ALY6140- Capstone Project

Exploratory Data Analytics
of HbA1c value at the time of
admission & readmittance rate
of patients

Submitted by-
Supreeth Murugesh,
Harshit Gaur &
Jeseeka Shah

Abstract

The analysis of the test HbA1c will be very significant as this would set a baseline for the change during the hospitalization of the patients. The values of HbA1c were considered during the time of admission of inpatient, outpatients, and emergency. After careful analysis, we found that males take the test more as compared to the female population other than those having diabetic symptoms. The dataset that we considered was UCI ML Repository which initially had the 101,766 encounters with 55 features and then after data cleaning we narrowed it down to 17000 encounters and 28 features. The HbA1c values indicate that the range that falls between greater than 7 and less than 8 are in the pre-diabetic range and greater than 8 belongs to the diabetic range. The statistical models we have used shows the probability between readmission and HbA1c measures depends on the primary diagnosis that is diagnosis1 in the dataset. Furthermore, the outcome pointed that greater attention given to the diabetes reflection in HbA1c value determination resulted in the improvement of the patient by implementation of change in medication. To arrive at this conclusion, we use three different supervised learning models. The random forest classifier, K nearest neighbour classifier, and logistic regression model. All the three models pointed out the accuracy to be around 62 percent and 72.2 percent patients who had HbA1c values more than normal values were readmitted and 27.8 percent patients with normal values were readmitted.

Introduction

The topic and dataset that we have chosen is to baseline the care strategies that are used at hospital to cure the diabetic patients. Our study has included the dataset from the Health Fact database, being a national data warehouse that contains comprehensive data from across the United States.

In our initial proposal we have mentioned that the dataset represents 130 hospitals, and we will be using 101,766 for exploratory data analysis that satisfy the five factors such as inpatient encountered, diabetic encountered, duration of stay, lab results and medication used during the stay at the hospital.

Data Description and EDA process

The data used in our project is part of the Heath Fact database which stores the clinical records of many hospitals in United States. The database contains values which are part of electronic medical records and includes the data like number of inpatients, outpatients, emergency, demographics (age, sex, and race), different stages of diagnoses, laboratory data, in-hospital procedures recorded under the ICD-9-CM codes and mortality [1].

In this project, we are interested in finding out the early readmission, the dataset consists of a column named ‘readmitted’ that determines the duration after which the patient was admitted to the hospital. There are 3 parameters to this column, ‘No’ readmit stands for no patient was admitted again after the discharge and the other two are show the number readmits based on before 30 days or after 30 days of discharge. The 30 days is set as per the bill cycle used by funding agency. The medication measure named Haemoglobin A1c (HbA1C) is an important measure of glucose control which is used to measure the performance diabetes care. The HbA1C helps to assess the efficiency of on-going therapy and make amendments in the theory if HbA1c > 8 and if it is in between the range 8>HbA1C>7 then the result is normal.

