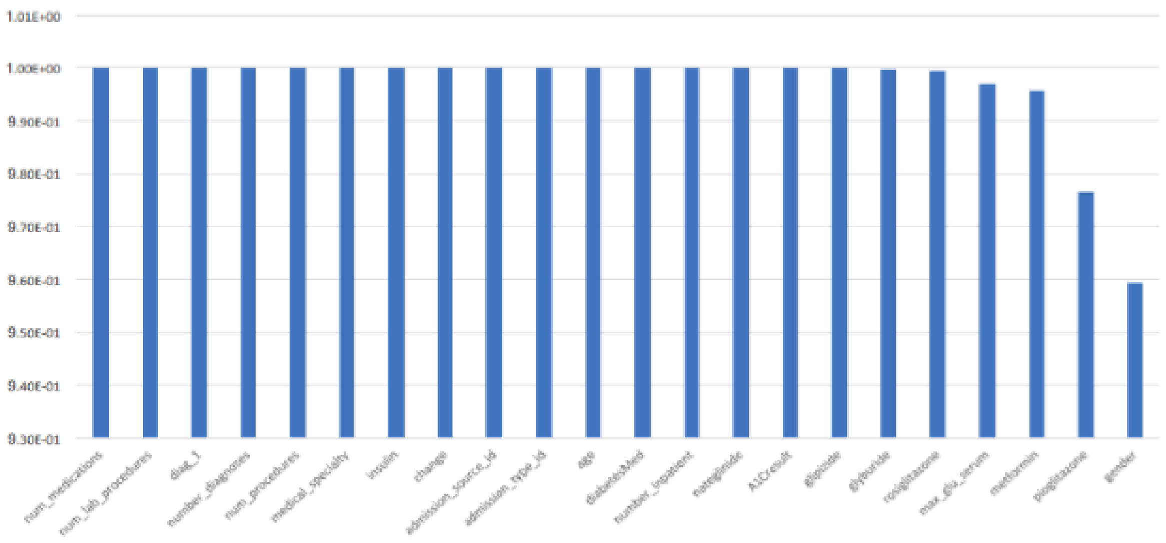


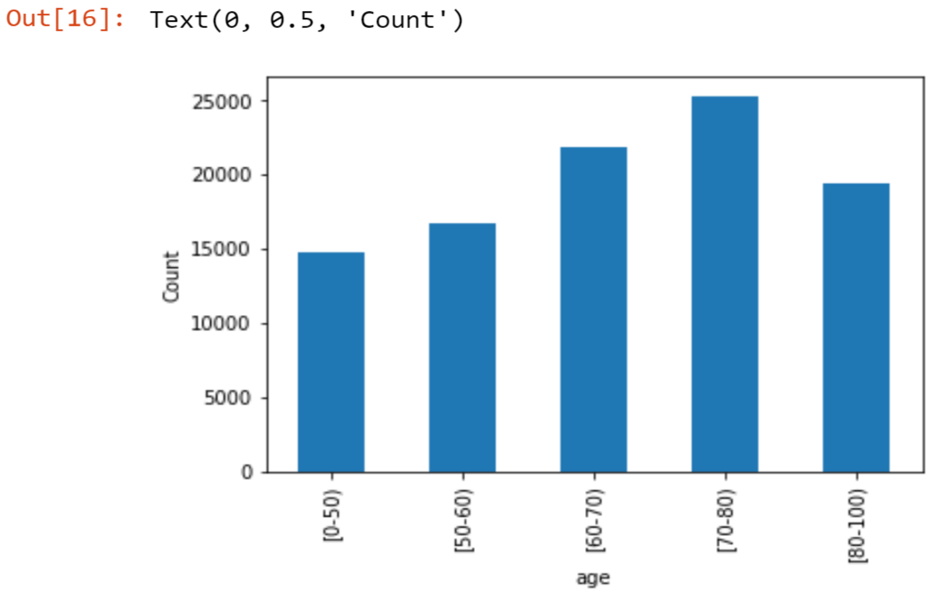
Fig 2: Feature wise impact on patient readmission Rates

