Journal Publication of International Research for Engineering and Management (JOIREM)

Hospital prediction using data mining

MS. Janvi Ahire,
MS. Rupali Nikam, MS. Shruti Parkhe, MS. Vaishnavi Pathak

Prof.S.R Jadhav
HOD Prof.V.D Badgujar

Department of computer Engineering, Brahma Valley College of engineering and research Institute Nashik

---

Abstract - One of the most critical problems in healthcare is predicting the likelihood of hospital readmission in case of chronic diseases such as diabetes to be able to allocate necessary resources such as beds, rooms, specialists, and medical staff, for an acceptable quality of service. Unfortunately relatively few research studies in the literature attempted to tackle this problem; the majority of the research studies are concerned with predicting the likelihood of the diseases themselves. Numerous machine learning techniques are suitable for prediction. Nevertheless, there is also shortage in adequate comparative studies that specify the most suitable techniques for the prediction process. Towards this goal, this paper presents a comparative study among five common techniques in theliterature for predicting the likelihood of hospital readmission in case of diabetic patients. Those techniques are logistic regression (LR) analysis, multi-layer perceptron (MLP), Naïve Bayesian (NB) classifier, decision tree, and support vector machine (SVM).The comparative study is based on realistic data gathered from a number of hospitals in the United States. The comparative study revealed that SVM showed best performance, while the NB classifier and LR analysis were the worst.

Key Words: Decision tree; hospital readmission; logistic regression; machine learning; multi-layer perceptron; Naïve Bayesian classifier; support vector machines

1.INTRODUCTION

Nowadays, numerous chronic diseases, such as diabetes, are widespread in the world; and the number of patients is increasing continuously. The estimated number of diabeticadults in 2014 was 422 million versus 108 million in 1980 [1]. Such patients visit hospitals frequently, requiring continuous preparation for ensuring the availability of required resources including hospital beds, rooms, and enough medical staff for anacceptable quality of service. Accordingly, predicting the likelihood of readmission of a given patient is of ultimate importance. In fact readmission during a one month period (30days) of discharge indicates "a high-priority healthcare qualitymeasure" and the goal is to address this problem [2].

Machine learning, which is one of the most important branches of artificial intelligence, provides methods and techniques for learning from experience [3]. Researchers oftenuse it for complex statistical analysis tasks [4]. It is a wide multidisciplinary domain which is based on numerous

disciplines including, but not limited to, data processing, statistics, algebra, knowledge analytics, information theory,control theory, biology, statistics, cognitive science, philosophy, and complexity of computations. This field plays an important role in term of discovering valuable knowledge from databases which could contain records of supplymaintenance, medical records, financial transactions, applications of loans, etc. [5].

As indicated in Fig. 1, machine learning techniques can be broadly classified into three main categories [3]. Supervised learning techniques involve learning from training data, guidedby the data scientist. There are two basic types of learning missions: classification and regression. Models of classificationattempt to predict distinguished classes, such as blood groups, while models of regression prognosticate numerical values [3]. In unsupervised learning, on the other hand, the system could attempt to find hidden data patterns, associations among features or variables, or data trends [3], [4]. The main objective of

unsupervised learning is the ability to specify hidden structures or data distributions without being subject to supervision or the prior categorization of the training data [6]. Finally, in reinforcement learning the system attempts to learn through interactions (trial and error) with a dynamic environment. During this learning mode, the computer programprovides access to a dynamic environment in order to perform a specific objective. It is worth noting that in this case, the system does not have prior

knowledge regarding the environment‟s behavior, and the only way to figure it out is through trial and error [3], [7], [8].

