(Available on SSRN. SSRN is an open-access online preprint community, owned by Elsevier.)

International Conference on Contents, Computing & Communication (ICCCC-2022)

26th and 27th February,2022

Prediction of Frequent Hospitalisation of Diabetic Patients

Ranjit M. Gawande #1, Dr. Varsha H. Patil *2

*#1 Asst. Prof. Department of Computer Engineering Matoshri College of Engineering & Research Center Nashik (M.S.) ranjitgawande@gmail.com, *2 Vice-Principal & HOD Department of Computer Engineering Matoshri College of Engineering & Research Center Nashik (M.S.) varsha.patil@gmail.com*

Abstract

Type1 and type2 diabetes is a chronic disorder of glucose-insulin metabolism characterized mainly by high blood glucose concentrations. The role of glycated hemoglobin A1c (HbA1c), is to determine the chronic condition of diabetic patients during the last 2-3 months, it is one of the important measures that indicate the average glucose level of the patients. If the clinical treatment history of the patients is made available then it will be helpful for prediction and decision making for the medical staff about the frequent occurrence of patients requiring hospitalization services. Predicting readmissions of patients will help the hospital calculate and manage the quality of patient care. The purpose of the study is to compare the performance of different classifier techniques in predicting frequent admission for diabetic patients, especially those with comorbid conditions. We also compare the performance of the classification model using the performance metric and use customized data preprocessing techniques for precise feature selection to increase the accuracy of classifiers. Among the supervised learning classification approaches, such as Naive Bayes, Logistic regression, Random Forest, and KNN the accuracy of Gradient boosting classifiers accuracy level is significantly more precise than other classifiers in this study for the prediction of readmission of diabetic patients.

Keywords: Diabetes mellitus, Supervised K-Nearest Neighbour (KNN), Logistic regression, Random Forest, Naive Bayes.

1. Introduction. The healthcare system is shifting to value-based care, and the Centers for Medicare & Medicaid Services (CMS), which is part of the New York State Department of Health and Human Services (HHS), has developed a number of programs to improve the quality of exact patient treatment. The identification of frequent hospitalization of diabetes patients [3] is one of these programs. The goal is to cut reimbursement to hospitals that have higher-than-average readmission rates. One answer to this problem is to develop interventions that provide extra help to patients who are at a higher risk of readmission. But how do we find out who these people are? We can utilize data science predictive modeling to assist prioritize patients.

It is observed that the patients having diabetes are having a greater risk of frequent hospitalization. Diabetes is a clinical condition that is caused due to inefficient management of insulin levels in the blood. As per the WHO [6], approximately 1 in 10 patients are suffering from this disease. Patients with diabetes have double the chances of being frequently hospitalized than the common population. With this objective, we will focus on predicting frequent readmission for patients with diabetes.

Using input data from the UCI machine learning repository, we employed Machine Learning using Scikit-Learn methods to categorize hospital readmissions of diabetes patients in this study [8]. This study's findings show the best performance in the classification of hospital readmissions under a range of trial settings. The purpose of this research is to find the best algorithm for detecting diabetes patients' hospital readmissions, as well as the optimal combination of preprocessing techniques. In the future, researchers may be able to create a suggestion system for diabetic patients' treatment programs.

1.1 Related Work

The ability of machines to manage data is the subject of machine learning [17]. The purpose of Machine Learning is to construct a system that can learn its own patterns without the need for human involvement, based on a training test. Machine Learning has applications in a variety of disciplines, including education [18], gaming [19], and this study uses it in medicine. Supervised Learning, Unsupervised Learning, and Reinforcement Learning are the three types of learning methods used in Machine Learning. Supervised Learning is a systematic learning method for labeling test data based on the model discovered by learning from the training data. Unsupervised learning, unlike supervised learning, is an unstructured learning method that uses groups of data to form classes rather than labels. Reinforcement Learning, on the other hand, entails the system learning something by doing an action and observing the results [20].

Essentially, machine learning works by learning from examples in the same manner that people do and then answering a related question. This learning process in the domain of ML makes use of data, which is referred to as the dataset train. Machine Learning, unlike static packages, was intended to produce programs that can learn on their own. Regression, clustering, and classification are examples of problems that can be tackled with Machine Learning. Classification is a method of grouping data that has been determined by its classification. Sklearn, an open-source machine learning package for the Python computer language [21], is used in this research categorization method. It can handle a variety of classification, regression, and clustering methods, such as support vector machines, random forests, gradient boosting, and k-means, and it's built to work with the Python libraries NumPy and SciPy.