This is a relapse issue. Here we are attempting to decide the span of remain of a patient on day 0, when the patient gets readmitted, in light of the information accessible by then of time.

$$T=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - Y_i)^2} \quad \text{Eq (1)}$$

For remaining patients

3. For task 2 (Age, count), the relationship examination brought about 20 huge highlights as shown in fig 3.



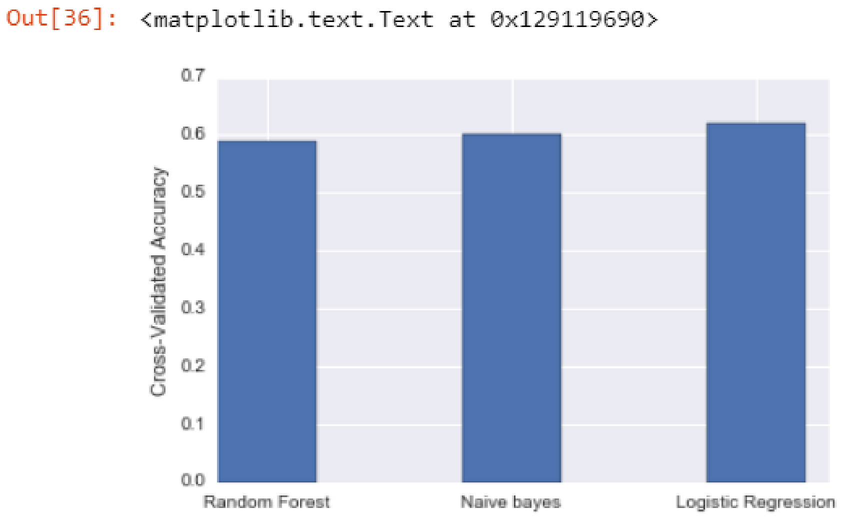
```
Out[16]: Text(0, 0.5, 'Count')
```

Fig 3:count of readmissions for different age groups

To improve our Future selection, we considered the class lopsidedness issue in our dataset. Class lopsidedness is an issue which happens in the AI methods, where we find the complete class information which is positive is did not match the another class information i.e., negative. In the wake of investigating the various highlights in the dataset, we dispensed with the highlights which had class awkwardness. All the dispensed with highlights are recorded beneath:

(3) Classification

Naive Bayes classifier depends on the Bayesian hypothesis, and is appropriate especially when the dimensionality of the info is high. The Bayesian arrangement is a regulated learning technique which is regarded as one of the measurable strategy for grouping. Expect a probabilistic model which enables us to find the risk in the model by deciding finding the probabilities. It also takes care of indicative and prescient problems. Regardless of its effortlessness, this classifier can beat increasingly complex classifiers. We utilized "naive Bayes" library in R to create a Naïve Bayes classifier. This classifier accomplished a precision of 62.15% as shown in fig 4



```
Out[36]: <matplotlib.text.Text at 0x129119690>
```

Fig 4: Cross validate accuracy of algorithms used.

Plotting the results:
Relative Feature importance of all the features as shown in fig 5:

