Predicting Hospital Readmission in Patients with Diabetes

Sumithra HariGuruprasad

SCS_3252_017 Big Data Management Systems & Tools

Objective:

To predict whether a patient with diabetes will be readmitted to the hospital.

Information On Data:

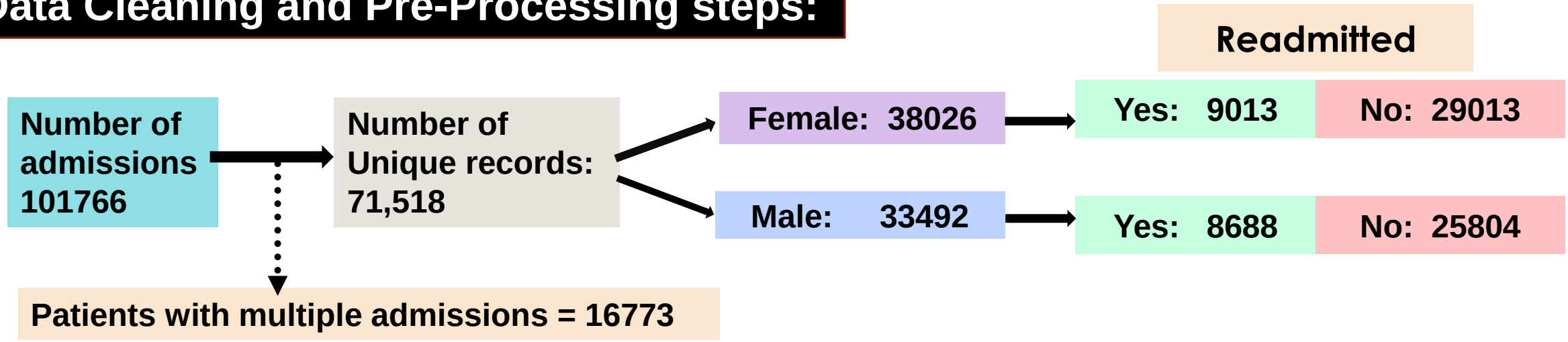
The dataset that is used comes from UCI ML repository representing 10 years (1999-2008) of clinical care at hospitals

It consists of over 100000 hospital admissions from patients with diabetes which includes 46 features representing patients and hospital outcomes

Information extracted from database contains variables of type categorical, numerical and nominal

Some of the important features are Race, Gender, Age, Admission type, Time in Hospital, Medical Specialty of admitting physician, Number of Lab test performed, A1C test result, Number of Diagnosis, Number of medications, Diabetic Medications, Number of Outpatient, Inpatient, and Emergency visits in the year before the hospitalization, etc.

Data Cleaning and Pre-Processing steps:

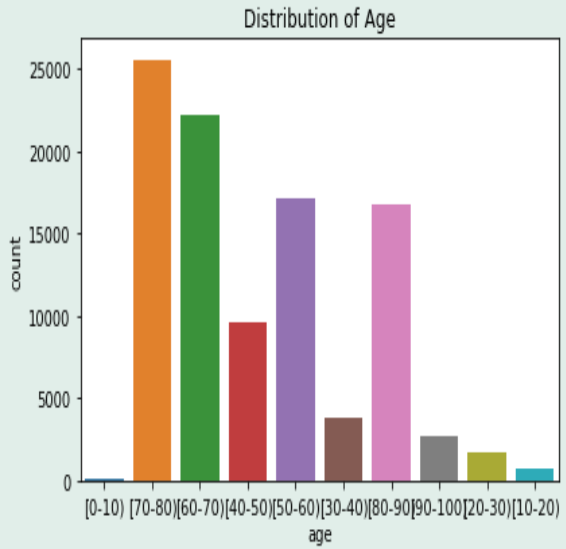


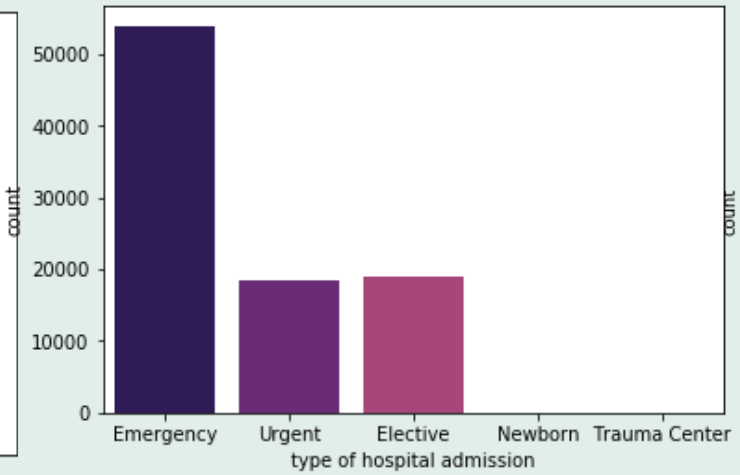
- Admission type column had values '**Not Mapped**' and "**Not Available**' which represent null values. So these two values were assigned **Null**. Discharge disposition description had values containing terms like '**Expired**' and '**Hospice**'(**8221)** were removed from the data.

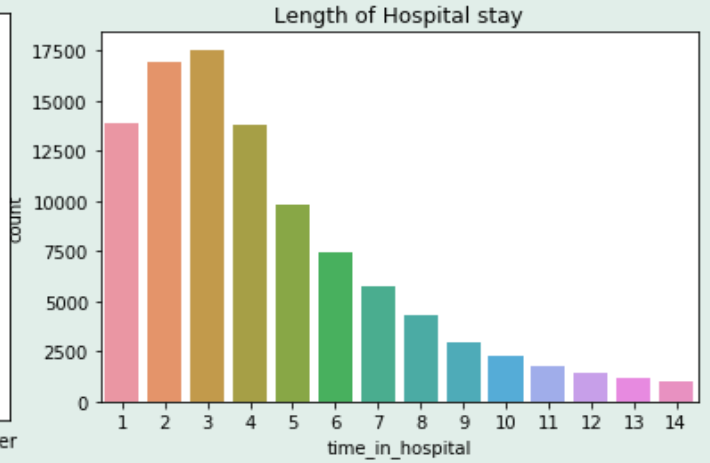
- 23 features of medications containing the values '**Up**', '**Down**', '**Steady**', updated to "1" (taking the medication) and '**No** to "0" (not taking the medication). Further '0' and '1' were converted into Boolean values to calculate the proportion of each medication.

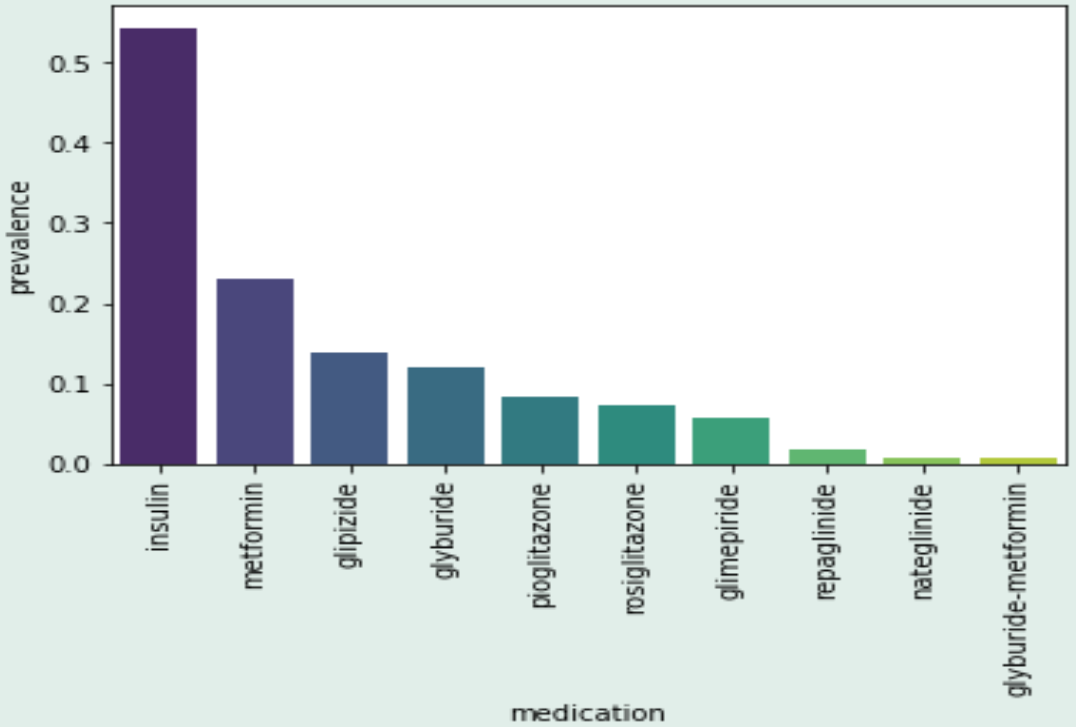
- Maximum Number of missing values found in '**Max_glu_serum**', '**A1Cresult**' and '**Medical specialty**' i.e.., **88814, 74507** and **44774** respectively.

