International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

# Classifying High Risk Diabetic Patients using Supervised Machine Learning Models

**Ritambhara Jha**

Email: *jha.ritambhara[at]gmail.com*

**Abstract:** ***Hospital re - admissions are a major healthcare concern, primarily in terms of the quality services delivered to hospitalized patients and the accompanying healthcare expenditures. In 2018, 20 percent of adult hospital re - admissions were associated with four conditions at index admission: septicemia, heart failure, diabetes, and chronic obstructive pulmonary disease (COPD) [10]. Our goal is to examine these studies and focus on the frequency of re - admissions, their causes and their usefulness as a measure of care quality. This paper compares four popular approaches in the literature to classify the diabetic patients into Re - admissible or Non - Re - admissible.***

**Keywords:** Classification ML models, Hospital Re - admissions, Healthcare Expenditures, Care Quality, Identifying Diabetic Patients

## 1. Introduction

In 2021, hospital spending accounted for 31.1% of total health spending in the United States. This was close to a third of the total [1]. Even after the pandemic, the healthcare system is facing an extreme shortage of staffs, nurses, doctors and healthcare workers. High staffing demands for ED, OR, and ICU overlap.

Anesthesiologists, general, thoracic, and vascular surgeons are in immediate demand. ICU admissions occur simultaneously with ongoing patient arrival to the ED [2]. Under such circumstances, my objective is to efficiently utilize the hospital resources by providing required care to the patients and reduce the inpatient readmission overloading.

## 2. Related Work

Several researchers have applied machine learning techniques to situations other than predicting hospital readmission chances. It is critical to avoid hospital readmissions in order to save expenditures. In 2012, the ACA amended the Social Security Act to include section 1886 (q), which established the Hospital Readmissions Reduction Program (HRRP) [3]. This decreases payment to hospitals with a high number of readmissions. Because of the HRRP's financial incentive system, the first 30 days after patient release are essential in care management. Reduced readmission rates will benefit patient welfare, quality of treatment, and providers' bottom lines. Further studies recommended developing a prediction model for diabetes in Indian females utilizing three machine learning techniques: random forests (RFs), LR, and SVMs, in addition to the characteristics that cause diabetes. Their comparative analysis revealed that RFs outperformed the others [4]. Few research papers have tackled the subject of predicting hospital readmission risk. For instance, Strack et al. used statistical models for this purpose [5].

## 3. Approach

### Dataset

The data set is publicly available on the UC Irvine machine learning repository website. The data was provided to UC Irvine on behalf of Virginia Commonwealth University's Center for Clinical and Translational Research. This data set is a synopsis of the Health Facts database that complies with the Health Insurance Portability and Accountability Act (HIPAA) [6]. Each entry contains over 50 variables that indicate a patient's current status as they were admitted to the healthcare provider. The variables presented encompass a wide variety of data points about a patient: race, gender, age, admission type, time in hospital, number of lab tests conducted, number of drugs administered, diabetes prescriptions, emergency visits in the year preceding the hospitalization, and so on. The hospital stay duration was for a minimum of one day and a maximum of fourteen days.

### Understanding and Cleaning the dataset

As it is typical of real - world data, the original dataset contained incomplete, redundant, and noisy information. Cleaning of data was performed by removing duplicate data rows for the same patient number. Replacing junk/unusable values like '?' with NaN to find all the missing values. Dropping of columns with a high percentage of missing values, which had zero to insignificant impact on the target variable, was executed. For instance - Weight (with 97 percent of values missing). Payer code was removed since it had a high percentage of missing values, it was not considered relevant to the outcome.

### Data Visualization

Figures 1 to 3 are few of the data visualizations executed to attain complete understanding of the dataset.