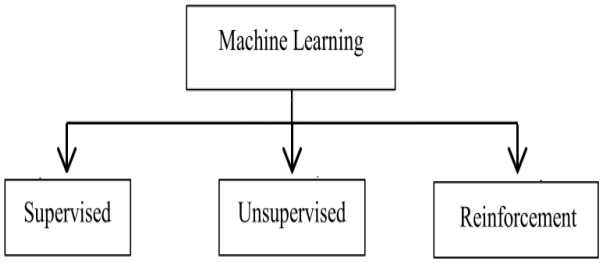
According to Kaelbling et al., the term healthcare informatics refers to the combination between machinelearning and healthcare with the purpose of specifying interest patterns [9]. In addition to this, it has the potential for establishing a good relationship between patients and doctors, and minimizing the increasing cost of healthcare [10]. The goalof this paper is to apply machine learning techniques, and specifically prediction techniques, for predicting the likelihoodof readmission of patients to hospitals. This problem hasn‟t been adequately addressed in the literature. In fact most research efforts are oriented towards prediction of diseases. Machine learning includes numerous analytic techniques for prediction and the literature lacks adequate comparative studies

that assist in selecting a suitable technique for this purpose. Our research is based on a large data set collected by numerousUnited States hospitals [11], [12]. In short, this paper has two main contributions as follows:

Analyzing five most common machine learning techniques for prediction and providing a comparative study among them.

Addressing the problem of patient readmission to hospitals, since it has been rarely addressed by researchers.

Organization of the rest of the paper is as follows: First, wepresent background about the machine learning techniquesconsidered in this research. This is followed by related work tohighlight the contributions of the paper. We then present our methodology and discuss the results of the experiments. Finally, we



sum up this work via a conclusion and discussion of possible future work.

This section discusses the five basic machine learning techniques employed in this research study.

A. *Logistic Regression Analysis*

Regression is a statistical notion that can be used to identifythe relationship weight between one variable called the dependent variable and a group of other changeable variables denoted as the independent variables. Logistic regression (LR)is a non-linear regression model, used to estimate the likelihood that an event will occur as a function of others [13].

B. *Artificial Neural Network*

An Artificial Neural Network (ANN) is a computational model which attempts to emulate the human brain parallel processing nature. An ANN is a network of strongly interconnected processing elements (neurons), which operate inparallel [14] inspired by the biological nervous systems [15]. ANNs are broadly used in many researches because they are capable of modeling non-linear systems, where relationships among variables are either unknown of quite complicated [14].An example of an ANN is the Multi-Layer

Perceptron (MLP), which is typically formed of three layers of neurons (input layer, output layer, and hidden layer) and its neurons use non- linear functions for data processing [16].

C. *Naïve Bayesian Classifier*

Naïve Bayesian (NB) classifier relies on applying Bayes‟ theorem to estimate the most probable membership of a given event in one of a set of possible classes. It is described as being naïve, since it assumes independence among variables used inthe classification process [15], [17], [18].

D. *Support Vector Machine*

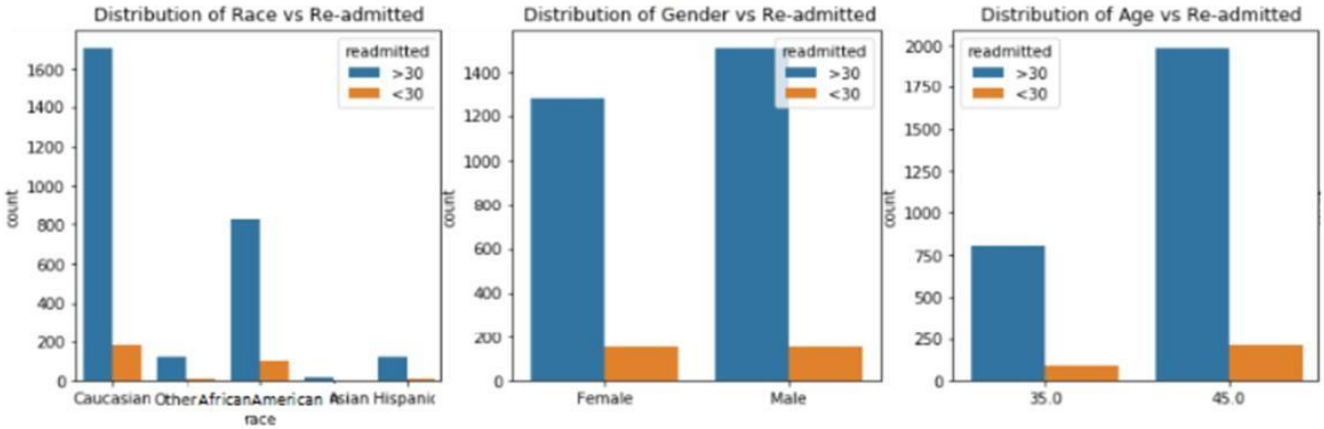
Support vector machines (SVMs) are supervised learning models, which can be applied for classification analysis and regression analysis. They have been proposed by Vapnik in 1995. They can perform both linear and non-linear classification tasks [5], [12], [17], [19]. Decision trees are one of the most famous techniques in machine learning. A decision tree relies on classification by using attribute values for making decisions. In general, a decision tree is a group of nodes, leaves, a root and branches [20]. Many algorithms have been proposed in the literature forimplementing decision trees. One important algorithm is CART (Classification and Regression Tree). It is used for dealing with continuous and categorical variables [8], [21].

