REPORT

Big Data & Predictive Analytics

Classification & Clustering

Prepared by

Prashant Singh
(Pre-final Data science & AI engineer
, IIIT Naya Raipur)

Dataset Overview

Using Diabetic patient data for developing a Machine Learning model for predicting hospital readmission within 30 days.

Hospital readmission is a real-world problem and an on-going topic for improving health care quality and a patient's experience, while ensuring cost-effectiveness. Information on the Hospital Readmissions Reduction Program (HRRP) is publicly available in CMS, Center for Medicare and Medicaid Services, web site.

The dataset, Diabetes 130-US hospitals for years 1999-2008 Data Set, was downloaded from UCI Machine Learning Repository. It represents 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks with 100,000 observations and 50 features representing patient and hospital outcomes.

The developed Machine Learning model is based on R and employs the package, SuperLearer, with ensemble learning to optimize the results. For computation needs, most of the ensemble learning ran on a Microsoft Azure public cloud, an E16 Virtual Machine with 16 vcpus and 128 GB RAM, as shown below. For a training set of 10,000 observations and 21 predictors, in general the model took about 2 to 3 hours to train and more than 6 hours to carry out 10-fold cross-validation with two algorithms. The demand for computing resources was significant.

Basic Explanation

It is important to know if a patient will be readmitted in some hospital. The reason is that you can change the treatment, in order to avoid a readmission.

In this database, you have 3 different outputs:

1. No readmission;
2. A readmission in less than 30 days (this situation is not good, because maybe your treatment was not appropriate);
3. A readmission in more than 30 days (this one is not so good as well the last one, however, the reason can be the state of the patient.

In this context, you can see different objective functions for the problem. You can try to figure out situations where the patient will not be readmitted, or if they are going to be readmitted in less than 30 days (because the problem can affect the treatment), etc… Make your choice and let's help them create new approaches for the problem.

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

Understanding Problem And Constraint

It is estimated that 9.3% of the population in the United States have diabetes , 28% of which are undiagnosed. The 30-day readmission rate of diabetic patients is 14.4 to 22.7 % . Estimates of readmission rates beyond 30 days after hospital discharge are even higher, with over 26 % of diabetic patients being readmitted within 3 months and 30 % within 1 year. Costs associated with the hospitalization of diabetic patients in the USA were $124 billion, of which an estimated $25 billion was attributable to 30-day readmissions assuming a 20 % readmission rate. Therefore, reducing 30-day readmissions of patients with diabetes has the potential to greatly reduce healthcare costs while simultaneously improving care.

Constraints:

- **Interpretability of model is important :** Interpretability is always important in healthcare domain if model predict that some patient will readmit but can't explain why it came to this conclusion the doctor will be clueless about such decision and also doctor won't be able to tell the patient why he needs to readmit practically it will create lots of inconvenience to doctor as well as patient.
- **Latency is not strictly important :** Most of the health care related applications are not latency dependent.

- **The cost of misclassification is high :** If the patient doesn't need to readmit if the model says “yes to readmit” that will put a financial burden on the patient. If a patient needs to readmit but the model says “no to readmit” then that will cause readmission cost to the hospital so, misclassification rate should be as low as possible.

Feature Description

- **Encounter ID** Unique identifier of an encounter
- **Patient number** Unique identifier of a patient
- **Race** Values: Caucasian, Asian, African American, Hispanic, and other
- **Gender** Values: male, female, and unknown/invalid
- **Age** Grouped in 10-year intervals: 0, 10), 10, 20), …, 90, 100)
- **Weight** Weight in pounds
- **Admission type** Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available
- **Discharge disposition** Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available
- **Admission source** Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital
- **Time in hospital** Integer number of days between admission and discharge
- **Payer code** Integer identifier corresponding to 23 distinct values, for example, Blue Cross/Blue Shield, Medicare, and self-pay Medical