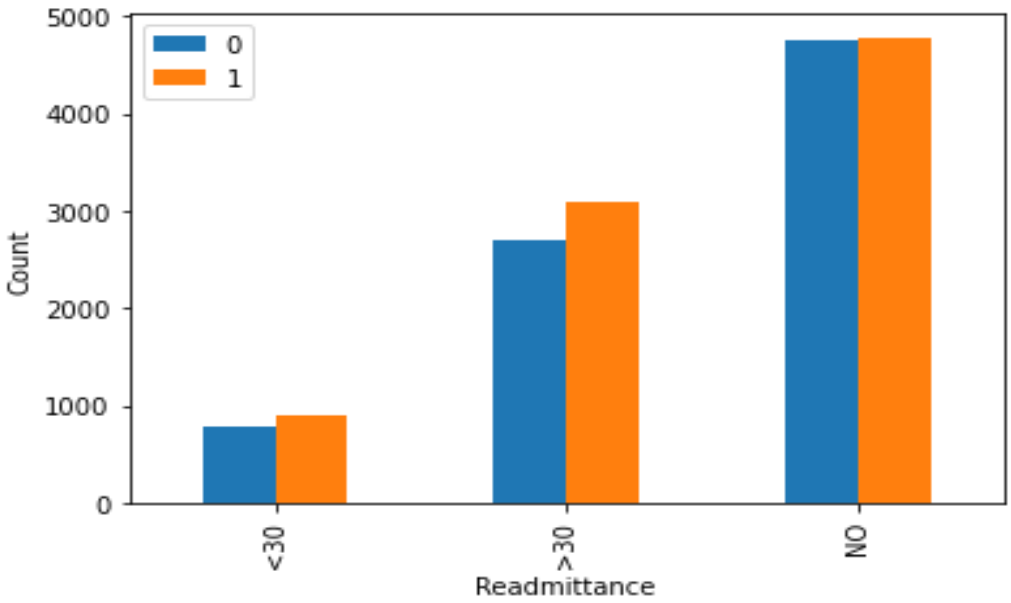

Figure 1: Mapping gender with readmittance by taking the total count.

The figure 1. above shows the count of female and male that were readmitted. The number of females that were readmitted after the duration greater or less than 30 days is more than that of the male populations. This depicts that the female population are more like to be readmitted overall.

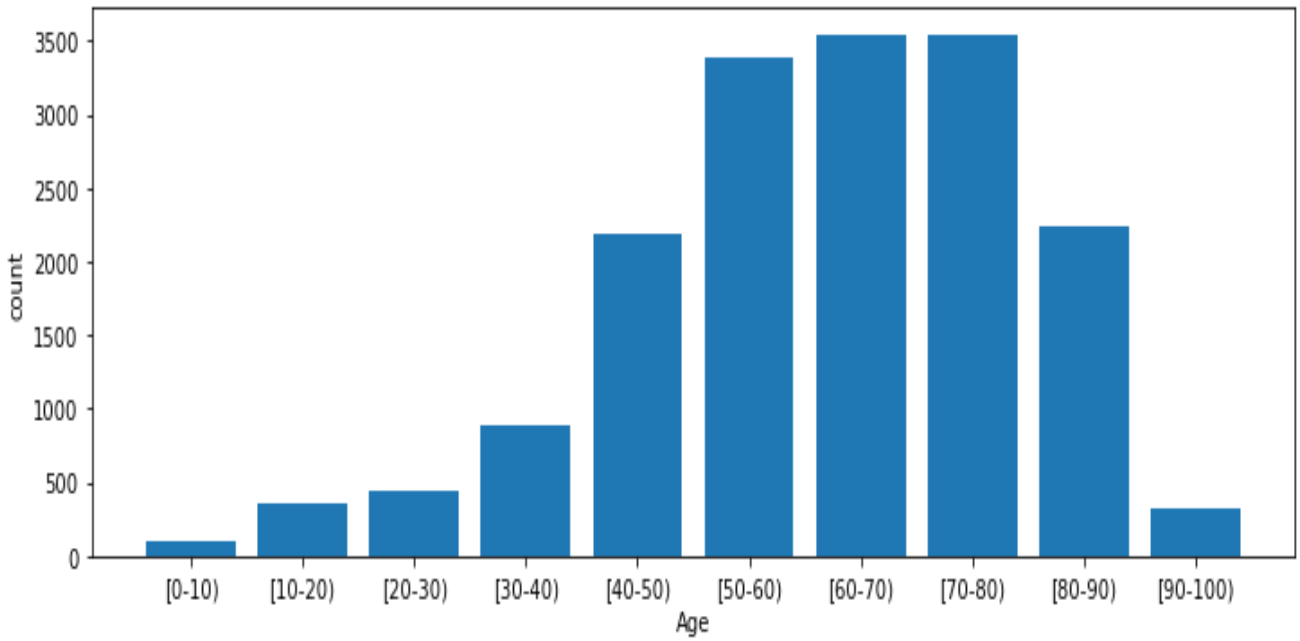

Figure 2: Relation between Age and Readmittance in terms of Count

The above graph shows that number readmittance increases as the age increases and then drops. The age group between 70-80 has the highest readmittance and the age group below 10 has the least readmittance.