RELATED WORK:

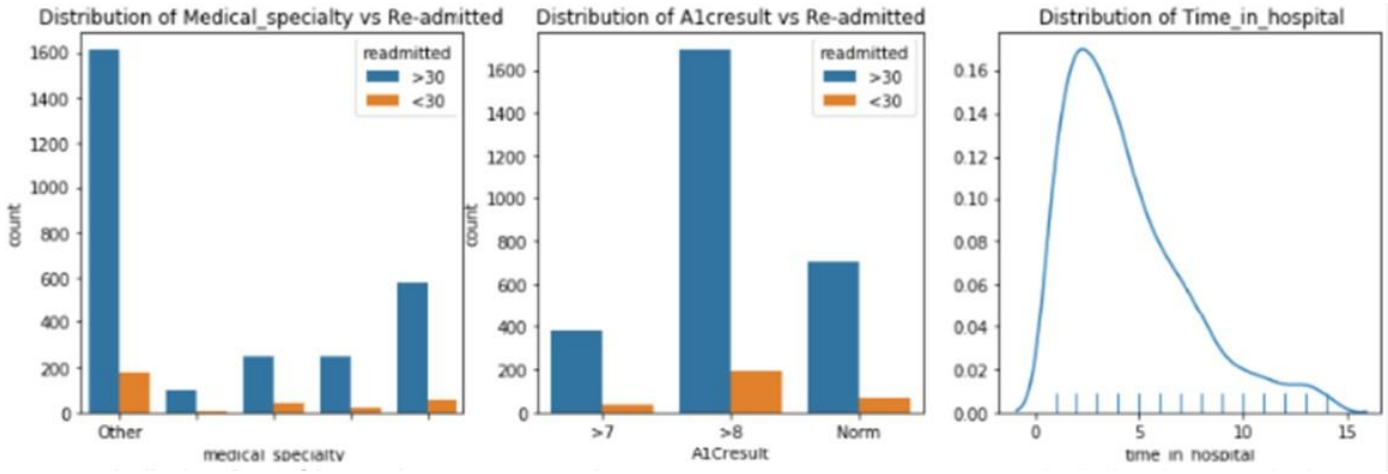
Many researchers attempted to use machine learning techniques in healthcare problems other than hospital readmission likelihood prediction. For example, Arun and Sittidech used K-Nearest Neighbor (KNN), NB, and decision trees with boosting, bagging, and ensemble learning in diabetesclassification. Their experiments confirmed that the highest accuracy is obtained by applying bagging with decision trees [22]. On the other hand, Perveen et al. attempted to improve the performance of such algorithms using AdaBoost. The evaluation of experimental outcomes showed that AdaBoost had better performance in comparison to bagging [23]. Orabi et al. [24] suggested integrating regression with randomization for predicting diabetes cases according to age, with an accuracyof 84%. Other researchers proposed building a predictive model using three machine learning techniques, which arerandom forests (RFs), LR, and SVMs; for predicting diabetes in Indians females, in addition to the factors causing diabetes. Their comparative study concluded that RFs had the best performance among the others [25].