- **Medical specialty** Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family/general practice, and surgeon
- **Number of lab procedures** Number of lab tests performed during the encounter
- **Number of procedures** Numeric Number of procedures (other than lab tests) performed during the encounter
- **Number of medications** Number of distinct generic names administered during the encounter
- **Number of outpatient visits** Number of outpatient visits of the patient in the year preceding the encounter
- **Number of emergency visits** Number of emergency visits of the patient in the year preceding the encounter
- **Number of inpatient visits** Number of inpatient visits of the patient in the year preceding the encounter
- **Diagnosis 1** The primary diagnosis (coded as first three digits of ICD 9); 848 distinct values
- **Diagnosis 2** Secondary diagnosis (coded as first three digits of ICD 9); 923 distinct values
- **Diagnosis 3** Additional secondary diagnosis (coded as first three digits of ICD 9); 954 distinct values
- **Number of diagnoses** Number of diagnoses entered to the system 0%
- **Glucose serum test result** Indicates the range of the result or if the test was not taken. Values: “>200,” “>300,” “normal,” and “none” if not measured
- **A1c test result** Indicates the range of the result or if the test was not taken. Values: “>8” if the result was greater than 8%, “>7” if the result was greater than 7% but less than 8%, “normal” if the result was less than 7%, and “none” if not measured.
- **Change of medications** Indicates if there was a change in diabetic medications (either dosage or generic name). Values: “change” and “no change”

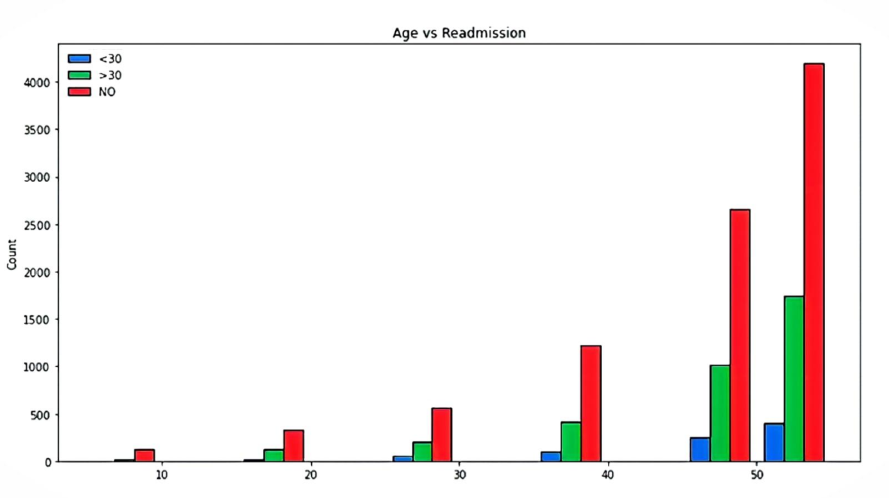
- **Diabetes medications** Indicates if there was any diabetic medication prescribed. Values: “yes” and “no”
- 24 different kinds of medical drugs.
- **Readmitted** Days to inpatient readmission. Values: “<30” if the patient was readmitted in less than 30 days, “>30” if the patient was readmitted in more than 30 days, and “No” for no record of readmission

Data Cleaning & Transformation

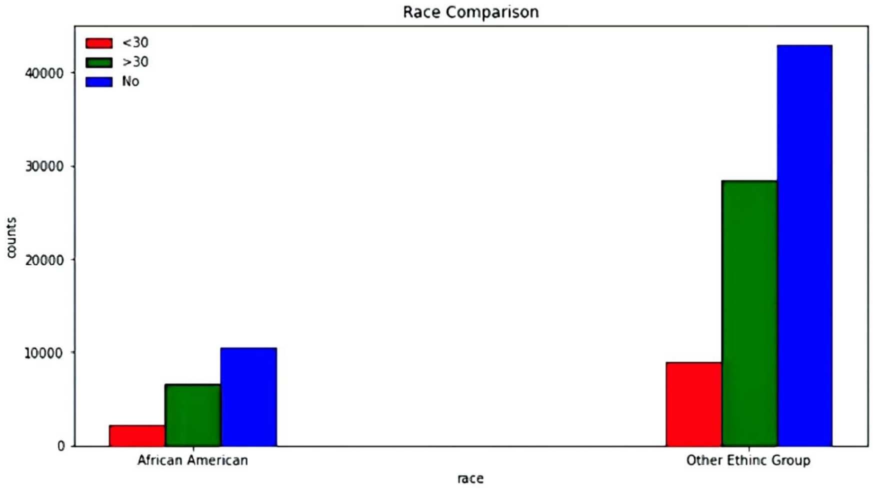
- Replacing all missing values (‘?’) with numpy.NaN .
- Dropping all columns with more than 50% values missing .
- Dropping all columns with more than 90% duplicate values .
- Transforming age column with mid values in given range .
- Fill missing values in diag_1, diag_2, diag_3 & dropping rest columns with missing values .
- Identifying Numerical & Categorical Features .
- Removing Outliers in Numerical Columns .
- Removing duplicate values in 'patient_nbr' column .