2. Proposed system for Prediction of Frequent Hospitalisation of Diabetic Patients

The basic objective of this study is to Predict if a patient with diabetes will be readmitted to the hospital within 30 days. In this study, we build a model for predicting readmission using sklearn incorporated with Python. The steps involved in model building and data preprocessing are mentioned below.

1. Data exploration
2. Feature selection
3. Building training data then test samples data
4. Model selection
5. Model evaluation
6. Validate the results

The initial data collected from the UCI Machine Learning repository found that there are many missing observations present in the mentioned dataset. The system architecture (Figure 1.1) shows the phases involved in the proposed system. In the preprocessing phase, we carried out dimensionality reduction techniques by considering the features which contributed more to achieving our aim of classifying future records of diabetic patients.

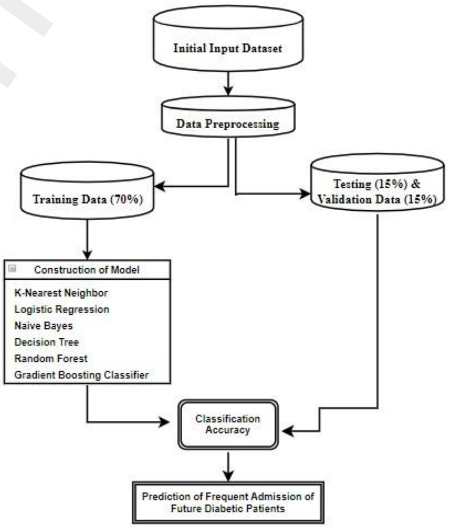


Figure 1.1 System Architecture of Prediction of Frequent Hospitalisation of Diabetic Patients

2.1 Feature Engineering

To identify the important features which contributed to building the predictive model we will add a new variable to each data frame. This mechanism enables us to keep track of which columns of the data frame will be a part of the predictive model features [9]. While experimenting with data we categorize the features into three types such as numerical features, categorical features, and extra features. In the preprocessing stage, we replaced the missing observations with a question mark.

i) Numerical Features

Any numeric real value that gives significant information is useful for determining the feature of the class. Most of the time these features do not need any modification. The easiest way to find missingness in numerical features is by using the panda function IsNull( ) followed by sum( ) which will return a count of missing observations[13].

ii) Categorical Features

The third type of feature is categorical variables, which are discrete in nature and belong to a finite set of categories. Categorical variables are non-numeric data such as race and gender. The simplest way to turn these non-numerical data into variables is to use a technique known as one-hot encoding which simply adds another category type for the unknown by combining the ‘fillna’ function with the ‘get_dummies’ function provided by pandas[18].

iii) Extra Features

There are many data values that are missing in nature. The presence of a variable might be predictive regardless of the value. The features like age and weight need to be converted into categorical data types. The order of data is natural so as to make more sense of the data.[12] To map age from 0 to 90 by 10s for the numerical data. The breakdown of the features is 8 numerical features, 133 categorical features, and 2 extra features

3. Building Training Data and Test Samples

In machine learning, the training and testing process for the classification of biological datasets is critical. The procedures that should be employed at each step should be carefully chosen by the researcher [8]. We have investigated our data and developed features from categorical data since the data preprocessing step. It's now time to separate our data. The goal of separating the data is to see how well the model performs on data that hasn't been seen before. As a result, the samples have been divided into two categories: training samples and test samples. Samples are utilized to train the model in the training phase, whereas in the test phase, samples are kept apart from all decisions and used to assess the model's overall performance. We will divide the study into three parts: 70 percent training, 15 percent validation, and 15 percent testing. All positive samples will be stacked on top of each other during validation and testing to ensure that they originate from similar distributions[17].

By using pandas.DataFrame.drop() method one can drop/remove/delete rows from DataFrame.Thus the rows that were not part of the sample get removed and this same idea is used to get the training data[14]. At this stage, we can analyze what percent of our groups are hospitalized within 30 days. This is known as prevalence. Ideally, all three groups would have a similar prevalence. From the sklearn, the package called pickle can be used to create a scaler for the test data [13] further it can transform into data matrices. At this stage, we will not check the performance until the model selection.