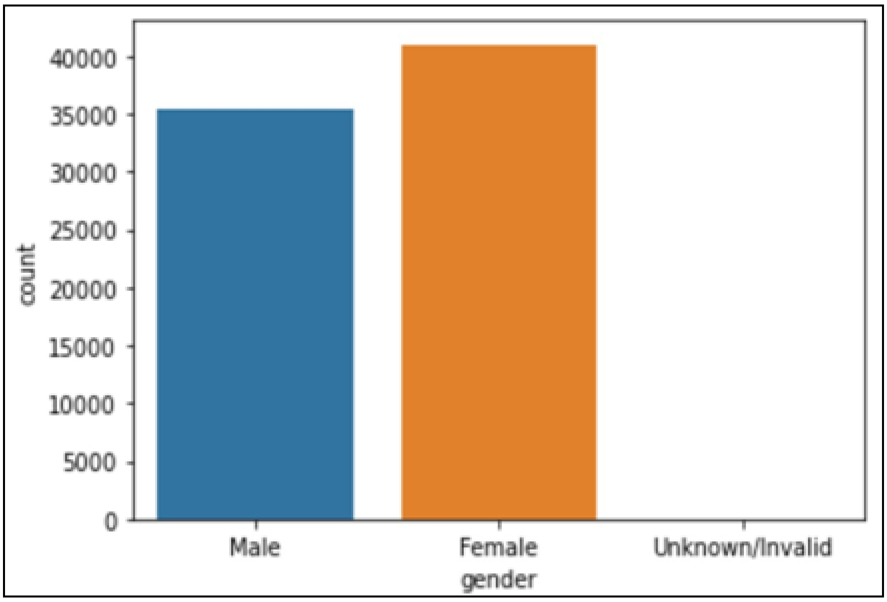


**Figure 1:** Data Count as per the Gender

### Pre - processing and Scaling

For each value in a feature, MinMaxScaler subtracts the minimum value in the feature and then divides by the range. The range is the difference between the original maximum and original minimum. MinMaxScaler preserves the shape of the original distribution. Furthermore, it does not reduce the importance of outliers.
Therefore, it was the best suited scaling technique for our dataset.

### Feature Selection

Attribute or feature selection is performed to ensure that only significant attributes that are potentially associated with the diabetic condition are retained. In a nutshell, it reduces overfitting of the model, model accuracy and reduced training time. Computing the feature importance score and correlation, helps in picking the required features for an optimal model, observed in Figure 4.

Following columns were dropped while creating models - ‘’Unnamed: 0’, ’patient$_nbr', '\ encounter_id', '\ glimepiride-pioglitazone', 'citoglipton', 'examide', 'glipizide''], axis = 1$)

Feature importance scores play an important role in a predictive modeling project, including providing insight into the data, insight into the model. The basis for dimensionality reduction and feature selection leads to improved efficiency and effectiveness of a predictive model.