- Target Variable Readmitted had values '**>30**' and '**<30**' converted to '**1**' and a value '**NO**' to '**0**'.







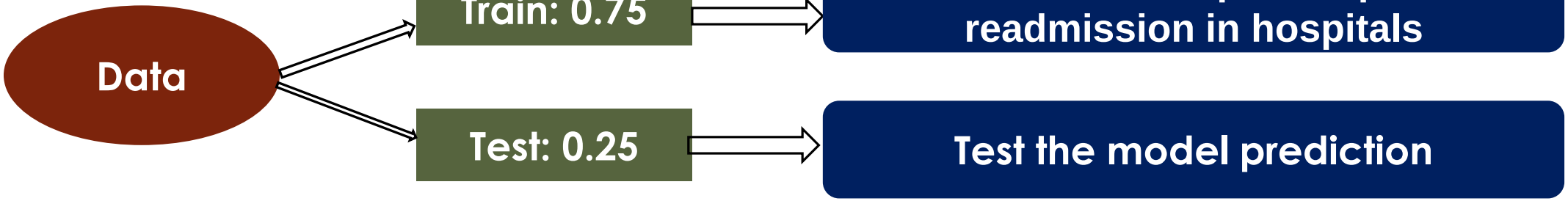


Feature Selection and Encoding

Variables such as admission type, discharge disposition description, encounter id, patient number , weight, few medications ,medical specialty, AIC result ,Maximum glucose serum were dropped due to either irrelevance to our analysis or it contained more than 75% null values.

StringIndexer – converts categorical variables to a column of label indices, **OneHotEncoderEstimator** -transforms multiple columns into an one-hot-encoded output vector column ,**VectorAssembler** - transform multiple columns into a single vector column.

Model Preparation



Models

- Random Forest Classifier
- Gradient Boosted Trees
- Logistic Regression
- Decision Tree

Evaluation:

MultiClassificationEvaluator- To calculate TP,FP, TN and FN
BinaryClassificationEvaluator – To Calculate Area Under ROC

Accuracy Score: 0.62

Precision: 0.63

Recall: 0.48

Confusion Matrix

	0	1
0	9118	2853
1	5562	5031

	0	1
0	TN	FP
1	FN	TP

Conclusion:

- Patients who were female, Caucasians , in the age of 60+, having no procedures done on them and who were not on diabetes medication and who had 0 inpatient visits, if admitted in emergency had more chance of getting readmitted.

- One fourth of the total number of patients were not on diabetes medication. Of the hospital admissions where A1C result was measured, almost half had a A1C level greater than 8, which suggests that the patient's diabetes was poorly managed. Also, the availability of A1C data is sparse in our dataset.

- The number of patients who had 0 inpatient days were higher and they were the one who got readmitted in hospitals and one third of those patients were readmitted after 30 days.

- We can try to improve the accuracy on our models by feature creation, better handling of null values using imputer and trying out different models like Naïve Bayes or KNN algorithm.

- Reducing readmission rates of diabetic patients has the potential to greatly reduce health care costs while simultaneously improving care.


Thank You!