# 30-Day Diabetes Readmission

Amy Liu, Emmanuel Ekwebelem, Stephanie Low, Wenjin Kuang

1

## TABLE OF CONTENTS



2

# 01.

## Project Scope

3

**Project name:**
Reducing 30-day Readmission Rate for diabetic patients

**Project Objective:**
To reduce the 30-day readmission rate for diabetic patients by analyzing a large dataset of diabetic patient records and designing a predictive model using machine learning algorithms to identify patients at high risk of readmission.

**Success Criteria:**

- Identifies leading variables of readmission
- Predictive model is accurate and reliable
- Project milestones are met on time
- Successful integration of predictive models into hospital's workflow
- Stakeholders are satisfied

**Approach:**
Python program will be used to analyze the dataset of diabetic patient records to determine the leading variables of readmission. A predictive model that identifies high risk patients will be designed and deployed. Tableau program will be used to visualize the work done in building the diabetes remission predictive model.

4

## Scope Statement

**Project Name:** Reducing 30-day Readmission Rate for diabetic patients

**Product Characteristics and Requirements:**

This project will use Python to analyze a dataset of diabetic patient records and design a predictive analysis through machine learning algorithms to identify patients at high risk of readmission so the hospital can take corrective actions and reduce the 30-day readmission rate.

**Out of Scope:**

This project will not consider additional datasets to be used for predictive analysis and will only analyze for factors contributing to the reduction of the 30-day readmission rate for diabetic patients

**Product User Acceptance Criteria:**

- The model has a high degree of accuracy and reliability in identifying high-risk patients
- Model results in improved patient outcomes through a reduction in the 30-day readmission rate for diabetic patients
- The model is HIPAA compliant and protects patients' data

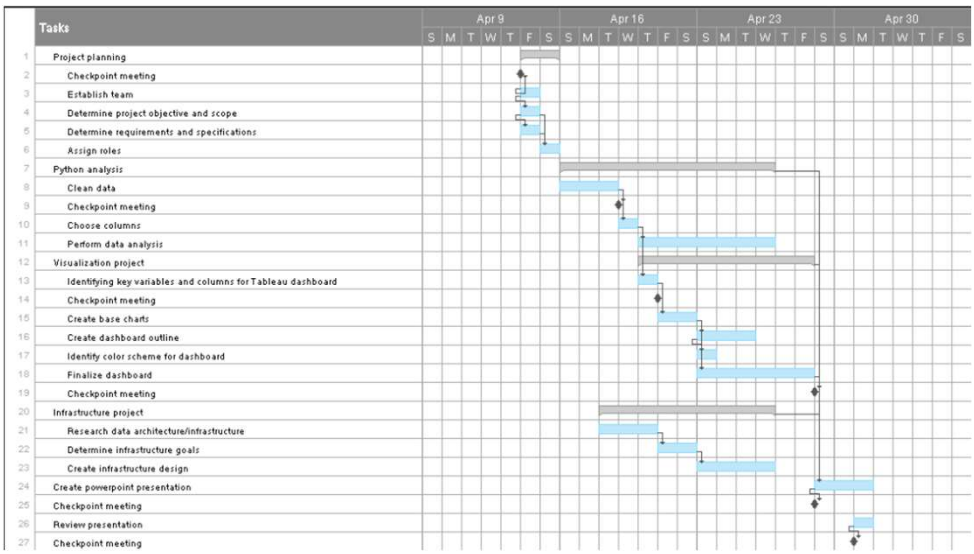
## Work Breakdown Structure (WBS)

**Project Name:** Reducing 30-day Readmission Rate for diabetic patients

1. Project planning
   - 1.1. Project objective and scope
   - 1.2. Project stakeholders
   - 1.3. Project budget
   - 1.4. Project team
   - 1.5. Project schedule
   - 1.6. Risk register
2. Project analysis
   - 2.1. Needs assessment
   - 2.2. Requirements and specifications
   - 2.3. Constraints and limitations
3. Powerpoint presentation
   - 3.1. Python project
     - 3.1.1. Cleaned dataset
       - 3.1.1.1. Variables required to predict readmission
     - 3.1.2. Predictive analysis
       - 3.1.2.1. Correlation
       - 3.1.2.2. Linear regression
       - 3.1.2.3. Logistic regression
       - 3.1.2.4. Random forest
       - 3.1.2.5. KMean
       - 3.1.2.6. Decision Tree
   - 3.2. Infrastructure project
     - 3.2.1. Infrastructure design
   - 3.3. Visualization project
     - 3.3.1. Tableau dashboard