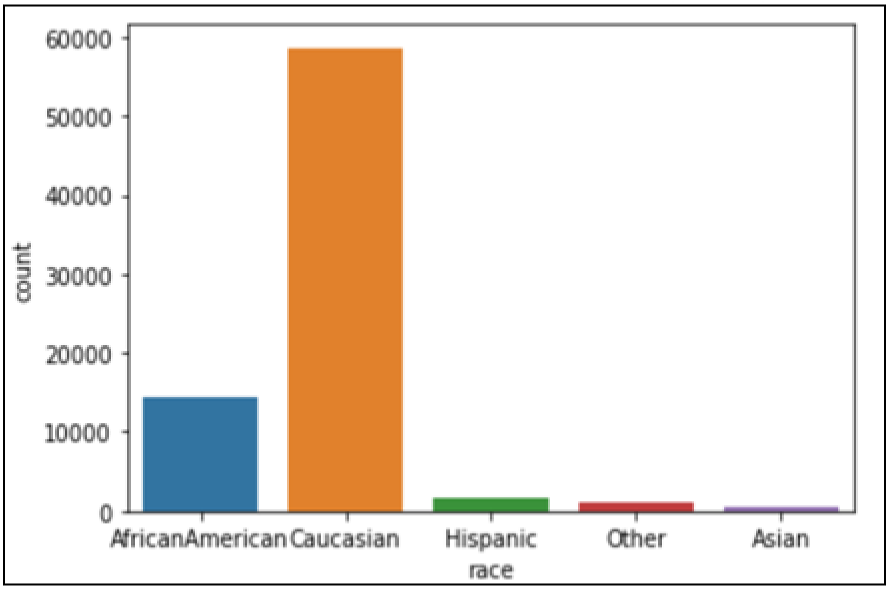


**Figure 2:** Data Count as per Demographics

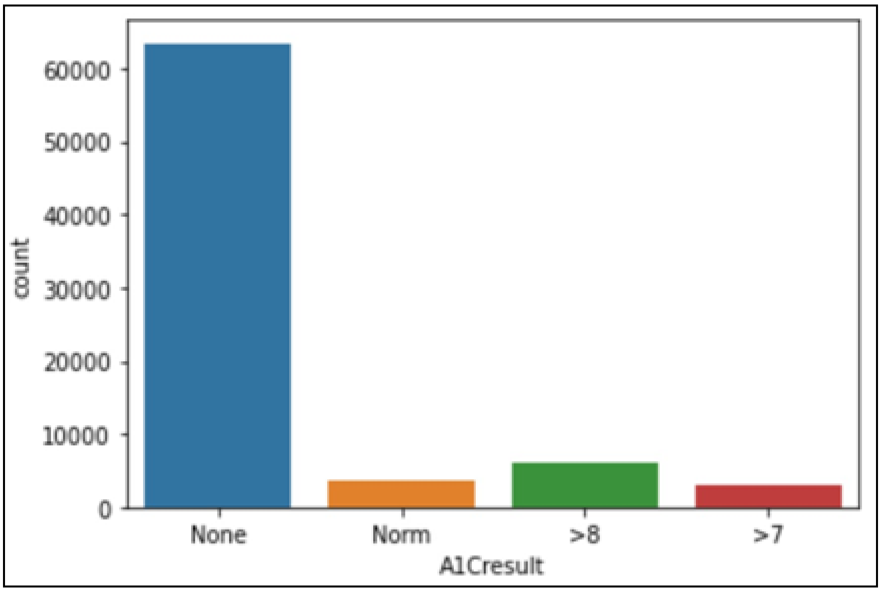


**Figure 3:** A1C Count

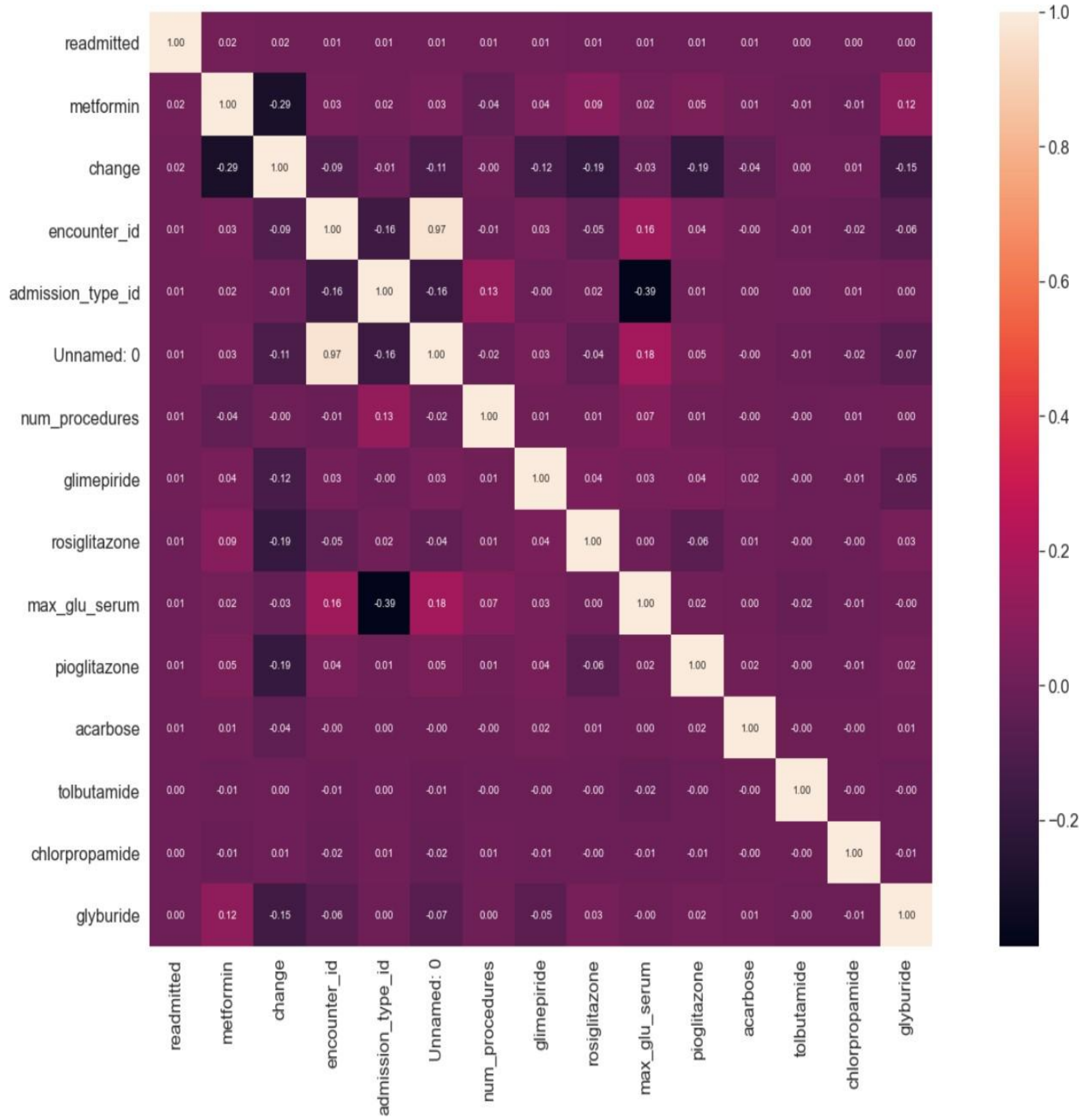
## 4. Implementation - Model Creation

This section includes the four fundamental machine learning techniques used in this research study.

a) Decision Trees are a non - parametric supervised learning method used for classification. The objective is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. [7]

b) K Nearest Neighbors is an algorithm that stores all available cases and classifies new cases based on a similarity measure. Small K values may lead to overfitting, whereas large K values will have cleaner boundaries but smaller variation. Thus, using grid search to get the value of K was appropriate. This model likewise performed well in terms of accuracy but performed poorly in terms of recall.

c) Logistic Regression is a supervised learning algorithm used to predict a dependent categorical target variable.



