Project :

1. Description of Data:
2. What data is
   1. This dataset is for hospital admission for diabetic patients.
3. Where it comes from
   1. The dataset is preprocessed from an electronic medical records database based on the paper. The database contains data systematically collected from participating institutions electronic medical records and includes encounter data (emergency, outpatient, and inpatient), provider specialty, demographics (age, sex, and race), diagnoses and in-hospital procedures documented by ICD-9-CM codes, laboratory data, pharmacy data, in-hospital mortality, and hospital characteristics.
4. Extent (size, number of years covered, etc.)
   1. The dataset has 101,766 rows with 55 features

II. Question 1 — predicting patients readmission of people with diabetes

1. The data that will be used for this question (specific identification of fields, especially target)
   1. Target variable: readmitted
   2. Other fields: encounter\_id, patient\_nb, race, gender, age, weight, admission\_type\_id, discharge\_disposition\_id, admission\_source\_id, time\_in\_hospital, payer\_code, medical\_specialty, num\_lab\_procedures, num\_procedures, num\_medications, number\_outpatient, number\_emergency, number\_inpatient, diag\_1, diag\_2, diag\_3, number\_diagnoses, max\_glu\_serum, A1Cresult, metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, citoglipton, insulin, glyburide-metformin, glipizide-metformin, glimepiride-pioglitazone, metformin-rosiglitazone, metformin-pioglitazone, change, diabetesMed
2. Description of what you do with the data before modeling
3. Exploratory Data Analysis
4. Remove people who died or went to hospice. (SQL Joins)
5. Checking data imbalance.
6. Univariate distribution of all the features in the dataset.
   1. Look for feature distribution in each class.
   2. Or look for feature distribution regardless of the class.
7. Variable Transformations

Missing values.

* In terms of the percentage of missing values, the following columns can be extracted: Weight, Payer Code, and Medical Specialty
* For Diagnosis 3, the missing values are 1%. We might remove the entire row considering how much the dataset will be reduced to.

Log transformation of numerical variables.

1. Clustering and PCA, if applicable

Clustering on medication (OFFICE HOURS)

1. Feature Engineering

* Binarizing
* Encoding the categorical features to numerical values.
* Use correlation matrix to find variables that have the highest correlation with the target variable readmitted
* Add or Multiply variables to see if the resultant feature has a better correlation with the target variable

1. Description of Modeling to be done
   1. Type of model
      1. Logistic regression baseline model with only significant features from correlation matrix.
      2. xgboost
      3. Naive Bayes model/QDA
      4. SVM
      5. Decision trees
   2. What will be used as a baseline model
      1. Logistic Regression
2. Analysis of models - Description of how to
   1. Error analysis to be done (OFFICE HOURS)
      1. Test error on test data
   2. Comparison to baseline
      1. Accuracy, Specificity, Sensitivity
      2. Use confusion matrix
   3. Visualization of results
      1. Plot ROC curves for the models.