Data Visualization

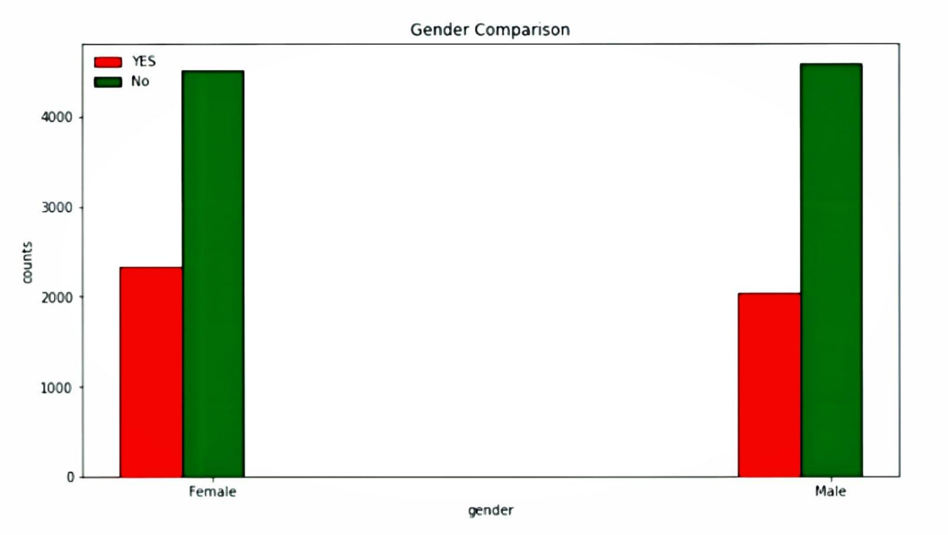
1. **Impact of Age on Readmission**：Most of the readmission of patients is from higher age .



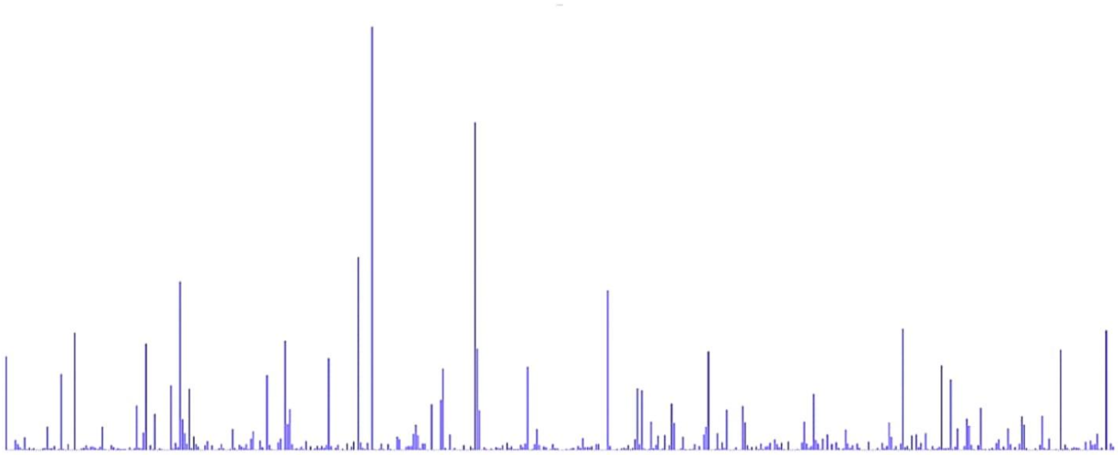
2. **Impact of African_Americans vs Other Ethnic Groups on Readmission**：African_Americans are less likely to readmit as compared to other ethnic groups .



3. **Impact of Women vs Men on Readmission** : Both women and men are equally likely to get readmitted with slight more chance for women readmission



4. Impact of Diagnosis type on Readmission



Models Used & Their Evaluation

1. K-fold cross validation using Decision tree and K nearest neighbor classifier models :

Cross-validation is a statistical method used to estimate the skill of machine learning models. It is commonly used in applied machine learning to compare and select a model for a given predictive modeling problem because it is easy to understand, easy to implement, and results in skill estimates that generally have a lower bias than other methods.