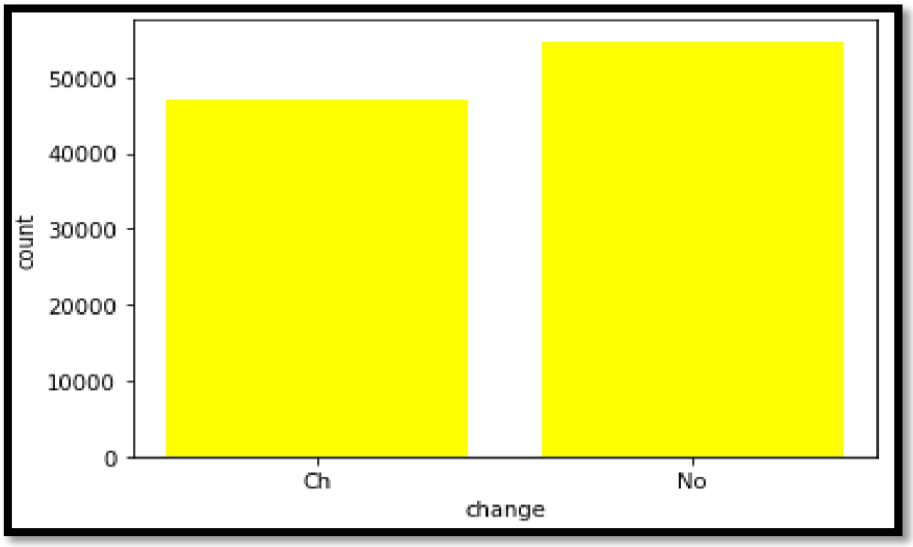

Figure 3 : Count indicating change in diabetic medication

The change in the medication is taken place based on the result of the HBA1C test result and its frequency of retaking the test. The change in medication indicates the change can be increase or reduction in the medication as well as change in the generic name of the dosage.

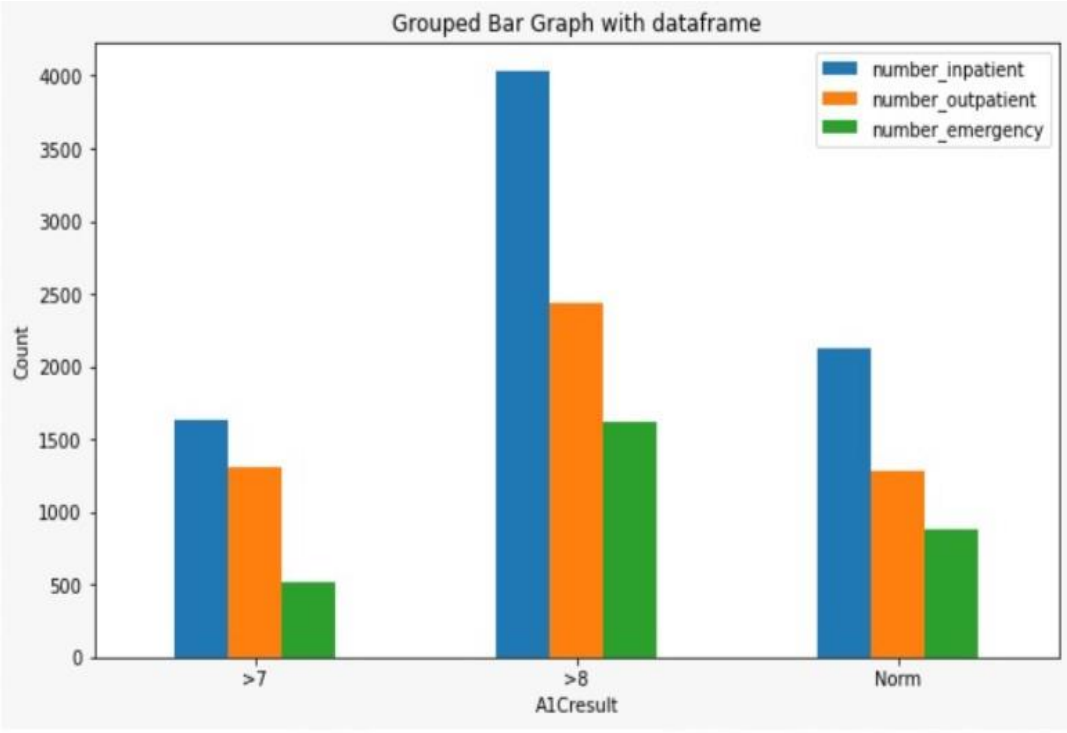

Figure 4: Count indicating change in diabetic medication

The Figure 4 shows the relation between number of inpatients, outpatients, and emergency against A1c test result. Haemoglobin A1c (HbA1c) is a measure of glucose control which is considered while

measuring the efficiency of the hospital's diabetes care centre. The above graph is plotted to determine if HBA1C test was taken while admission of the patient and used also to make the change in the treatment if the value is greater than 8 percent on the current regime.

