INTRODUCTION

It was reported that in 2011 more than 3.3 million patients were readmitted in the US within 30 days of being discharged, and they were associated with about $41 billion in hospital costs. The need for readmission indicates that inadequate care was provided to the patient at the time of first admission. The readmission rate has become an important metric measuring the overall quality of a hospital.

Diabetes is the 7th leading cause of death and affects about 23.6 million people in the US. 1.4 million Americans are diagnosed with diabetes every year. Hospital readmission being a major concern in diabetes care, over 250 million dollars was spent on treatment of readmitted diabetic patients in 2011. Early identification of patients facing a high risk of readmission can enable healthcare providers to conduct additional investigations and possibly prevent future readmissions.

In this project, I build a machine learning classifier model to predict diabetes patients with high risk of readmission. Note that higher sensitivity (recall) is more desirable for hospitals because it is more crucial to correctly identify "high risk" patients who are likely to be readmitted than identifying "low risk" patients.

PROPOSED METHODOLOGY

CLASSIFICATION

Classification is a fundamental task in machine learning that involves predicting the class or category of an input based on its features. It is a supervised learning approach where the algorithm learns from labeled training data to make predictions on new, unseen data.

In classification, the input data consists of a set of features (also called attributes or independent variables) and their corresponding class labels (dependent variable). The goal is to build a model that can accurately assign the correct class label to unseen instances based on their features.

Classification models learn patterns and relationships in the training data to generalize and make predictions. There are various classification algorithms, each with its own underlying principles and assumptions.

TYPES OF CLASSIFICATION MODELS USED:

Logistic Regression: Logistic regression models the relationship between the features and the probability of belonging to a specific class. It assumes a linear relationship between the features and uses a logistic function to map the linear combination to a probability distribution.

Decision Trees: Decision tree models use a tree-like structure to make decisions based on feature values. The tree is constructed by recursively splitting the data based on feature conditions that optimize a certain criterion, such as maximizing information gain or Gini impurity.

Random Forest: Random forest is an ensemble learning method that combines multiple decision trees. Each tree is trained on a random subset of features and

makes predictions. The final prediction is determined by aggregating the predictions of all individual trees.

Naive Bayes: Naive Bayes is a probabilistic classifier based on Bayes' theorem. It assumes that the features are conditionally independent given the class label. It calculates the probability of each class given the features and selects the class with the highest probability.

DATASET DESCRIPTION

The dataset represents 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. It includes 50 features representing 101766 diabetes patient and hospital outcomes. Information was extracted from the database for encounters that satisfied the following criteria:

It is an inpatient encounter (a hospital admission).

It is a diabetic encounter, that is, one during which any kind of diabetes was entered to the system as a diagnosis.

The length of stay was at least 1 day and at most 14 days.

Laboratory tests were performed during the encounter.

Medications were administered during the encounter.

The data contains such attributes as patient number, race, gender, age, admission type, time in hospital, medical specialty of admitting physician, number of lab test performed, HbA1c test result, diagnosis, number of medication, diabetic medications, number of outpatient, inpatient, and emergency visits in the year before the hospitalization, etc.

GLIMPSE OF DATA

	encounter_id	patient_nbr	race	gender	age	weight	admission_type_id	discharge_disposition_id	admission_source_id
0	2278392	8222157	Caucasian	Female	[0-10)	NaN	6	25	1
1	149190	55629189	Caucasian	Female	[10-20)	NaN	1	1	7
2	64410	86047875	AfricanAmerican	Female	[20-30)	NaN	1	1	7
3	500364	82442376	Caucasian	Male	[30-40)	NaN	1	1	7
4	16680	42519267	Caucasian	Male	[40-50)	NaN	1	1	7