Relatively few research studies addressed the problem of hospital readmission likelihood prediction. For example, Strack et al. used statistical models for this purpose [12]. Other researchers focused on comparing different machine learning techniques for addressing this problem. For example, Kerexeta

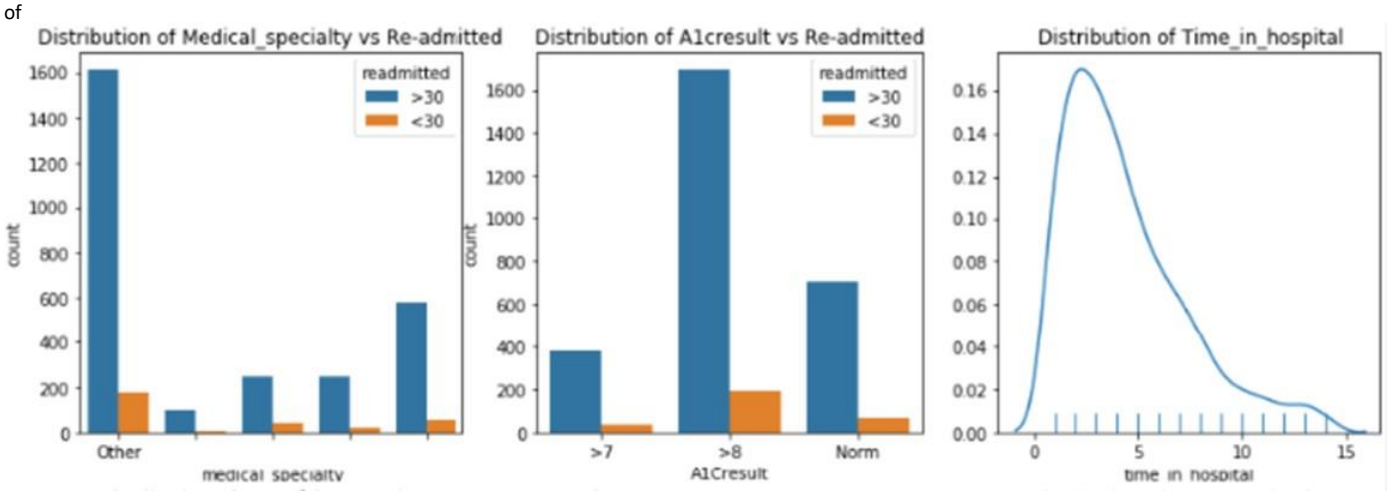
[26] proposed two approaches. In the first, they combinedsupervised and unsupervised classification techniques, while inthe latter, they combined NB and decision trees. They showed that the former approach had a better performance incomparison to the latter in terms of readmission predictioTo sum up, relatively few research efforts in healthcare areconcerned with the problem of prediction of hospital readmission likelihood. Additionally, there is a shortage of adequate comparative studies for comparing machine learning techniques used for prediction. Hence, this paper attempts to tackle those two problems by comparing five common machinelearning techniques for tackling the problem of hospital readmission