| | readmitted | metformin | change | encounter_id | admission_type_id | Unnamed: 0 | num_procedures | glimepiride | rosiglitazone | max_glu_serum | pioglitazone | acarbose | tolbutamide | chlorpropamide | glyburide |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| readmitted | 1.00 | 0.02 | 0.02 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.00 | 0.00 | 0.00 |
| metformin | 0.02 | 1.00 | -0.29 | 0.03 | 0.02 | 0.03 | -0.04 | 0.04 | 0.09 | 0.02 | 0.05 | 0.01 | -0.01 | -0.01 | 0.12 |
| change | 0.02 | -0.29 | 1.00 | -0.09 | -0.01 | -0.11 | -0.00 | -0.12 | -0.19 | -0.03 | -0.19 | -0.04 | 0.00 | 0.01 | -0.15 |
| encounter_id | 0.01 | 0.03 | -0.09 | 1.00 | -0.16 | 0.97 | -0.01 | 0.03 | -0.05 | 0.16 | 0.04 | -0.00 | -0.01 | -0.02 | -0.06 |
| admission_type_id | 0.01 | 0.02 | -0.01 | -0.16 | 1.00 | -0.16 | 0.13 | -0.00 | 0.02 | -0.39 | 0.01 | 0.00 | 0.00 | 0.01 | 0.00 |
| Unnamed: 0 | 0.01 | 0.03 | -0.11 | 0.97 | -0.16 | 1.00 | -0.02 | 0.03 | -0.04 | 0.18 | 0.05 | -0.00 | -0.01 | -0.02 | -0.07 |
| num_procedures | 0.01 | -0.04 | -0.00 | -0.01 | 0.13 | -0.02 | 1.00 | 0.01 | 0.01 | 0.07 | 0.01 | -0.00 | -0.00 | 0.01 | 0.00 |
| glimepiride | 0.01 | 0.04 | -0.12 | 0.03 | -0.00 | 0.03 | 0.01 | 1.00 | 0.04 | 0.03 | 0.04 | 0.02 | -0.00 | -0.01 | -0.05 |
| rosiglitazone | 0.01 | 0.09 | -0.19 | -0.05 | 0.02 | -0.04 | 0.01 | 0.04 | 1.00 | 0.00 | -0.06 | 0.01 | -0.00 | -0.00 | 0.03 |
| max_glu_serum | 0.01 | 0.02 | -0.03 | 0.16 | -0.39 | 0.18 | 0.07 | 0.03 | 0.00 | 1.00 | 0.02 | 0.00 | -0.02 | -0.01 | -0.00 |
| pioglitazone | 0.01 | 0.05 | -0.19 | 0.04 | 0.01 | 0.05 | 0.01 | 0.04 | -0.06 | 0.02 | 1.00 | 0.02 | -0.00 | -0.01 | 0.02 |
| acarbose | 0.01 | 0.01 | -0.04 | -0.00 | 0.00 | -0.00 | -0.00 | 0.02 | 0.01 | 0.00 | 0.02 | 1.00 | -0.00 | -0.00 | 0.01 |
| tolbutamide | 0.00 | -0.01 | 0.00 | -0.01 | 0.00 | -0.01 | -0.00 | -0.00 | -0.00 | -0.02 | -0.00 | -0.00 | 1.00 | -0.00 | -0.00 |
| chlorpropamide | 0.00 | -0.01 | 0.01 | -0.02 | 0.01 | -0.02 | 0.01 | -0.01 | -0.00 | -0.01 | -0.01 | -0.00 | -0.00 | 1.00 | -0.01 |
| glyburide | 0.00 | 0.12 | -0.15 | -0.06 | 0.00 | -0.07 | 0.00 | -0.05 | 0.03 | -0.00 | 0.02 | 0.01 | -0.00 | -0.01 | 1.00 |

**Figure 4:** Feature Correlation

d) Random Forest (RF) algorithms form a family of classification methods that rely on the combination of several decision trees. [8]

## 5. Result

Model evaluation metrics explain the performance of each model.

a) Accuracy - It is the ratio of number of correct predictions to the total number of input samples.
b) Misclassification Rate (Error Rate) - proportion of instances misclassified over the whole set of instances
c) Precision - ratio between the True Positives and all the Positives which signifies the correctness of the model classification
d) Recall - Recall gives us the true positive rate (TPR), which is the measure of the model correctly identifying True Positives
e) AUC - Area Under the Curve has a range of [0, 1]. The greater the value, the better is the performance of our model. AUC - ROC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability.

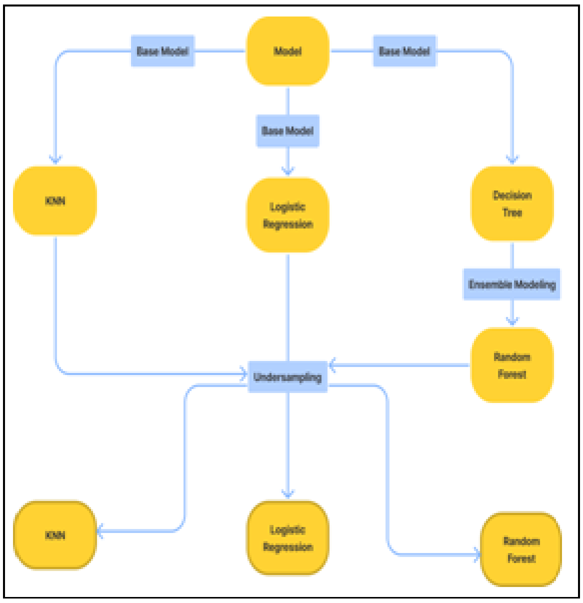


**Figure 5:** Model Flow

**Model Comparison**

Accuracy seems satisfactory but precision and recall values are low for all the models, observed in Figure 6. Thus, the model’s accuracy can be misleading.