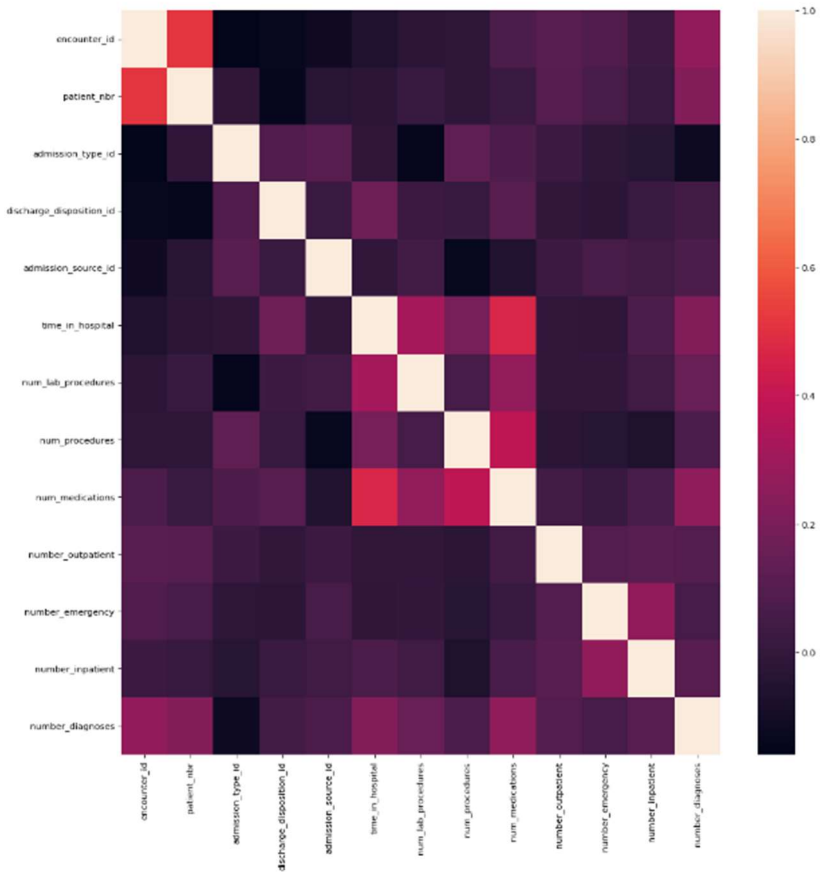
citoglipton	insulin	glyburide-metformin	glipizide-metformin	glimepiride-pioglitazone	metformin-rosiglitazone	metformin-pioglitazone	change	diabetesMed	readmitted
No	No	No	No	No	No	No	No	No	NO
No	Up	No	No	No	No	No	Ch	Yes	>30
No	No	No	No	No	No	No	No	Yes	NO
No	Up	No	No	No	No	No	Ch	Yes	NO
No	Steady	No	No	No	No	No	Ch	Yes	NO

Table 1: Description of Columns

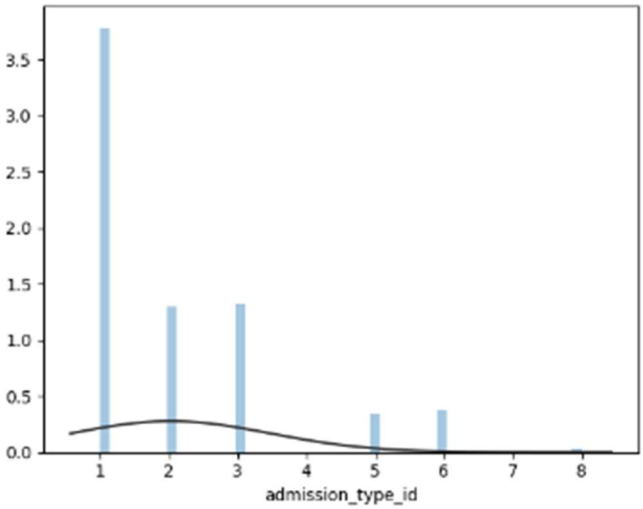
Feature name	Type	Description and values	% missing
Encounter ID	Numeric	Unique identifier of an encounter	0%
Patient number	Numeric	Unique identifier of a patient	0%
Race	Nominal	Values: Caucasian, Asian, African American, Hispanic, and other	2%
Gender	Nominal	Values: male, female, and unknown/invalid	0%
Age	Nominal	Grouped in 10-year intervals: [0, 10), [10, 20), . . ., [90, 100)	0%
Weight	Numeric	Weight in pounds.	97%
Admission type	Nominal	Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available	0%
Discharge disposition	Nominal	Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available	0%
Admission source	Nominal	Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital	0%
Time in hospital	Numeric	Integer number of days between admission and discharge	0%
Payer code	Nominal	Integer identifier corresponding to 23 distinct values, for example, Blue Cross\Blue Shield, Medicare, and self-pay	52%
Medical specialty	Nominal	Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family\general practice, and surgeon	53%
Number of lab procedures	Numeric	Number of lab tests performed during the encounter	0%
Number of procedures	Numeric	Number of procedures (other than lab tests) performed during the encounter	0%
Number of medications	Numeric	Number of distinct generic names administered during the encounter	0%
Number of outpatient visits	Numeric	Number of outpatient visits of the patient in the year preceding the encounter	0%
Number of emergency visits	Numeric	Number of emergency visits of the patient in the year preceding the encounter	0%
Number of inpatient visits	Numeric	Number of inpatient visits of the patient in the year preceding the encounter	0%
Diagnosis 1	Nominal	The primary diagnosis (coded as first three digits of ICD9); 848 distinct values	0%
Diagnosis 2	Nominal	Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values	0%
Diagnosis 3	Nominal	Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values	1%
Number of diagnoses	Numeric	Number of diagnoses entered to the system	0%
Glucose serum test result	Nominal	Indicates the range of the result or if the test was not taken. Values: “>200,” “>300,” “normal,” and “none” if not measured	0%
		Indicates the range of the result or if the test was not taken. Values: “>8” if the result	

