# Analysis of Hospital Readmission Rates Based on Haemoglobin Measurements

**Shashank Holla**
Luddy School of Informatics, Computing, and Engineering
Indiana University
Bloomington, Indiana
sholla@iu.edu

**Nagadarshan Nanjundaswamy**
Luddy School of Informatics, Computing, and Engineering
Indiana University
Bloomington, Indiana
nananjun@iu.edu

**Nikhil Mahadevaswamy**
Luddy School of Informatics, Computing, and Engineering
Indiana University
Bloomington, Indiana
nmahade@iu.edu

**Abstract**

Diabetes is a chronic health condition that affects the majority of the population around the world. Centers for Disease Control and Prevention (CDC), diabetes is a chronic health condition that affects how our body turns food into energy. Very little works have been done on this dataset achieving an accuracy of less than 70%. In this project, we have experimented with 8 different models including two deep learning models. Of all the models, the Gradient Boosting classifier outperforms other classifiers with an accuracy of 58.92% for three-class classification and 64.08% for binary classification

***Keywords—MLP, Gradient Boost, Diabetes, Decision Tree, etc.***

## I. Introduction

Diabetes is one of the most prevalent health conditions around the world. From the statistics available from American Diabetes Association, in the year 2018, nearly 34 million people in the United States (10.5% of the population) had diabetes. Also, there are 1.5 million Americans diagnosed with diabetes every year.

Per the Centers for Disease Control and Prevention (CDC), diabetes is a chronic health condition that affects how our body turns food into energy. The food that we eat is broken down into sugar (glucose) and released into the bloodstream. When the blood sugar goes up, it signals the pancreas to release insulin. Insulin is a hormone that lets the blood sugar be used as energy in our body's cells. With diabetes, the body doesn't make enough insulin or can't use the insulin that is produced. With insufficient insulin, blood sugar stays in the bloodstream which, over time, can cause serious health problems such as heart disease, vision loss, and kidney disease.

This document is divided into 7 sections. Section 1 is the introduction. Section 2 is the literature. In section 3, we describe the dataset. Section 4 briefs about the experiments we made. Section 5 shows the results of the experiments. In Section 6, we have provided the conclusions and further scope of this project. Section 7 has the references.

## II. Literature

In one of the work, data preprocessing was done by dropping 29 features in total, out of which 22 features were related to medications. The main focus of this work was on six major features which were found to have high impact on diabetes patient readmission: number of lab procedures, number of medications administered during the encounter, time spent in hospital, number of procedures other than lab tests, number of diagnoses, and number of inpatient visits. Features with few missing values such as race and gender have also been dropped. This work has been implemented on 3 different machine learning models, out of which the logistic regression classifier stands out with 62% accuracy[1].

Another work has considered one encounter per patient. This is based on the understanding that multiple inpatient visits cannot be considered as statistically independent. It has also removed examples that resulted in either discharge to a hospice or patient death. In this work, the probability of readmission is calculated for cases

where the patient has normal to high blood sugar levels with varying medications. The probability of readmission is then calculated for cases where the patient has normal to high blood sugar levels with varying medications[2]. This is the only work published on this dataset. The rest of them are referenced from the github.com website.

In one another work, features such as age, medicine dosages, and diagnosis classes were compressed into smaller categories. This work has also implemented oversampling techniques such as SMOTE to increase the data volume for the minority class. This work has implemented Adaboost, Logistic Regression out of which the latter has given the highest accuracy of 59% [3]

## III. Dataset

The Diabetes dataset represents 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. It includes over 50 features representing patient and hospital outcomes. The collected data contains attributes such as patient number, race, gender, admission type, time in hospital, the medical specialty of admitting physician, HbA1c test result, diagnosis, number of diabetic medications, and the number of outpatients and inpatients emergency visits in the year before hospitalization.

The diabetes information tabulated are for inpatient diabetic encounters with length of hospital stay between 1 day and at most 14 days. 101, 766 such encounters are identified and 50 such features describing the demographics, diagnoses, diabetic medications, the number of visits are tabulated.

The target feature of interest here is the readmitted state. The possible values of this feature is Not readmitted (54,864) , <30 days of readmission (35,545) and >30 days of readmission (11,357). Other independent features in the dataset includes Admission type (consists of 9 distinct values, for example, emergency, urgent, elective, newborn, and not available), number of lab tests, Admission source (consists of 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital), 24 other features for medications like metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, etc.