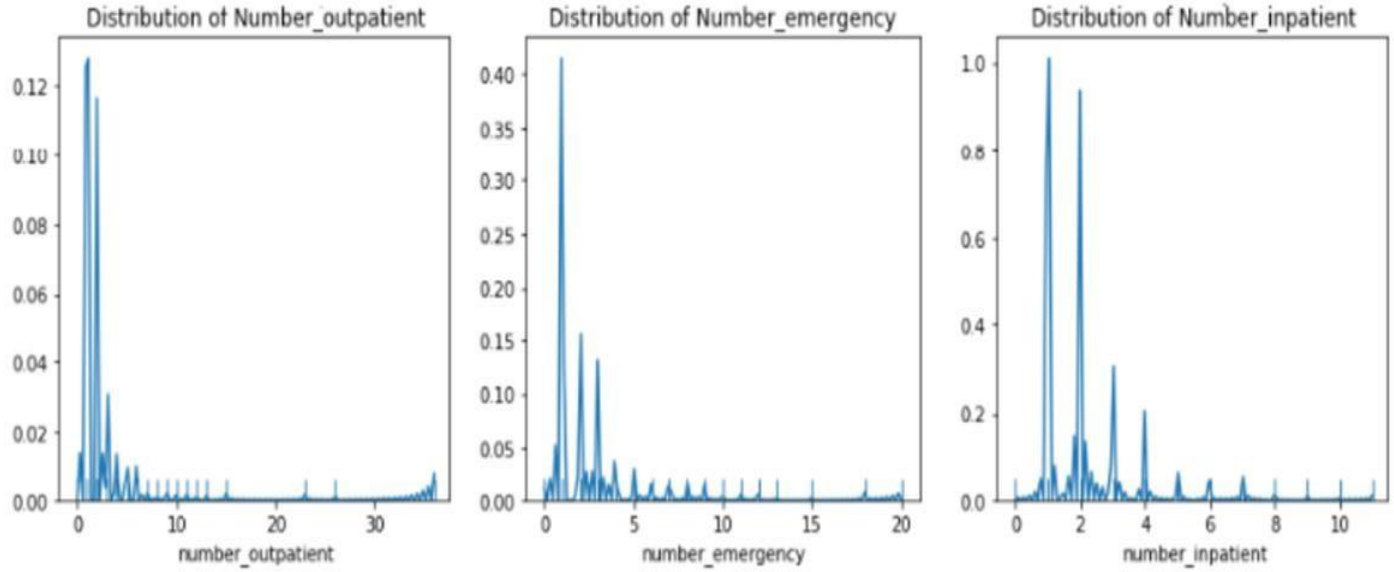


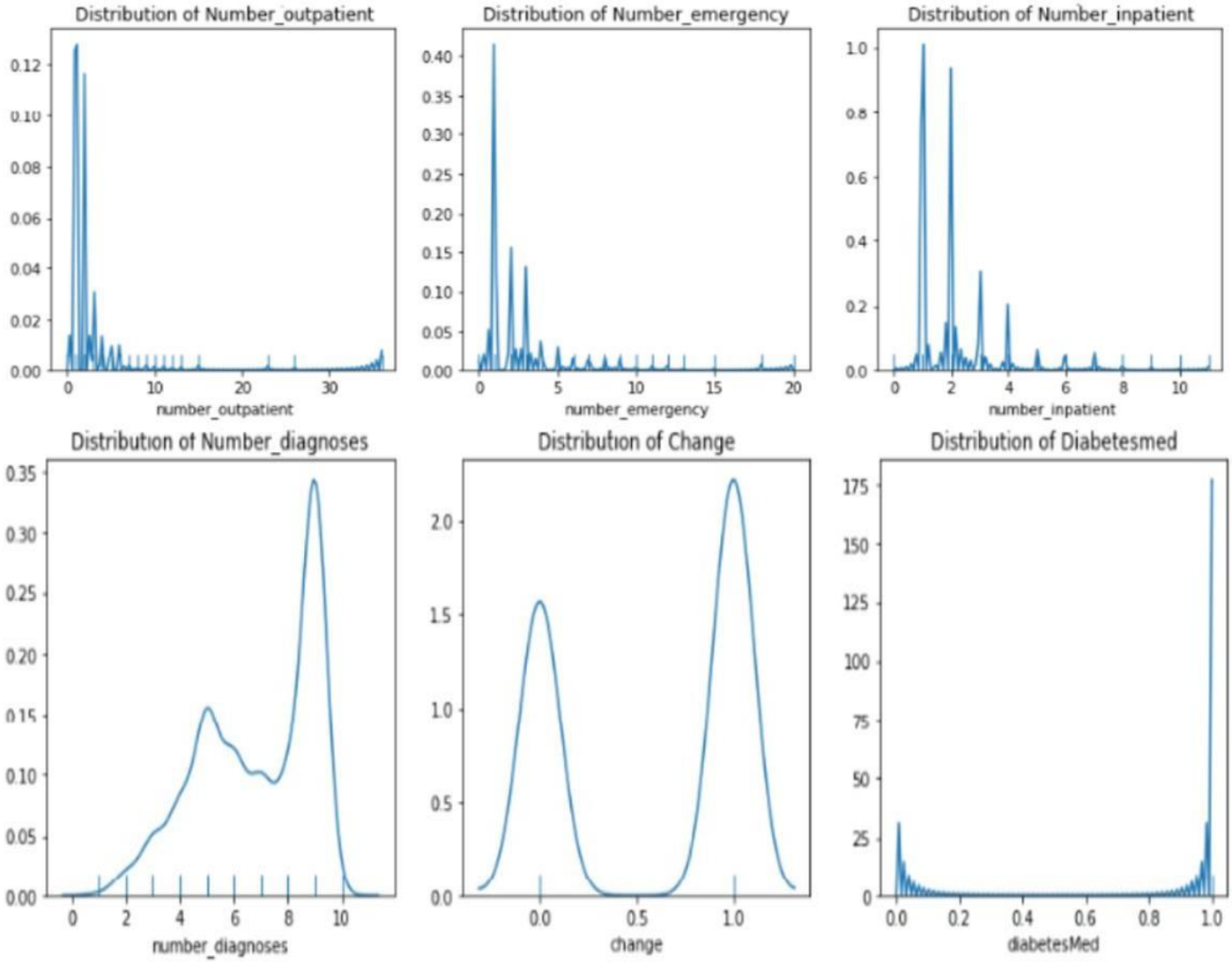
likelihood prediction based on real data.