DATA VISUALIZATION

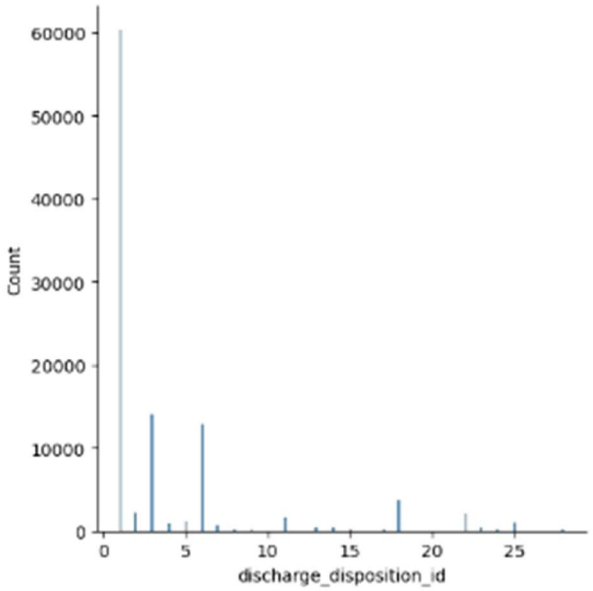
1)A correlation heatmap



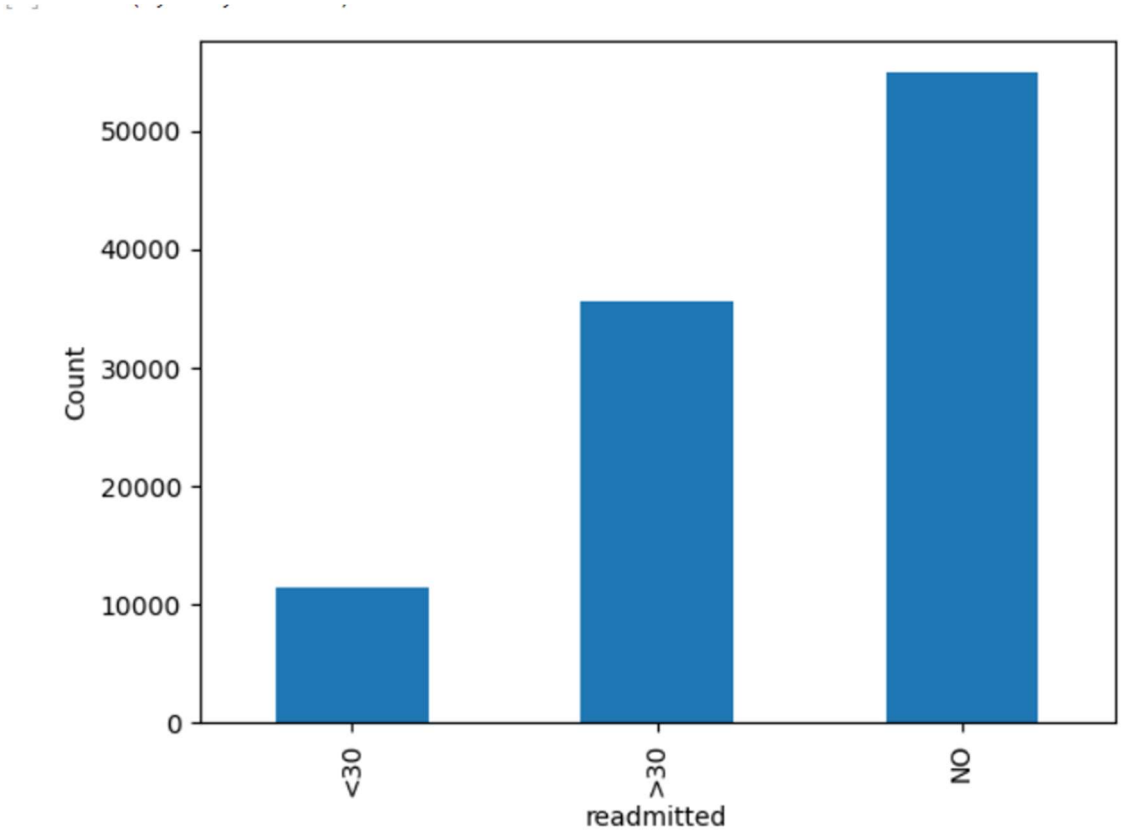
2)Admission type id to count



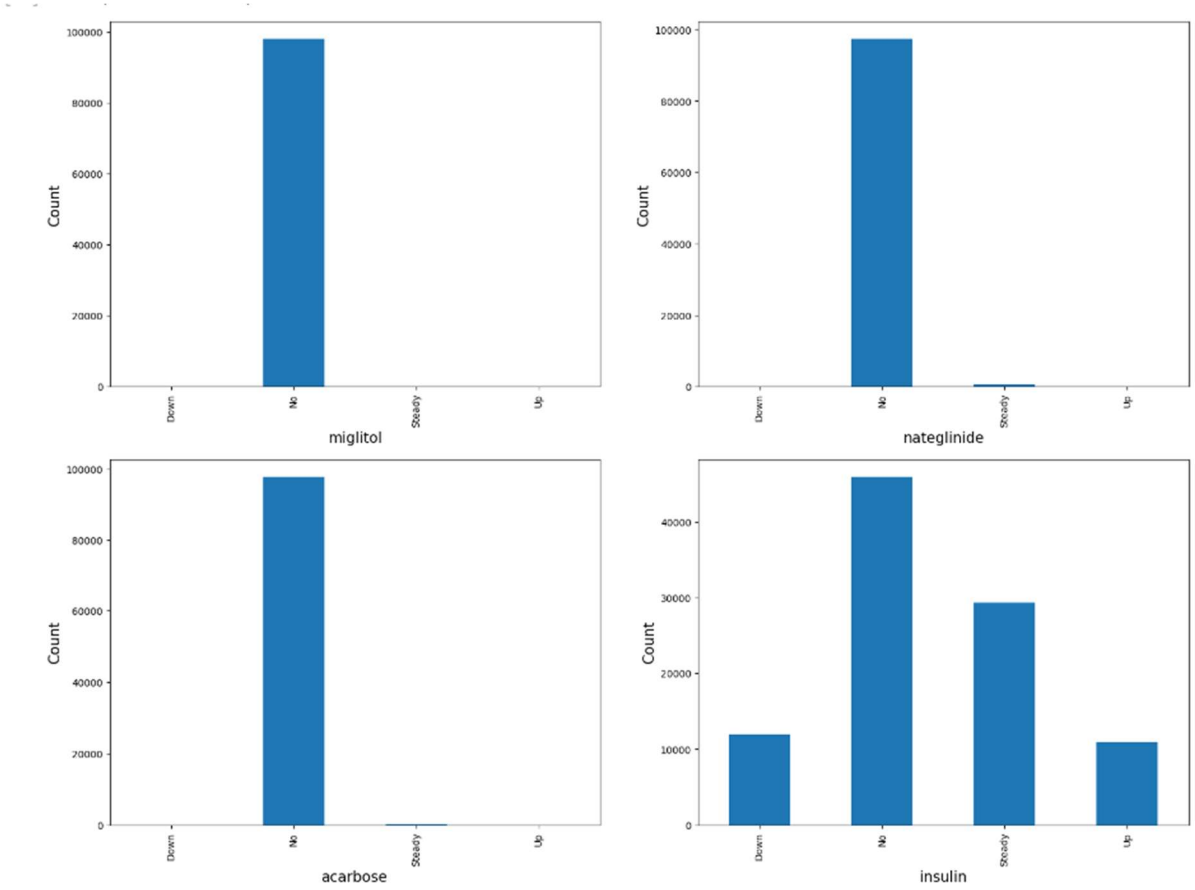
3)Discharge disposition id to count



4)Readmitted to count



5)Different medications of diabetes



APPLYING MACHINE LEARNING ALGORITHMS

In [33]:
```
# Random Forest
from sklearn.ensemble import RandomForestClassifier
clf1 = RandomForestClassifier()
RF_score = cross_val_score(clf1, X_cv, y_cv, cv=10, scoring='accuracy').mean()
RF_score
```