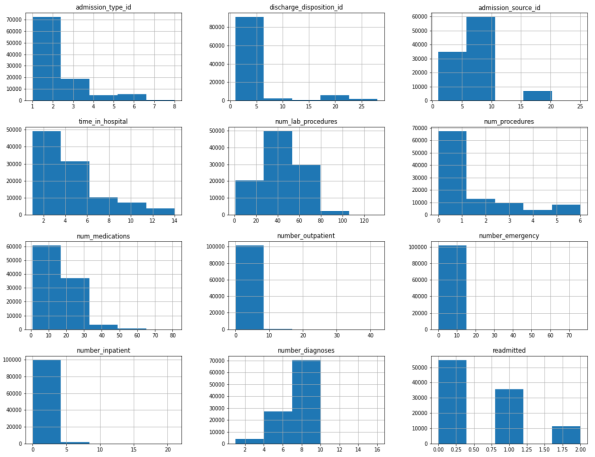


Figure 3.1: Histogram showing distribution of features

## IV. Experiments

The hospital readmission and the problem statement translate to a classification problem. Here we are classifying the given input sample to one of the three classes, which are, (1) the patient never readmits to the hospital, (2) the patient readmits to the hospital less than 30 times in 10 years, (3) the patient readmits to the hospital more than 30 times in 10 years.

The problem statement is also considered as a binary classification problem of classifying patients by- 1) No - the patient was not admitted to the hospital. 2) Yes - the patient was admitted to the hospital.

Patient-specific information such as Patient Number, Encounter ID which do not provide meaningful information for classification is dropped. Diagnosis columns have 700 or more unique categories. Those diagnosis columns have been dropped. Categorical columns are pre-processed with one-hot encoding. Numerical features are normalized with standard scaling while for convolution neural network learning these columns are scaled by considering the

min-max scaler. While splitting the dataset for train/test sets, data has stratified to retain the class distribution.

The preprocessed data is given to the 8 classifiers, which are (1) K Nearest Neighbor, (2) Logistic Regression, (3) Decision Tree, (4) Random Forest, (5) Gradient Boost, (6) Multilevel Perceptron, (7) Ensemble Model (MLP and SVM), (8) Convolutional Neural Network.

Table 5.1: 3-class classification problem

| Sl No. | Model | Train Accuracy | Test Accuracy |
|---|---|---|---|
| 1 | K Nearest Neighbor | 0.6231 | 0.5501 |
| 2 | Logistic Regression | 0.5709 | 0.5759 |
| 3 | Decision Tree | 0.5971 | 0.5762 |
| 4 | Random Forest | 0.9998 | 0.5780 |
| 5 | Gradient Boost | 0.5879 | 0.5892 |
| 6 | Multi-level Perceptron | 0.5888 | 0.5809 |
| 7 | Ensemble Model (MLP and SVM) | 0.5921 | 0.5820 |
| 8 | Convolutional Neural Network | 0.5940 | 0.5828 |

Table 5.2: Binary classification problem

| Sl No. | Model | Train Accuracy | Test Accuracy |
|---|---|---|---|
| 1 | K Nearest Neighbor | 0.6758 | 0.5974 |
| 2 | Logistic Regression | 0.6183 | 0.6245 |
| 3 | Decision Tree | 0.6502 | 0.6303 |
| 4 | Random Forest | 0.9998 | 0.6325 |
| 5 | Gradient Boost | 0.6388 | 0.6408 |
| 6 | Multi-level Perceptron | 0.6505 | 0.6365 |
| 7 | Ensemble Model (MLP and SVM) | 0.6517 | 0.6386 |
| 8 | Convolutional Neural Network | 0.6533 | 0.6393 |

## V. Results

From the experiment runs for three-class classification, Gradient Boost Classifier was observed to have the highest test accuracy of 58.92% as shown in Table 5.1. From the model executions, One of the possible causes for the low accuracy could be the imbalance in the class distribution. Pearson correlation calculation suggests the numerical features to have a very weak correlation with the target feature which could another reason for the low performance.