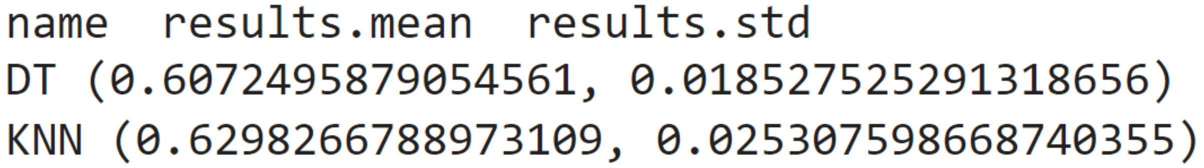
Decision tree and KNN model is used in our K-fold cross validation.

Accuracy achieved with these models ->

Decision Tree : 60%

KNN : 62%

Outputs:



```
name  results.mean  results.std
DT (0.6072495879054561, 0.018527525291318656)
KNN (0.6298266788973109, 0.025307598668740355)
```

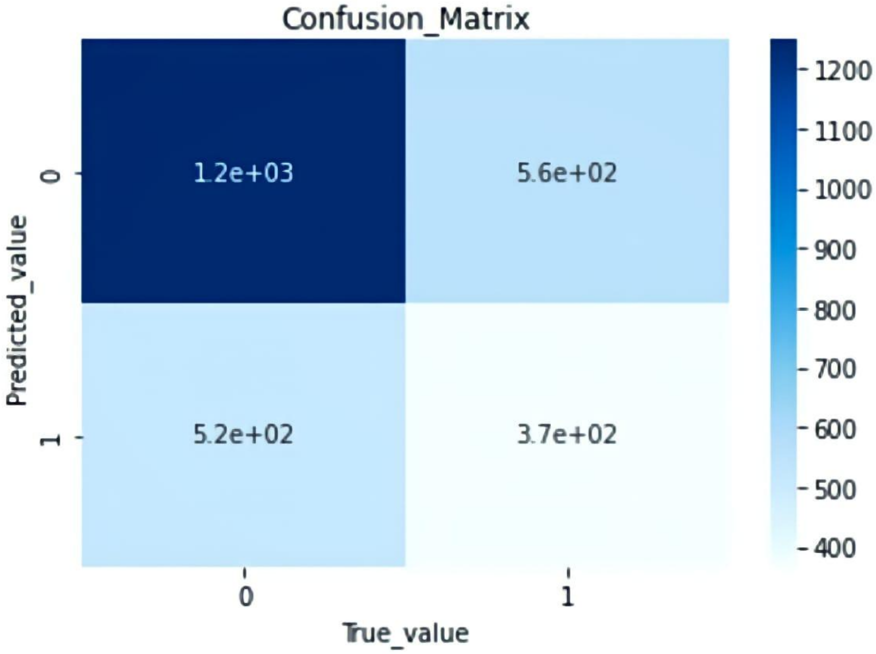
2. Plotting Confusion matrix of Decision tree :

The confusion matrix, precision, recall, and F1 score gives better intuition of prediction results as compared to accuracy .