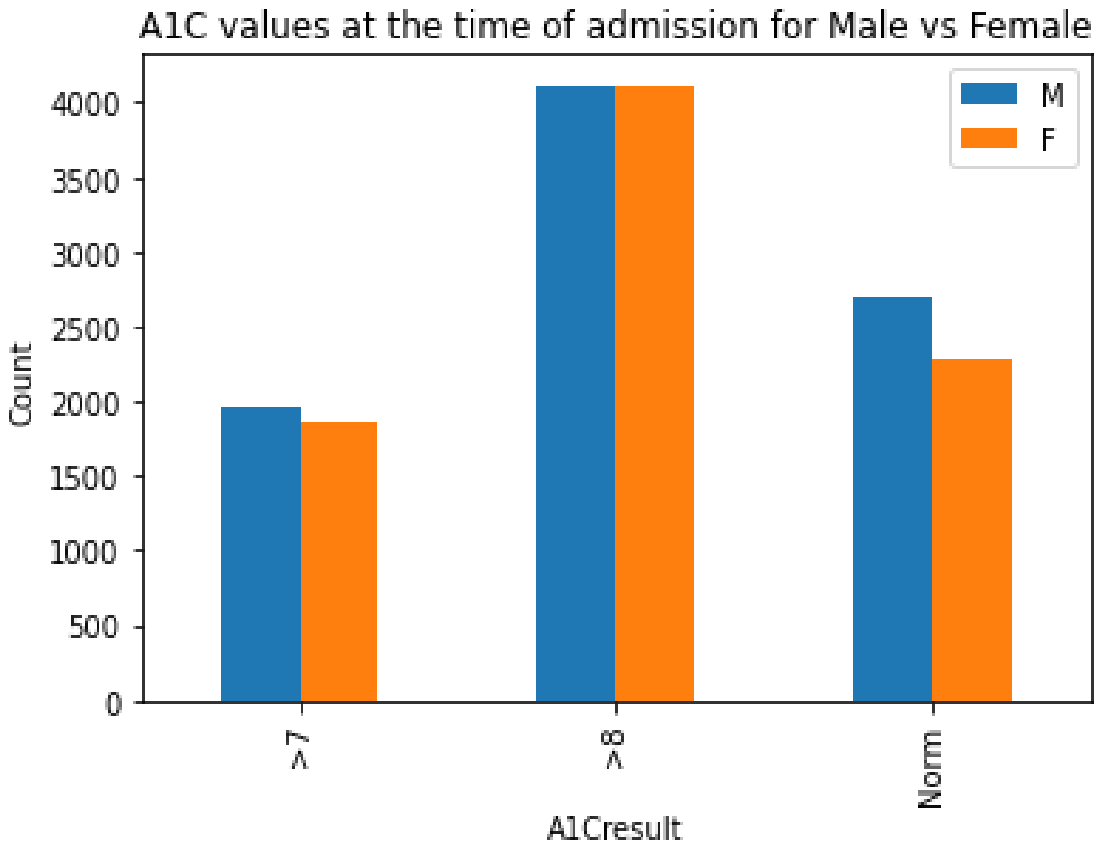

Figure 5 :Count indicating the HBA1C result of the male and female

The figure 5 depicts the HBA1C test result values are more at the range indicating the percentage greater than 8 for male and female implying the change in the medication.

The exploratory data analysis is carried out in Python using Plotly, Seaborn and Matplotlib. Our initial approach started with data cleaning and extraction. After looking into the data, we have considered to drop number of variables from original 50 variables. Firstly, we checked for maximum number of missing values, null values, non-significant values for our EDA and the unbalanced values that the categorical column has and have dropped them from the data frame.

Now, according to our problem statement we decided on keeping the 'readmitted' column as the target value for training and testing data. The column is taken as the target hence it cannot be included in either training or testing to predict the accuracy.