3.1 Model Selection: Baseline models

It is observed that approximately 80 to 90% of the efforts and time are spent on preprocessing of the dataset [9]. In this study, we compare the performance of the few machine learning models using default hyperparameters. Such as K-nearest neighbors, Logistic regression, Naive Bayes, Decision Tree, Random forest, and Gradient boosting classifier

i) K nearest neighbors (KNN)

One of the most basic machine learning models is KNN. The model looks at the k nearest data points for a given sample point and calculates the likelihood by counting the number of positive labels divided by K. The disadvantage of this model is that it is sensitive to K and that it takes a long time to evaluate if the number of trained samples is big [14]. From scikit-learn, one can import the K neighbors classifier to evaluate the model. The results obtained from all the selected models are summarised in the result analysis section.

ii) Logistic Regression

Fits a linear decision border between positive and negative samples in a typical machine learning model. The probability of the positive class is then calculated using a sigmoid function applied to this linear function. Logistic regression is an appropriate model to utilize when the characteristics are linearly separable. The advantage of logistic regression is that the model is interpretable [12], which means we can see which features are important in predicting positive or negative outcomes. One thing to keep in mind is that the modeling is sensitive to feature scaling, which is why we scaled the features before. We can perform the logistic regression by importing the logistic regression from the sklearn library of a linear model.

iii) Naive Bayes

The Naive Bayes is another supervised learning model occasionally used in machine learning. In Naive Bayes, we utilize Bayes Rule to calculate the probabilities. The “naive” part of this model is that it assumes all the features are independent [10] for solving the classification problem. This can be implemented using GaussianNB from the sklearn library of naive Bayes.

iv) Decision Tree

To get all the possible solutions to the problem at hand, it produces probable decision conditions with graphical representation. A decision tree is another class of popular machine learning models. Essentially, in this decision tree method, we utilize the methodology to divide the samples in a form of questions. Each query necessitates the analysis and splitting of samples that have a certain variable greater than a threshold. As a result, the final prediction is the percentage of positive samples in the tree's last leaf (final split). Machine learning is used to select the variable and threshold to employ at each split in this method. Tree-based approaches have the advantage of not assuming data structure and being able to capture nonlinear effects when the tree is deep enough [15].

v) Random Forest

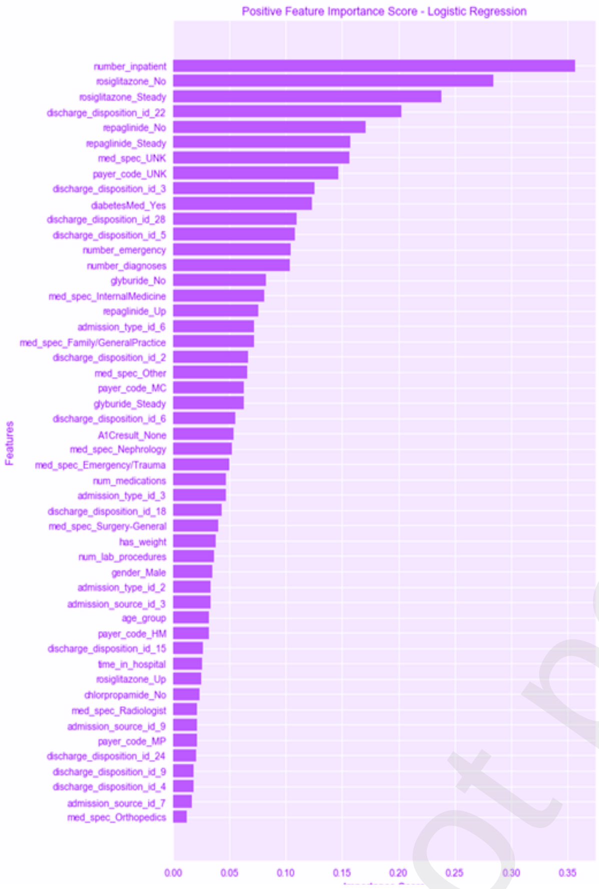
To solve the classification and regression problem a supervised algorithm random forest was used. Random forests reduce the overfitting problem of data by considering the majority of vote concept. Trees are generated and results are collected using random forest models. A random set of samples and a random number of attributes in each tree are used to de-correlate trees in a forest. Random forests outperform decision trees in most circumstances because they can generalize more easily [20].

vi) Gradient Boosting Classifier

Boosting is an approach to improving the performance of decision trees using machine learning techniques in regression and classification tasks. This method involves creating a bunch of shallow trees that will be used to correct errors in previously trained trees. The classifier that uses this technique along with a gradient descent algorithm (to control the learning rate) is known as a gradient boosting classifier. This “boosting” process continues iteratively, with the tree depth, learning rate, and a number of trees were optimized using repeated cross-validation [23]. To fit the gradient boosting classifier, we can use the sklearn.ensemble import GradientBoostingClassifier