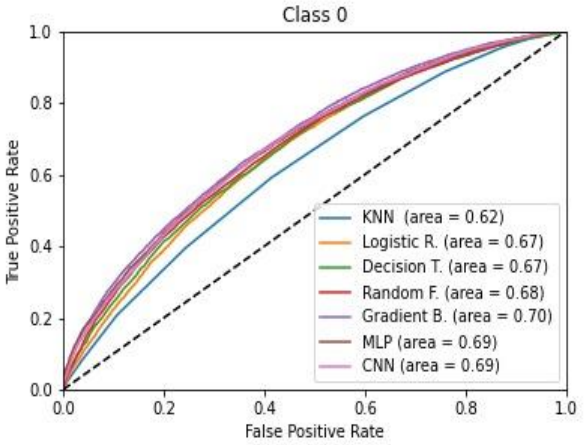


Figure 5.1: ROC-AUC Curve for class 0 for the 3 class classification problem

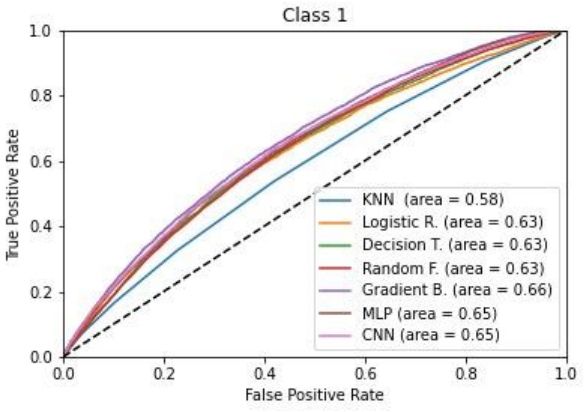


Figure 5.2: ROC-AUC Curve for class 1 for the 3 class classification problem

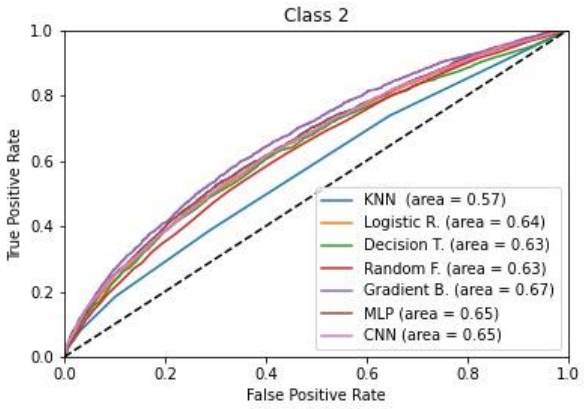


Figure 5.3: ROC-AUC Curve for class 1 for the 3 class classification problem

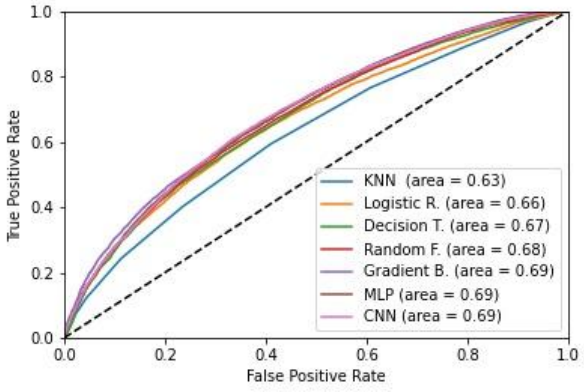


Figure 5.4: ROC-AUC Curve for class 1 for the 3 class classification problem

As we can see in the above figure 5.1, the roc-auc curve shows that the area under the curve for class 0 is highest for the Gradient Boost classifier with auc as 0.70. We can also see that the Gradient boost classifier also outperformed other classifiers with auc of 0.66 and 0.67 for classes 1 and 2 respectively. Even for the binary classification problem Gradient Boost classifier has the highest auc compared to other classifiers. Since Gradient Boost Classifier outperforms other classifiers in all the results, Gradient Boost Classifier is considered as one of the best classifiers for this problem.

## VI. Conclusions And Further Studies

As we can see in the results that Gradient Descent outperforms other classifiers in all the aspects. In this project, we have only considered few deep learning models, this problem can be further tackled with Recurrent Neural Networks. In our experiments, we have implemented a 2 layer convolutional neural network. Further this can be experimented with deeper CNN models. We also came across another algorithm, Image generator for tabular data(IGTD), which can be used to create images with tabular data and further those images can be given as input to the CNN models[4].

## VII. References

[1] https://github.com/swengzju/Predicting-Diabetes-Patient-Readmission/blob/master/Predicting%20Diabetes%20Patient%20Readmission.ipynb

[2] Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore, “Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records,” BioMed Research International, vol. 2014, Article ID 781670, 11 pages, 2014.

[3] https://github.com/angelmanzur/Diabetes_130Hospitals

[4] Zhu, Y., Brettin, T., Xia, F., Partin, A., Shukla, M., Yoo, H., Evrard, Y.A., Doroshow, J.H. and Stevens, R.L., 2021. Converting tabular data into images for deep learning with convolutional neural networks. Scientific reports, 11(1), pp.1-11.