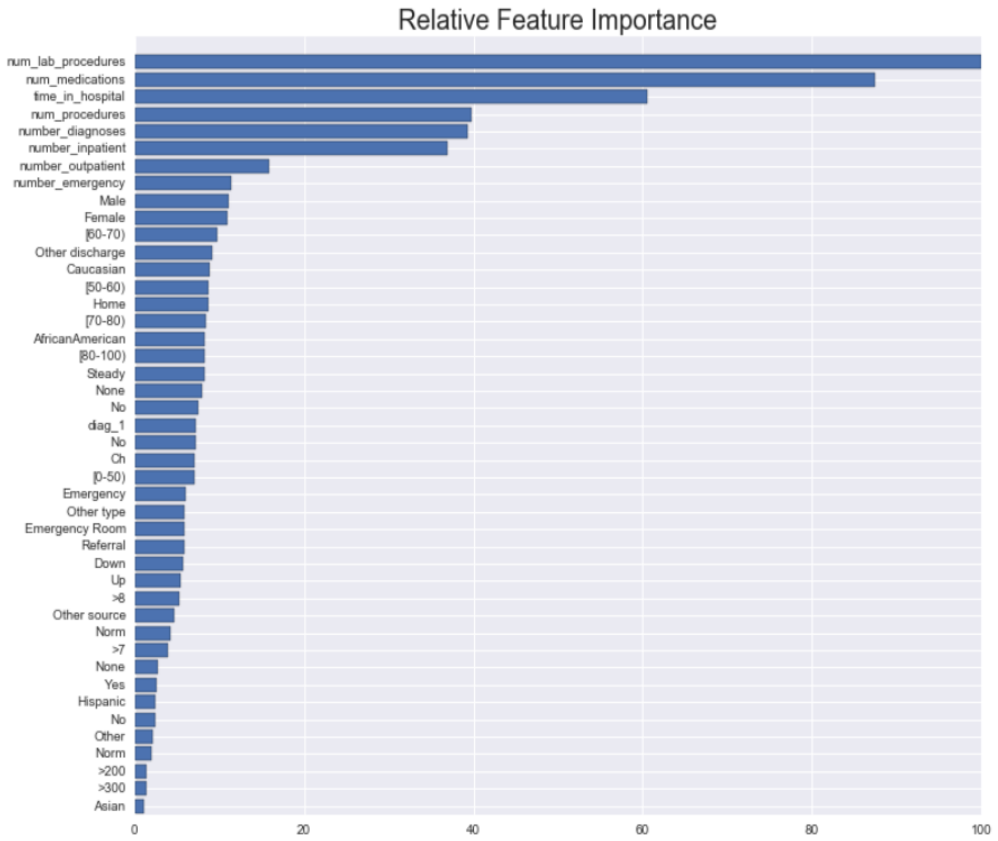


Fig 5: Relative feature importance of features used

The value of c for logistic regression against Cross validated accuracy is as shown in the fig 6:

```
Out[40]: <matplotlib.text.Text at 0x11a8f3410>
```

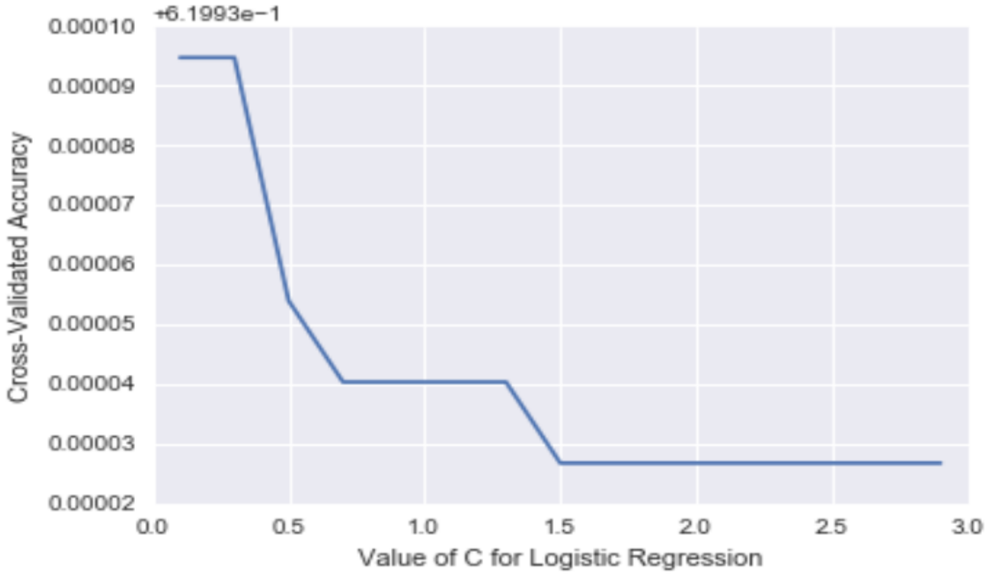


Fig 6: Logistic regression against cross validated accuracy

Plotting the ROC curve: The ROC curve for Diabetes Readmission i.e.,1-Specificity against Sensitivity is as shown in the fig 7