Models used:

After data cleaning and pre-processing the data, we decided to apply classification and regression models. Since we had 38+ features and it was really difficult for us to decide on what best subset of features we had to use to get the best achievable features. Hence, we decided to choose Random Forest Classifier to run on a loop to get the best subset of features with feature importance. The column 'Diag_1' diagnosis-1 was numerical values according to ICD-9 codes of the medical standards which later became useable feature. Since most of the useable columns we are using are categorical data, we applied one-hot encoding on the corresponding columns to make it useable to apply Classifier and Regression models.

Model 1- Random Forest Classifier

1. **Random Forest Classifier**: We first applied Random Forest Classifier to get to know the N-Estimator parameter by plotting against the mean error. We then applied the a for loop to run against the values of N-Estimator with a break of 50. N-Estimator value with least mean error was considered as the final value for R Random Forest Classifier. We plotted different N-estimator value with mean error to find out the best parameter value. For N-estimator value 250 we obtained an accuracy of 63.2% as shown in the table.

Accuracy obtained for different N-Estimators:

N-Estimator	Accuracy	N-Estimator	Accuracy
100	0.623	400	0.624
150	0.615	450	0.631
200	0.623	500	0.624
250	0.632	550	0.63
300	0.627	600	0.631
350	0.624	650	0.63