4. Results

The diabetic patient data utilized in the study was sourced from the UCI Machine Learning Repository[22]. The data on these diabetic patients is based on patient data from 130 diabetes care clinics in 130 US hospitals that are linked to other networks throughout a ten-year period (1999-2008). There are 50 attributes and 101,776 instances in this collection. After loading the data it is observed that around 11% of the population is rehospitalized. This represented an imbalanced classification problem. To treat this skewed data it is necessary to identify types of attributes that are present in the dataset. The dataset consists of a mixed type of categorical (non-numeric) and numerical data. The parameters such as encounter_id and patient_nbr: are just identifiers and not useful parameters. Similarly, age and weight were found to be categorical in this data set. The numerical variables such as admission_type_id,discharge_disposition_id,admission_source_id are having integer data type. By observation, we ignore the parameters which are having single data instances such as examine parameters, so we will not use these variables. To reduce the dimension of the dataset we group the ICD codes by numbering the diagnostic criteria[23].To understand which features are more contributing to frequent hospitalization of diabetic patients needs to be analyzed. Generally in Logistic Regression or Random Forests, this can be investigated. Figure 1.2 shows the importance of the positive score of a top 50 features using logistic regression. Similarly, figure 1.3 gives us more insight into 50 negative coefficients. This feature analysis helps to identify new feature ideas which help to understand high bias and high variance. The obtained list of the top 50 features will be used further for feature reduction thus enabling the high variance. This dimensionality reduction technique provides a more robust prediction model. After looking at these graphs, it's evident that a few fresh observations are needed to collect new data on several key traits. The most essential variable in both models, for example, is the number of inpatients, which is the number of inpatient visits in the previous year. This indicates that patients who have been in the hospital in the previous year are more likely to be readmitted. This drives them to seek additional information about their previous admissions. Discharge disposition id 22 is another example, which is used when a patient is discharged to a rehabilitation institution. We chose the gradient boosting classifier that performed the best on the validation set for our study. The advantage of a gradient boosting classifier is that there is no need to train the best classifier every time for the new predictions. The package pickle has been used to save the classifier.



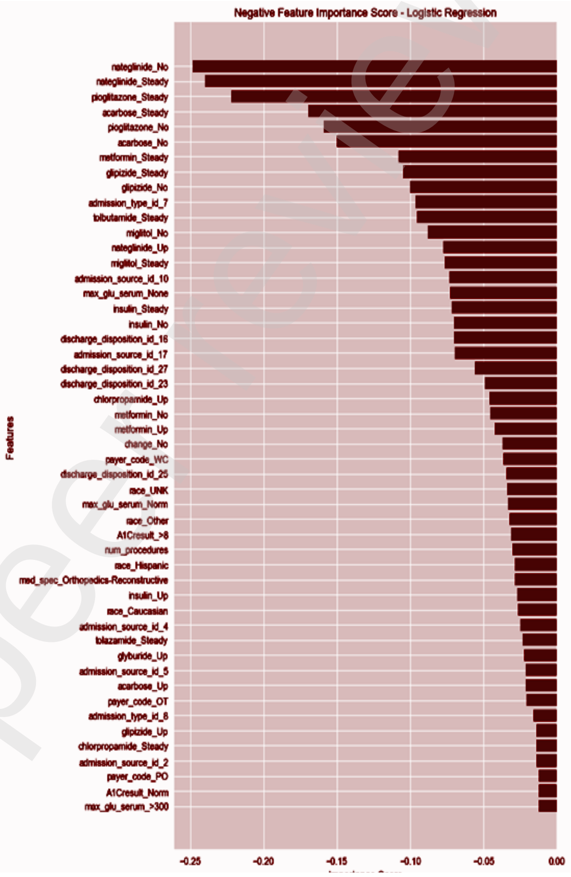
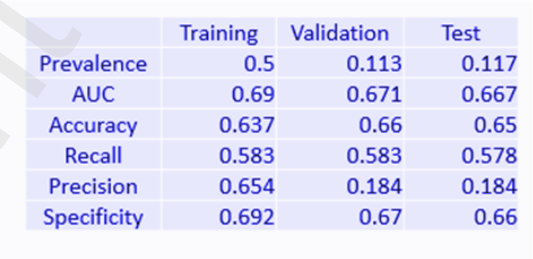


Figure 1 Score of Positive Features

Figure 2 Score of Negative Features

4.1 Model Evaluation



	Training	Validation	Test
Prevalence	0.5	0.113	0.117
AUC	0.69	0.671	0.667
Accuracy	0.637	0.66	0.65
Recall	0.583	0.583	0.578
Precision	0.654	0.184	0.184
Specificity	0.692	0.67	0.66

Table1. Evaluation of Gradient Boosting Classifier

After evaluation, it is found that the gradient boosting classifiers' performance is much better. We also carried out the evaluation based on the performance using the test set which is shown in the table1

Conclusion

As a result of this research, we developed a machine learning model that can predict which diabetic patients are most likely to be readmitted within 30 days. The model was a gradient boosting classifier with hyperparameters that were optimized. The findings of our study can be used to design a treatment recommendation system for diabetes patients in the future.