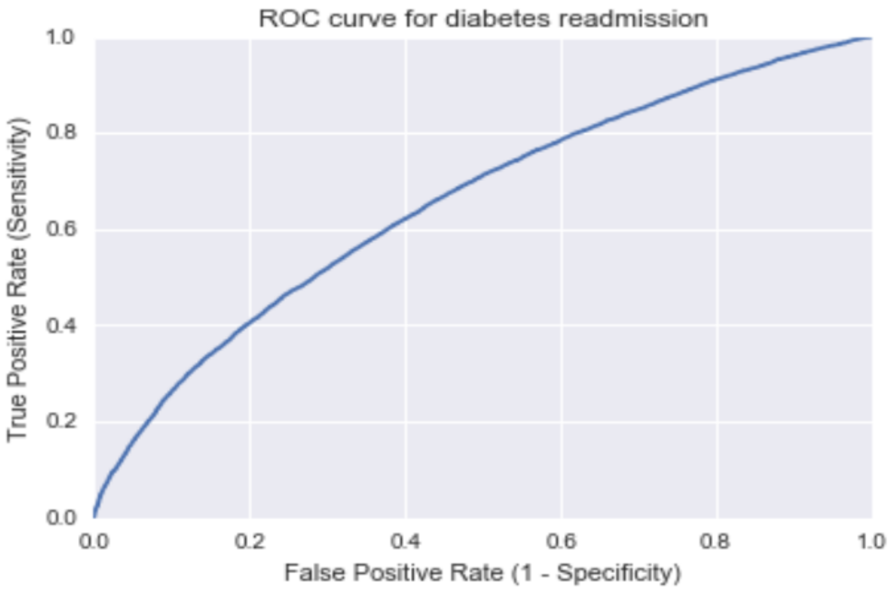


Fig 7: ROC curve for diabetes readmission rate

V. CONCLUSION

From the above results. we have used various diabetic dataset features for our technique to predict the readmission probability rates of patients. We compared our proposed technique with existing Machine Learning algorithms like Random Forest, K-means clustering, Support Vector Machine (SVM) and compared them with each other to find the accuracy for each algorithm and finally our technique is able to achieve 0.62 accuracy and 0.66 AUC score.