Table 1- N- Estimation with accuracy

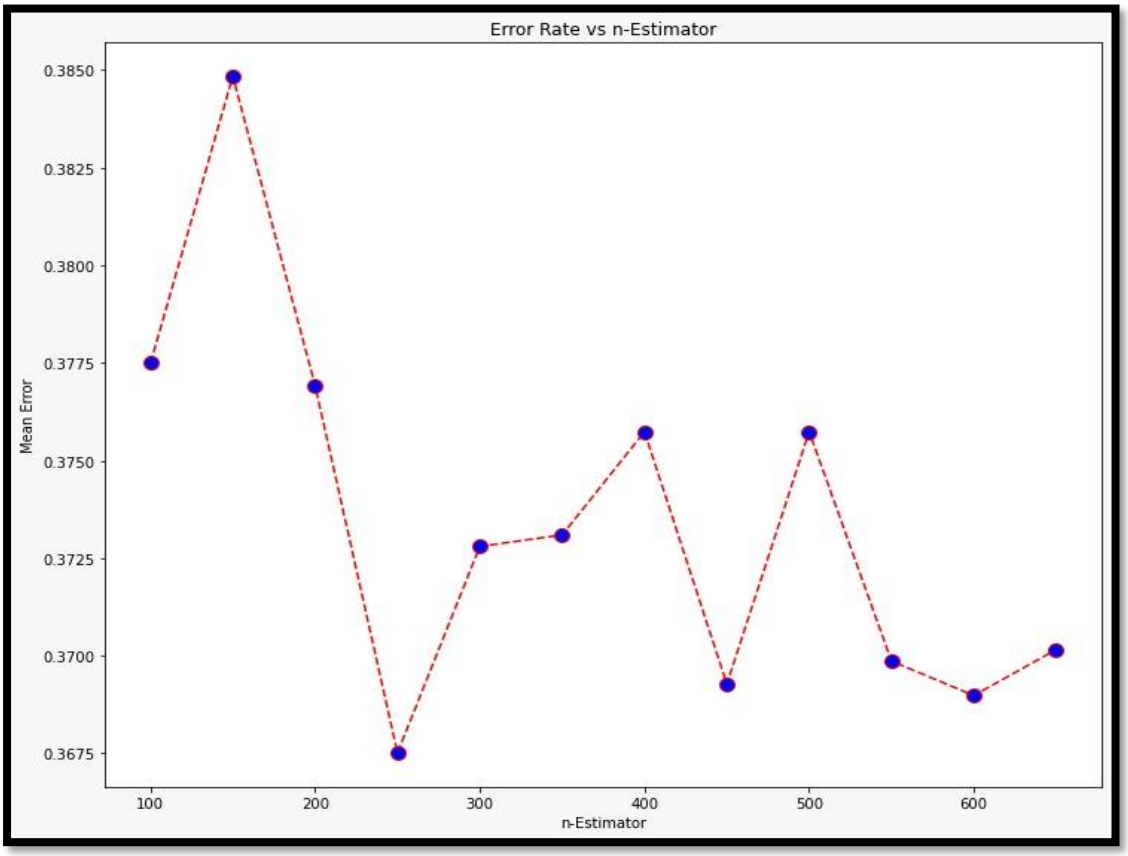

Figure 6: N estimator for random forest classifier

Model-2 -K Nearest Neighbour

2. **KNN Classifier**: Next, we applied KNN classifier, and we had the problem of deciding the K value for the classifier. We again applied for loop for k values ranging from 4 to 14 inclusive and plotted the mean error. The K value which had the minimum mean square error was finally taken as the value for the KNN classifier. To estimate the optimum K value we again used mean error and obtained the highest accuracy of 59% for a K value of 16 as show in the table below.

- **Accuracy obtained for different K values:**

K Value	Accuracy	K Value	Accuracy
4	0.557	12	0.572
6	0.563	14	0.588
8	0.573	16	0.59
10	0.576	18	0.587

Table 2 - K value with accuracy values

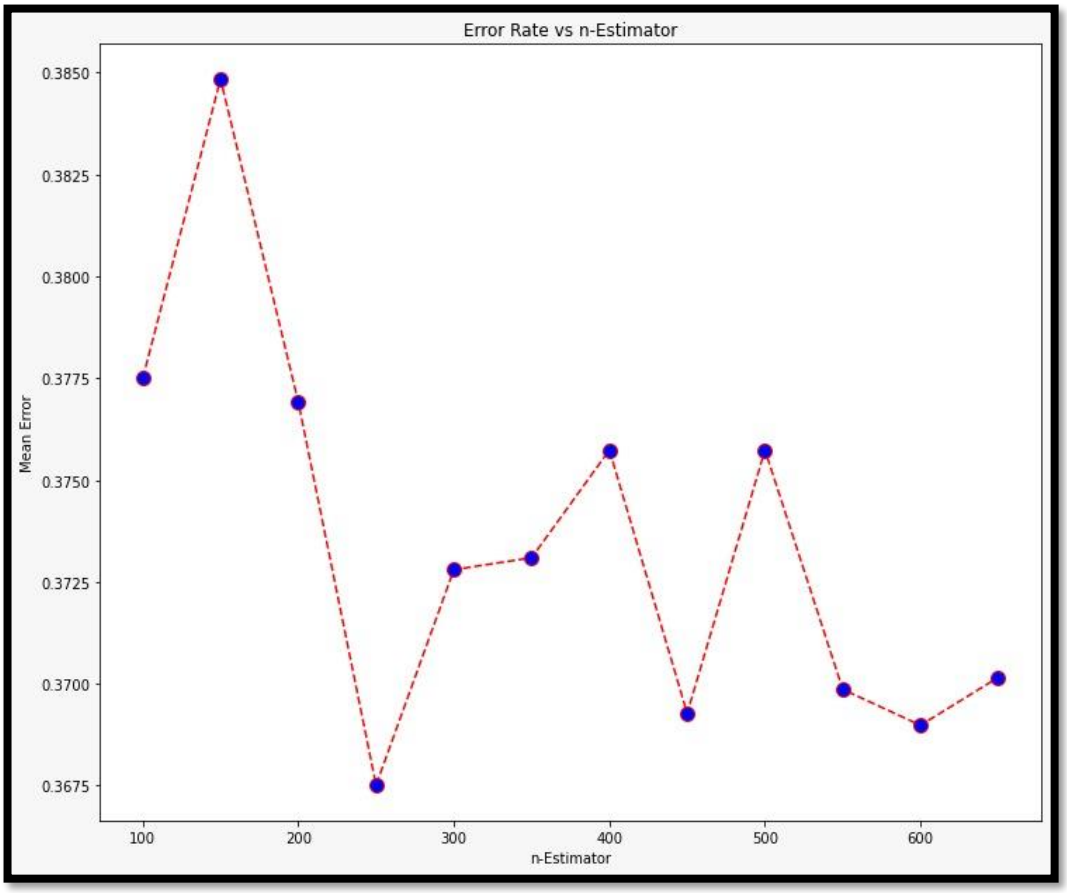

Figure 7 – Mean Error for different K values.