Out[33]: 0.6085885791020825

In [34]:
```
# Naive Bayes
from sklearn.naive_bayes import GaussianNB
clf2 = GaussianNB()
NB_score = cross_val_score(clf2, X_cv, y_cv, cv=10, scoring='accuracy').mean()
NB_score
```

Out[34]: 0.5989882709167862

In [35]:
```
# Logistic Regression
from sklearn.linear_model import LogisticRegression
clf3 = LogisticRegression()
LR_score = cross_val_score(clf3, X_cv, y_cv, cv=10, scoring='accuracy').mean()
LR_score
```

Out[35]: 0.6179578907127243

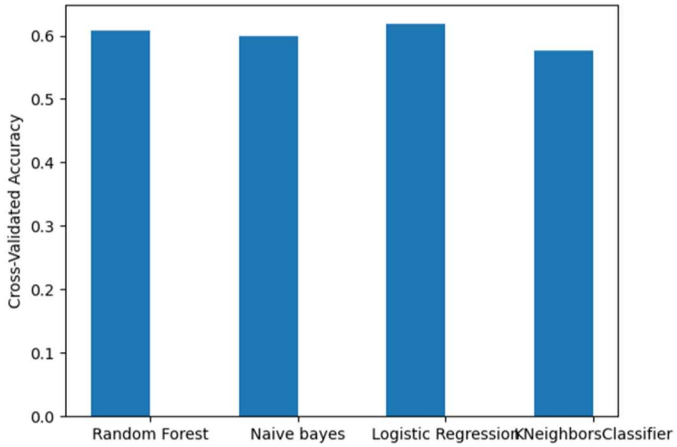
In [36]:
```
# KNeighbor
from sklearn.neighbors import KNeighborsClassifier
knc4 = KNeighborsClassifier()
kn_score = cross_val_score(knc4, X_cv, y_cv, cv=10, scoring='accuracy').mean()
kn_score
```

Out[36]: 0.5768368184302826

PERFORMANCE COMPARISON

In [38]:
```
# plot and compare the scores
# LR outperforms the other four a little bit
x_axis = np.arange(4)
y_axis = [RF_score, NB_score, LR_score,kn_score]
plt.bar(x_axis, y_axis, width=0.4)
plt.xticks(x_axis + 0.4/2., ('Random Forest', 'Naive bayes', 'Logistic Regression','KNeighborsClassifier'))
plt.ylabel('Cross-Validated Accuracy')
```

Out[38]:
```
Text(0, 0.5, 'Cross-Validated Accuracy')
```



In [39]:
```
# Logistic Regression on Top 6 features
# still be able to achieve good result with reduced running time
LR_score_top = cross_val_score(clf3, X_cv_top6, y_cv, cv=10, scoring='accuracy').mean()
LR_score_top
```

Out[39]:
```
0.6119610210918149
```

Conclusions

- Six major features are found to have high impact on diabetes patient readmission: number of lab procedures, number of medications administrated during the encounter, time spent in hospital, number of procedures other than lab tests, number of diagnoses, and number of inpatient visits.

- The logistic regression classifier modeling achieves 0.62 accuracy and 0.66 AUC score. The sensitivity of the modeling can be increased by adjusting the classification threshold.

- To correctly predict the readmission and avoid extra cost, hospitals should carefully examine the clinical data of patients and pay special attention to the above major features.

- Some other features might be worth collecting, for example, date of admission and family history.

- This analytic method can be applied to different diseases other than diabetes.

References

1. Strack, Beata, et al. "Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records." BioMed research international 2014 (2014).
2. Bhuvan, Malladihalli S., et al. "Identifying Diabetic Patients with High Risk of Readmission." arXiv preprint arXiv:1602.04257 (2016).
3. Sushmita, Shanu, et al. "Predicting 30-Day Risk and Cost of" All-Cause" Hospital Readmissions." Workshops at the Thirtieth AAAI Conference on Artificial Intelligence. 2016.