Accuracy = TP/TP+FP , Recall = TP/TP+FN

F1_score = 2*precision*recall/precision+recall

Final accuracy = 60.7%

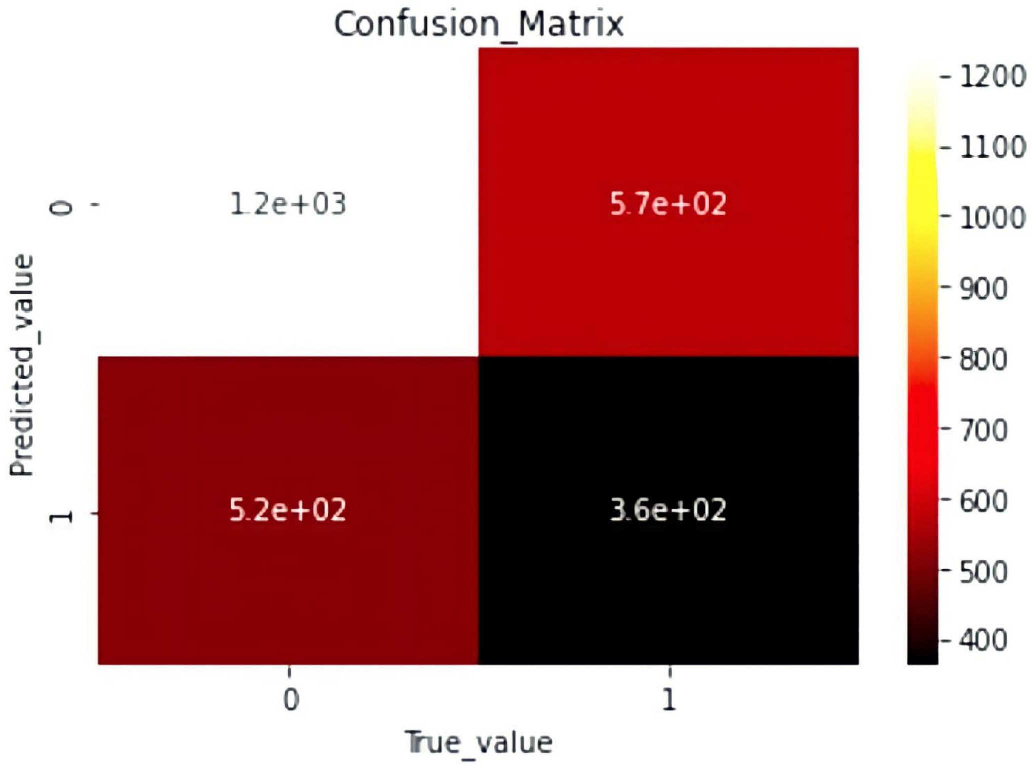


3. Improved accuracy with Logistic Regression model over diabetic dataset :

Using Logistic regression over diabetic dataset improves accuracy by almost 10%!!!

Even svm models give just 68% accuracy ! Logistic regression gives better accuracy than decision tree, KNN & svm classification models :)

Final Logistic Regression accuracy : 70.1%



K-Mean Clustering

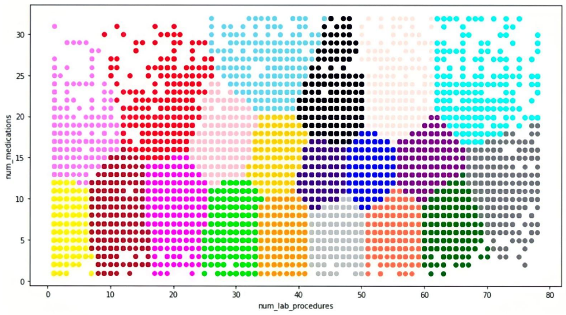
Applying K-mean clustering on each pair of features

Features : `['num_medications', 'number_outpatient', 'number_emergency',`

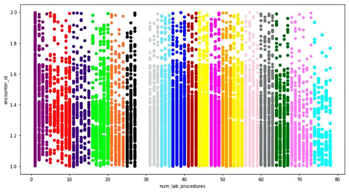
```
    'time_in_hospital', 'number_inpatient', 'encounter_id', 'age',

    'num_lab_procedures', 'number_diagnoses', 'num_procedures',]
```

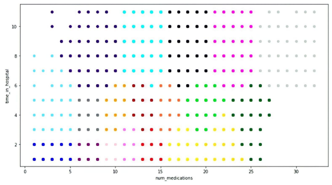
Visualizing Some important Clusters :



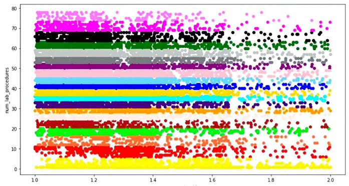
Num_lab_procedures - Num_medications



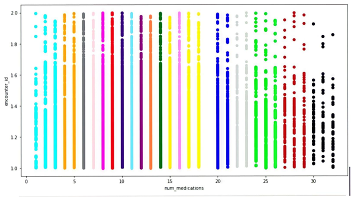
Num_lab_procedures - encounter_id



Num_medications - time_in_hospital



Encounter_id - Num_lab_procedures



Num_medications - Encounter_id

Conclusion

Applying various supervised classification models like logistic regression ,svm,KNN,Decision tree helped us to predict readmissions of diabetic patients !!!

Furthermore applying Unsupervised K-Mean clustering model over our Diabetic Dataset results in formation of various clusters ,visualizing these clusters gives us insight of what features or parameters we have to optimize to reduce the number of readmissions of diabetic patients .

Thanks