Model-3-Logestic Regression

3. **Logistic Regression**: The third model that was applied was Logistic Regression and obtained an accuracy of 63%. In the coming week we aim to compare the results of the different classifier and try to tune the models. The confusion matrix as shown in Figure 7 tells us that the model was comparatively successful in predicting true positive and performed comparatively less in False positive.

Accuracy of Logistic Regression : 62.72%

	PRED : READMITTED	PRED : NOT READMITTED
TRUE : READMITTED	1560	325
TRUE : NOT READMITTED	944	575

Figure 8 – Logistic Regression – accuracy acquired 62.72 percent

Model-4- Support Vector Machine

4. **Support Vector Machine:** A support vector machine is a machine learning model that is able to generalise between two different classes if the set of labelled data is provided in the training set to the algorithm. The main function of the SVM is to check for that hyperplane that is able to distinguish between the two classes. Two kernels were applied Linear and Polygon kernel and respective kernels. We obtained almost nearly same precision and recall for linear and polygonal kernels.

KERNAL: LINEAR

	precision	recall	f1-score	support
NO	0.59	0.92	0.72	1885
YES	0.69	0.22	0.34	1519
accuracy			0.61	3404
macro avg	0.64	0.57	0.53	3404
weighted avg	0.64	0.61	0.55	3404

Figure 9 – Support Vector Machine- Kernel value- Linear

KERNAL: POLY

	precision	recall	f1-score	support
NO	0.58	0.94	0.72	1885
YES	0.69	0.16	0.26	1519
accuracy			0.59	3404
macro avg	0.64	0.55	0.49	3404
weighted avg	0.63	0.59	0.51	3404

Figure 10 – Support Vector Machine- Kernel value- Poly

Conclusion

For the initial part, we have performed analysis of the data set to figure out the unbalanced features and numeric medical identification features which do not hold any significance in the application of modelling on this data set. Since, most of the features in this data set are categorical, we proceeded further with the implementation of One-Hot Encoding to make-ready our data set for modelling and afterwards performed feature engineering based on the importance matrix produced by Random Forest Classifier algorithm. We applied Random Forest Classifier & K-Nearest Neighbour models with the optimal N-Estimator & K-Neighbour values (both calculated using Mean Error Value graphs) respectively and found the accuracies to be around 65%. Later, we used the Logistic Regression & Support Vector Machine models to fit the training set and predict the testing data set in 80/20 rule. These models also gave out the accuracies around 65%. The plotting of confusion matrices of target outcomes (training & testing) after application of each of the 4 models inferred that the prediction rates of 'Readmitted' values are very high, but the same cannot be said for 'Not Readmitted' values. The 'True Positive' values are very high in comparison to 'False Positive' which signifies a good prediction of the models for the value, but the 'True Negative' values are not that high when compared against values of 'False Negative'.

Since, all the 4 models are giving the same kind of accuracies and generating the same kind of confusion matrices, we are inclining towards the introduction of Neural Network algorithm and model in our project and its application to our dataset. Hopefully, we can achieve far greater success with Neural Network learning algorithm in future.

Bibliography

1. Strack, B., DeShazo, J. P., Gennings, C., Olmo, J. L., Ventura, S., Cios, K. J., & Clore, J. N. (2014). Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records. BioMed Research International, 2014, 1–11. https://doi.org/10.1155/2014/781670
2. A. G. Pittas, R. D. Siegel, and J. Lau, "Insulin therapy for crit- ically ill hospitalized patients: a meta-analysis of randomized controlled trials," Archives of Internal Medicine, vol. 164, no. 18, pp. 2005–2011, 2004.
3. Hatwell, J., Gaber, M. M., & Azad, R. M. A. (2020). CHIRPS: Explaining random forest classification. Artificial Intelligence Review, 53(8), 5747–5788. https://doi.org/10.1007/s10462-020-09833-6
4. WARE, M., FRANK, E., HOLMES, G., HALL, M., & WITTEN, I. H. (2001). Interactive machine learning: letting users build classifiers. International Journal of HumanComputer Studies, 55(3), 281–292. https://doi.org/10.1006/ijhc.2001.0499
5. sklearn.ensemble.RandomForestClassifier. (n.d.). Scikit-Learn. https://scikitlearn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html