REFERENCES

1. R. Wallmann, J. Llorca, I. Gómez-Acebo, Á. C.Ortega,F.R. DOI 10.1016/j.jksuci.2016.11.00 Publisher: **Elsevier** Roldan, and T. Dierssen-Sotos,&quot;Expectation of 30-day cardiovascular related emergency-readmissions utilizing basic managerial emergency clinic information, & quot; Int. J.Cardiol., vol. 164, no. 2, pp. 193–200, 2013.
2. D. Goodman, E. Fisher, and C. Chang, "The Revolving Door: A Report on US Hospital Readmissions," Princeton, NJ Robert Wood Johnson Found., 2013.
3. P. Jain, Prognostic Diabetic Diseases human services the executives framework, no. May. FLORIDA ATLANTIC UNIVERSITY, 2014.
4. P.Raja Rajeswari, Supriya menon.M, A contemporary way for enhanced modeling of context aware privacy system in PPDM. Journal of Advanced Research in Dynamic and Control systems,Vol.10,01-issue,**July 2018.**
5. R. Behara, A. Agarwal, V. Rao, and C. Baechle, "Anticipating the Occurrence of Diabetes utilizing Analytics," in Models and Applications in the Decision Sciences: Best Papers from the 2015 Annual Conference, first ed., Pearson Press, 2016, pp. 187–193.
6. R. Behara, A. Agarwal, V. Rao, and C. Baechle,&quot;Prescient Analytics for Chronic Diabetes Care,&quot; in 2015 Annual Meeting of the Decision Sciences Institute Proceedings, 2015.
7. Mohammed Ismail B , K. Bhanu Prakash, M. Nagabhushana Rao” Collaborative Filtering-Based Recommendation of Online Social Voting” International journal of Engineering and Technology “ Volume 7 issue 3 1504-1507 **July 2018**
8. Mohammad Ismail K.Naga Lakshmi, Y. Kishore Reddy, M. Kireeti, T.Swathi” Design and Implementation of Student Chat Bot using AIML and LSA” International Journal of Innovative Technology and Exploring Engineering (IJITEE) Volume-8 Issue-6**,** ISSN: 2278-3075 PP 1742-1746 **April 2019**
9. R. Behara, A. Agarwal, F. Fatteh, and B. Furht, "Anticipating Hospital Readmission Risk for COPD Using EHR Information," in Handbook of Medical and Healthcare Technologies, Springer, 2013, pp. 297–308.
10. Mohammad Ismail, V.Harsha Vardhan, V.Aditya Mounika,K.Surya Padmini “An Effective Heart Disease Prediction Method Using Artificial Neural Network “International Journal of Innovative Technology and Exploring Engineering’ at Volume-8 Issue-8, pp 1529-1532 ISSN: 2278-3075 **June 2019**
11. Mohammed Ismail. B, Dr. B. Eswara Reddy, Dr. T. Bhaskara Reddy “Cuckoo Inspired Fast Search Algorithm for Fractal Image Encoding” Elsevier Journal of King Saud University Computer and Information Sciences volume 30 issue 4, ISSN: 1319-1578 Pages462–469 DOI 10.1016/j.jksuci.2016.11.00 Publisher: **Oct 2018**
12. J. H. Wasfy, G. Singal, C. O&#39;Brien, D. M.Blumenthal, K. F. Kennedy, J. B. Strom, J. A.Spertus, L. Mauri, S. L. T. Normand, and R. W. Yeh,&quot;Improving the Prediction of 30-Day Readmissionafter Percutaneous Coronary Intervention Using DataExtracted by Querying of the Electronic HealthRecord,&quot; Circ. Cardiovasc. Qual. Results, vol. 8, no.5, pp. 477–485, 2016.
13. K.Srinivas,Dr.Mohammed Ismail.B “Test case Prioritization With Special Emphasis On Automation Testing Using Hybrid Framework” Journal of Theoretical and Applied Information TechnologyVol.96. No 134180-4190 **July 2018**

AUTHOR PROFILE

Dr. Mohammed Ismail.B working as Professor CSE in KLEF Deemed to be University, received his Bachelor of Engineering Degree from Visvesvaraya Technological University, Belgaum India. M.Tech from JNTUH Hyderabad and a Ph.D. degree in Computer Science Engineering from JNTUA Anantapur (A.P) India in the area of Digital Image processing. He has 17 years of teaching (UG& PG) and research experience. He is Guiding 6 students for their doctoral studies in the area of image processing and machine learning. His Research Interests are Image processing, Machine Learning, and IOT. He has 30 Research Publications in reputed international/national journals and has presented 15 papers in International Conferences.

He has completed 4 sponsored research projects and 1 consultancy project and 1 Patent filled under his credit. He has guided 40 UG student projects and 4 P.G student projects. He has won awards from various Academic bodies and NGOs for his research and academic contributions. He is a life member of various academic bodies like CSI, ISTE, IAENG, and a member of IEEE. He has given various expert lectures in his research areas at various universities and colleges.

D.Praveen is a student of the Computer Science and Engineering Department at the Koneru Lakshmaiah Educational Foundation situated at Vaddeswaram, Guntur District.

K.Sai Srikanth is a student of the Computer Science and Engineering Department at the Koneru Lakshmaiah Educational Foundation situated at Vaddeswaram, Guntur District.