References

1) Maniruzzaman, M. et al. Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm. Comput. Methods Programs Biomed. 152, 23–34 (2017).

2) Bonaccorso, by G. Machine learning algorithms reference guide for popular algorithms for data science and machine learning.

3) Jiawei Han, M. K. Data Mining Concepts and Techniques.

4) Singh, J. Centers for Disease Control and Prevention. Indian Journal of Pharmacology vol. 36 268–269 https://www.cdc.gov/brfss/ (2004).

5) Wu, H., Yang, S., Huang, Z., He, J. & Wang, X. Type 2 diabetes mellitus prediction model based on data mining. Informatics Med. Unlocked 10, 100–107 (2018).

6) Ramos, M. et al. Diagnosis : A Case : Control Study (FS03-04-19). 664.

7) Choubey, D. K., Kumar, P., Tripathi, S. & Kumar, S. Performance evaluation of classification methods with PCA and PSO for diabetes. Netw. Model. Anal. Heal. Informatics Bioinforma. 9, (2020).

8) Lukmanto, R. B. & Irwansyah, E. The Early Detection of Diabetes Mellitus (DM) Using Fuzzy Hierarchical Model. Procedia Comput. Sci. 59, 312–319 (2015).

9) Gomes Filho, E., Pinheiro, P. R., Pinheiro, M. C. D., Nunes, L. C. & Gomes, L. B. G. Heterogeneous Methodology to Support the Early Diagnosis of Gestational Diabetes. IEEE Access 7, 67190–67199 (2019).

10) Utz, B. et al. Detection and initial management of gestational diabetes through primary health care services in Morocco: An effectiveness-implementation trial. PLoS One 13, 1–17 (2018).

11) Dryad. Data from: Genetic determinants for gestational diabetes mellitus and related metabolic traits in Mexican women. https://datadryad.org/stash/dataset/doi:10.5061/dryad.kq0k2.

12) Zheng, T. et al. A simple model to predict risk of gestational diabetes mellitus from 8 to 20 weeks of gestation in Chinese women. BMC Pregnancy Childbirth 19, 1–10 (2019).

13) Qiu, H. et al. Electronic Health Record Driven Prediction for Gestational Diabetes Mellitus in Early Pregnancy. Sci. Rep. 7, 1–13 (2017).

14) Sajeev, S. et al. Deep Learning to Improve Heart Disease Risk Prediction. 1, 96–103 (2019).

15) Rahman, R. A., Aziz, N. S. A., Kassim, M. & Yusof, M. I. IoT-based personal health care monitoring device for diabetic patients. ISCAIE 2017 - 2017 IEEE Symposium on Computer Applications and Industrial Electronics (2017). doi:10.1109/ISCAIE.2017.8074971.

16) Filho, E. G. et al. Support to early diagnosis of gestational diabetes aided by Bayesian networks. Adv. Intell. Syst. Comput. 985, 360–369 (2019).

17) Krishnan, D. R. et al. Evaluation of predisposing factors of Diabetes Mellitus post Gestational Diabetes Mellitus using Machine Learning Techniques. in 2019 IEEE Student Conference on Research and Development, SCOReD 2019 81–85 (Institute of Electrical and Electronics Engineers Inc., 2019). doi:10.1109/SCORED.2019.8896323.

18) Itani, S., Lecron, F. & Fortemps, P. Specifics of medical data mining for diagnosis aid: A survey. Expert Syst. Appl. 118, 300–314 (2019).

19) Aiello, E. M., Toffanin, C., Messori, M., Cobelli, C. & Magni, L. Postprandial glucose regulation via KNN meal classification in type 1 diabetes. IEEE Control Syst. Lett. 3, 230–235 (2019).

20) Li, Y., Li, H. & Yao, H. Analysis and Study of Diabetes Follow-Up Data Using a Data-Mining-Based Approach in New Urban Area of Urumqi, Xinjiang, China, 2016-2017. Comput. Math. Methods Med. 2018, (2018).

21) Huerta-Chagoya, A. et al. Genetic determinants for gestational diabetes mellitus and related metabolic traits in Mexican women. PLoS One 10, (2015).

22) Douali, N., Dollon, J. & Jaulent, M. C. Personalized prediction of gestational Diabetes using a clinical decision support system. IEEE Int. Conf. Fuzzy Syst. 2015-Novem, (2015).