Distribution of Medical_specialty vs Re-admitted
readmitted
>30
<30
count
Other
medical specialty
Distribution of A1cresult vs Re-admitted
>7
>8
Norm
A1Cresult
Distribution of Time_in_hospital
time in hospital

of


Distribution of Medical_specialty vs Re-admitted
readmitted
>30
<30
count
Other
medical specialty
Distribution of A1cresult vs Re-admitted
>7
>8
Norm
A1Cresult
Distribution of Time_in_hospital
time in hospital


Distribution of Number_outpatient
number_outpatient
Distribution of Number_emergency
number_emergency
Distribution of Number_inpatient
number_inpatient



METHODOLOGY:

Before starting the comparative study, it is important to understand the data, perform preprocessing if necessary, and select features appropriate for the experiments as depicted in Fig. 2. Those tasks are explained below. It is worth noting thatall the experiments were conducted using Python.

A. Data Preparation

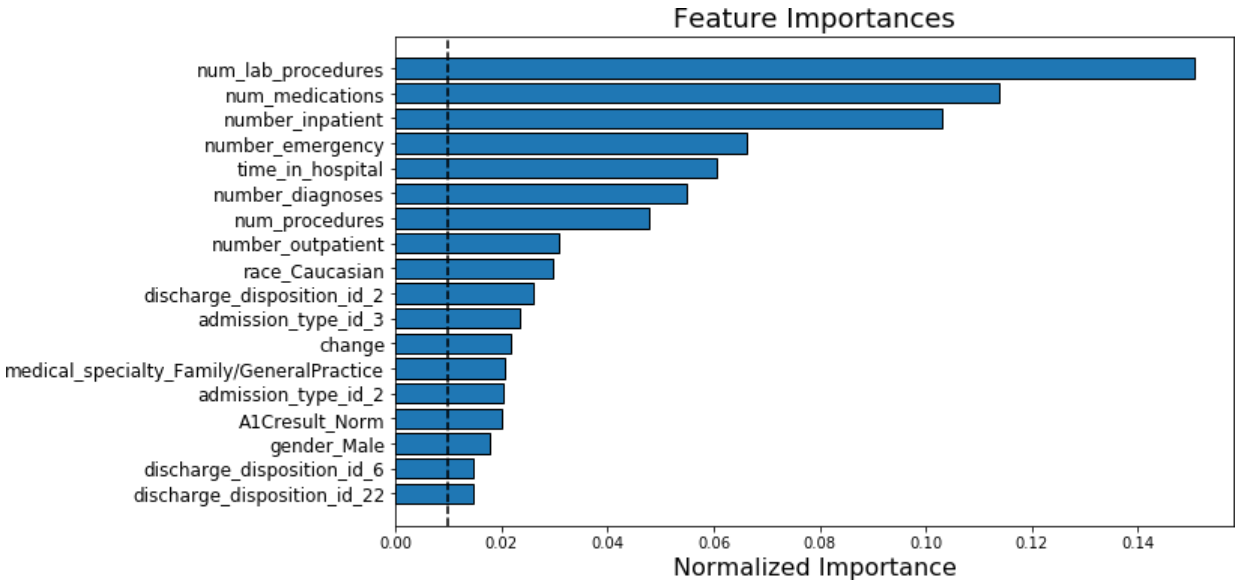
1) Understanding data: In this study, we exploited a sample of a diabetes patients‟ dataset, which has been extracted from many hospitals in the United States [11], [12]. This dataset includes 3090 instances in the age range of 30-50 and with 18 attributes. Table I depicts the variables of the dataset together with their descriptions. The scientific meanings of those variables are beyond the scope of this paper.Fig. 3 through 8 depict the distribution of those features.