| Model | Accuracy | Precision | Recall |
|---|---|---|---|
| KNN | 88% | 24.15% | 3.52% |
| Decision Tree | 79.61% | 15.53% | 18.63% |
| Random Forest | 88.88% | 70.83% | 0.59% |
| Logistics Regression | 88.13% | 35.39% | 7.64% |

**Figure 6:** Model Metrics

Upon close investigation of the dataset, data skewness was observed in Figure 7.

Realizing the extreme skewness of the data set, data balancing measure was performed with Under - sampling. Under sampling is a technique to balance uneven datasets by keeping all the data in the minority class and decreasing the size of the majority class [9]. After building and hyper - tuning multiple models and performing under - sampling, Random Forest delivers the finest outcome, with a balance of good recall and accuracy 8.

## 6. Conclusion

The problem of unplanned hospital readmission from patients whose conditions deteriorate soon after treatment, which is an enormous and expensive burden on the system, will only grow if not properly addressed. By using such predictive machine learning models, we can reduce the burden of the healthcare system and in turn we can take better care of the patients.

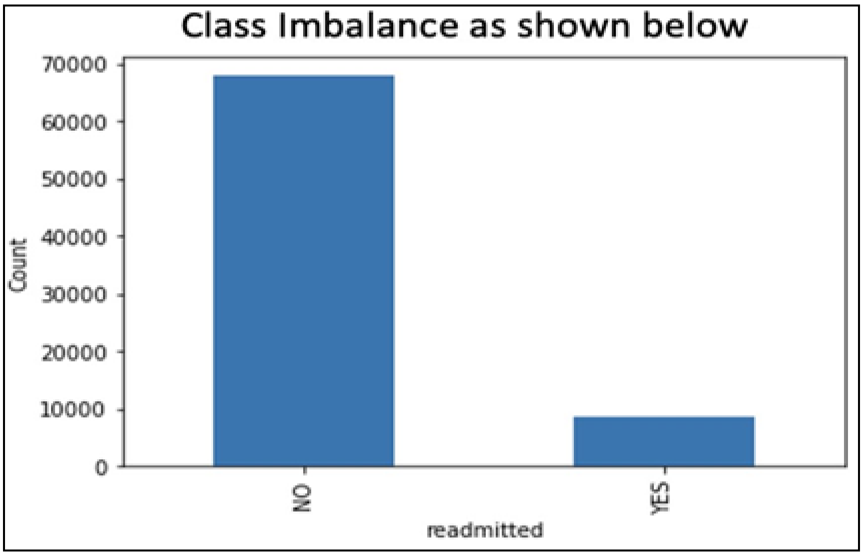


**Figure 7:** Data Skewness

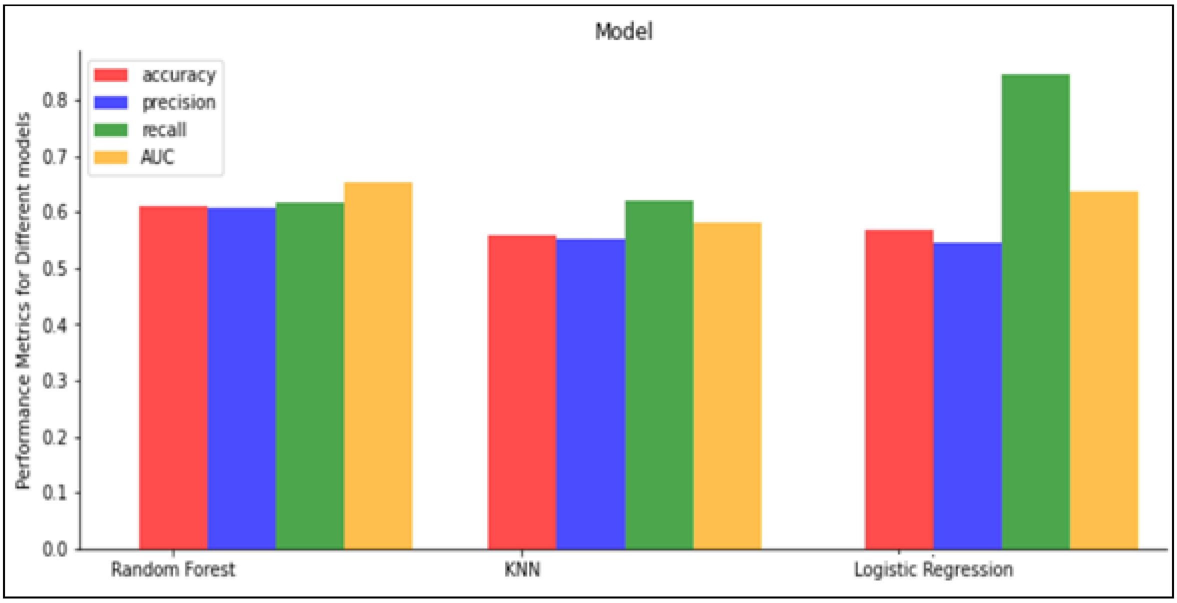


**Figure 8:** Model Evaluation Metrics

The final model reveals that the association between readmission risk and HbA1c level is highly dependent on the primary illness (diabetes is always one of the secondary diagnoses). Findings tend to back up the idea that paying more attention to glucose homeostasis during hospitalization is a good idea.

We can reduce the strain on the healthcare system and, as a result, take better care for patients by employing predictive machine learning models.

## 7. Future Work

This project was aimed only to classify the re - admission of diabetic patients. In the future, with relevant data sets collected from the hospitals across the countries, we can implement the re - admission classification for other pressing ailments like heart and lung diseases. This is a robust way to tackle the hospital re - admission problem.

## References

[1] https: //www. healthsystemtracker. org/

[2] Einav S, Aharonson - Daniel L, Weissman C, Freund HR, Peleg K; Israel Trauma Group. In - hospital resource utilization during multiple casualty incidents. Ann Surg.2006 Apr; 243 (4): 533 - 40. doi: 10.1097/01. sla.0000206417.58432.48. PMID: 16552206; PMCID: PMC1448970.

[3] C. for Medicare, M. S.7500 S. B. Baltimore, and M. Usa, “ReadmissionsReduction - Program, ” 27 - Apr - 2018. [Online]. Available: https: //www.cms. gov/medicare/medicare - fee - for - servicepayment/ acuteinpatientpps/ readmissions - reduction - program. html.

[4] D. Paul, ”Analysing Feature Importances for Diabetes Prediction using Machine Learning Debadri, ” 2018 IEEE 9th Annu. Inf. Technol. Electron. Mob. Commun. Conf., pp.924 - 928, 2018.

[5] 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.

[6] “UCI Machine Learning Repository: Diabetes 130 - US hospitals for years 1999 - 2008 Data Set. ” [Online]. Available: https: //archive. ics. uci. edu/ml/datasets/diabetes+130 - us+hospitals+for+years+1999 - 2008.

[7] https: //scikit - learn. org/stable/modules/tree. html

[8] Bernard, Simon, S´ebastien Adam, and Laurent Heutte. ”Dynamic random forests. ” Pattern Recognition Letters 33.12 (2012): 1580 - 1586.

[9] https: //www. masters in datascience. org/learning/statistics - data - science/undersampling/

[10] Weiss, Audrey J., and H. Joanna Jiang. ”Overview of clinical conditions with frequent and costly hospital readmissions by payer, 2018. ” (2021)