2) Data pre-processing: This is a very important stage which includes data transformation and cleaning. In data transformation, some variables were transformed from categorical to binary (0/1) such as (Change, DM, G, and A). Some other variables were transformed from integer to string such as AS, DI, and AS. In data cleaning, some values of categorical data were missing and had to be accounted for. Forthis purpose, we employed imputation (substitution) via the mode of the categorical data.

B. Feature Selection: In this step, we perform feature selection for dimensionality reduction. In other words, weselect the most relevant features. In this study, towards this goal, we assessed the impact of variables on our target. This helped us eliminate variables with low importance. Features which have

high influence on accuracy are the most important [27]. We used the Gradient Boosting technique [28] for categorical features. Table II demonstrates the average

Variable	Variables Abbreviation	Data type
Race	R	Categorical
Gender	G	Categorical
Age	A	Categorical
Admission type Id	AT	Integer
Discharge disposition Id	DI	Integer
Admission source Id	AS	Integer
Medical specialty	MS	Categorical
A1Cresult	A1Cresult	Categorical
Time in hospital	TH	Integer
Number of lab procedures	NL	Integer
Number of procedures	NP	Integer
Number of medications	NM	Integer
Number of outpatient	NO	Integer
Number of emergency	NE	Integer
Number of inpatient	NI	Integer
Number of diagnosis	ND	Integer
Change	Change	Categorical
DiabetesMed	DM	Categorical
DM	0.008867	Rejected



Naïve bayesian classifier: A NB model was created using Gaussian Naive Bayes, which assumes that the attributesfollow a natural distribution.

1) Multi-Layer perceptron: We built a MLP network using 18 inputs. The number of neurons in a hidden layer was

5. The function of the neurons was stochastic gradient descent.The maximum number of iterations was 300, and the two outputs were (readmitted < 30 and readmitted > 30). Table V illustrates that result of MLP weight matrix after training.

RESULTS AND DISCUSSION:

This work utilized various performance measures to compare the studied techniques [31]. Specifically, we relied onaccuracy, recall, precision, and F1 scores for this purpose. Those parameters are defined in terms of the true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) as indicated in equations (1) through (4). TPs are cases in which we predicted yes (they will be readmitted ina month period), and they were really readmitted. TNs are cases in which we predicted no, and they were not readmitted. On the other hand, FPs are cases in which we predicted yes, but they were not actually readmitted; Type I error. Finally, FNs are cases in which we predicted no, but they were actuallyreadmitted; Type II error.

CONCLUSIONS:

This paper presented a comparative study among five machine learning techniques; namely LR, MLP, NB classifier,decision trees, and SVMS; for predicting the likelihood of hospital readmission of diabetes patients. The study relied on real data collected from hospitals in the United States. Based on the study, the SVM provided the best performance. Nevertheless, the study will be extended to compare additional techniques and larger datasets will be considered as well.

REFERENCES:

[1] 1 G. Roglic, "Global report on diabetes.," World Heal. Organ., vol. 58, no. 12, pp. 1-88, 2016.

[2] D. Rubin, K. Donnell-Jackson, R. Jhingan, S. Golden, and A. Paranjape, "Early readmission among patients with diabetes: A qualitative assessment of contributing factors," J. Diabetes Complications, vol. 28, no. 6, pp. 869-873, 2014.

[3] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine Learning and Data Mining Methods in Diabetes Research," Comput. Struct. Biotechnol. J., vol. 15, pp. 104-116, 2017.

[4] P. Chowriappa, S. Dua, and Y. Todorov, "Machine Learning in Healthcare Informatics," vol. 56, pp. 1-23, 2014.

[5] T. Mitchell, "Machine learning (mcgraw-hill international editions computer science series)," 1997.

[6] E. Bose and K. Radhakrishnan, "Using Unsupervised Machine Learning to Identify Subgroups among Home Health Patients with Heart Failure Using Telehealth," CIN - Comput. Informatics Nurs., vol. 36, no. 5, pp. 242-248, 2018.

[7] L. Kaelbling, A. Littman, and A. Moore, "Reinforcement learning: A survey," J. Artif. Intell. Res., vol. 4, pp. 237-285, 1996.

[8] K. Shailaja, B. Seetharamulu, and M. Jabbar, "Machine Learning in Healthcare: A Review," 2018 Second Int. Conf. Electron. Commun. Aerosp. Technol., no. Iceca, pp. 910-914, 2018.

[9] J. Davies and J. Gibbons, "Machine Learning and Software Engineering in Health Informatics," in Proceedings of the First International Workshop on Realizing AI Synergies in Software Engineering, 2012, pp. 37-41.

[10] R. Bhardwaj, A. Nambiar, and D. Dutta, "A Study of Machine Learning in Healthcare," Proc. - Int. Comput. Softw. Appl. Conf., vol. 2, pp. 236- 241, 2017.

[11] A. Asuncion and D. Newman, "UCI Machine Learning Repository," 2007. [Online]. Available: https://archive.ics.uci.edu/ml/index.php.

[12] S. B. et al., "Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records," Biomed Res. Int., vol. 2014, 2014.

[13] A. H. Karp, "Using logistic regression to predict customer retention," Proc. Elev. Northeast SAS Users Gr. Conf. http//www. lexjansen. com/nesug/nesug98/solu/p095. pdf, 1998.

[14] F. Amato, A. López, E. M. Peña-Méndez, P. Va?hara, A. Hampl, and J. Havel, "Artificial neural networks in medical diagnosis," J. Appl. Biomed., vol. 11, no. 2, pp. 47-58, 2013.

[15] S. F., "Machine-Learning Techniques for Customer Retention: A Comparative Study," Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 2, pp. 273-281, 2018.

[16] N. Jothi, N. Rashid, and W. Husain, "Data Mining in Healthcare - A Review," Procedia Comput. Sci., vol. 72, pp. 306-313, 2015.

[17] D. Sisodia and D. Sisodia, "Prediction of Diabetes using Classification Algorithms," Procedia Comput. Sci., vol. 132, no. Iccids, pp. 1578-1585, 2018.

[18] A. Hazra, S. Kumar, and A. Gupta, "Study and Analysis of Breast Cancer Cell Detection using Naïve Bayes, SVM and Ensemble Algorithms," Int. J. Comput. Appl., vol. 145, no. 2, pp. 39-45, 2016.

[19] E. Holzschuh, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting*," Reports Prog. Phys., vol. 55, no. 7, pp. 1035-1091, 1992.

[20] R. Sharma, V. Sugumaran, H. Kumar, and M. Amarnath, "A comparative study of naive Bayes classifier and Bayes net classifier for fault diagnosis of roller bearing using sound signal," Int. J. Decis. Support Syst., vol. 1, no. 1, p. 115, 2015.

[21]