5

# Project Schedule 4/14/2023- 5/2/2023




Key Milestones:

4/14/2023
Establishing team

4/18/2023
Data cleaning completion

4/28/2023
Finalizing dashboard

5/1/2023
Completing presentation

6

# 02. Background

Diabetes is a chronic medical condition characterized by high blood sugar levels due to inadequate insulin production or ineffective use of the body's insulin. It can be caused-by or lead-to various complications affecting the heart, kidneys, eyes, and nerves. Proper management through medication, diet, exercise, and regular monitoring is essential to prevent long-term health issues.

7

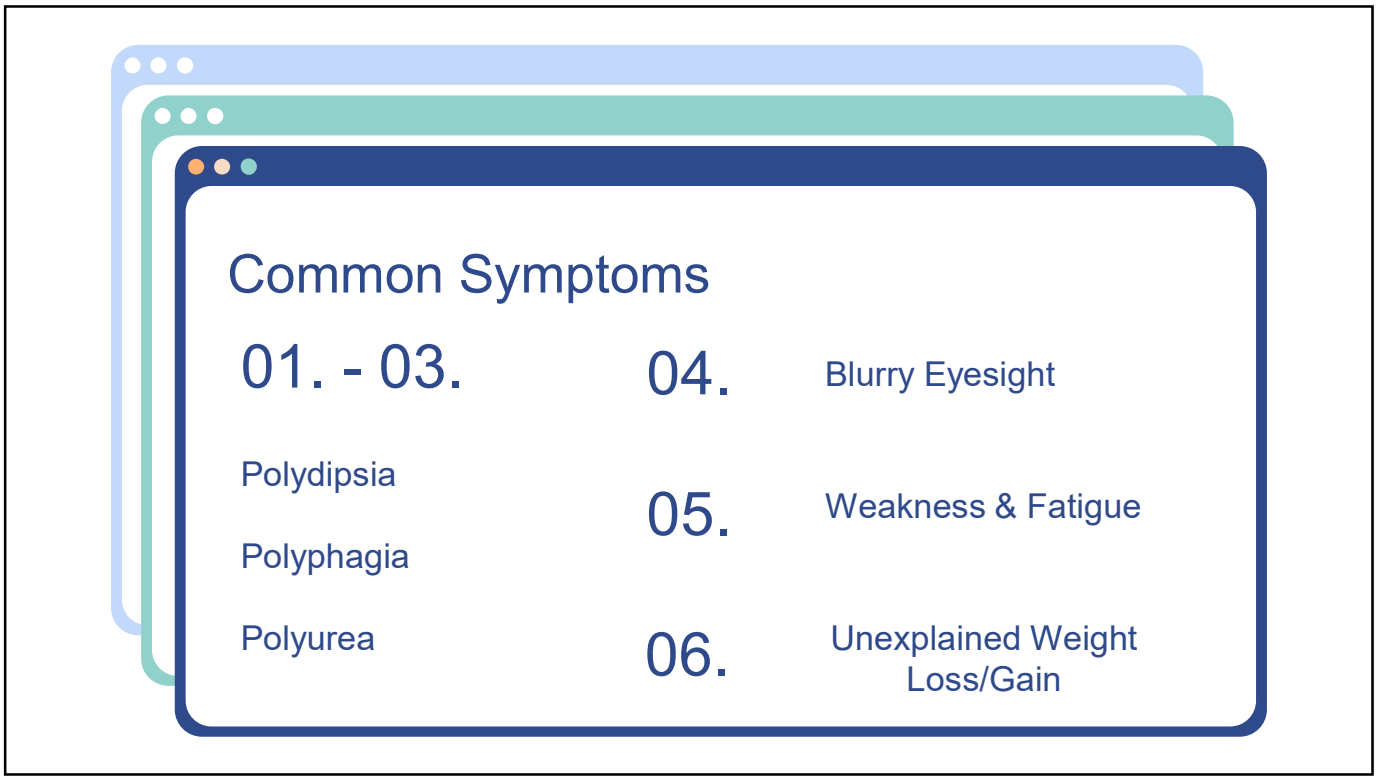
## Type 1

(Hereditary) Immune system attacks the cells that produce insulin, so the body is unable to make insulin

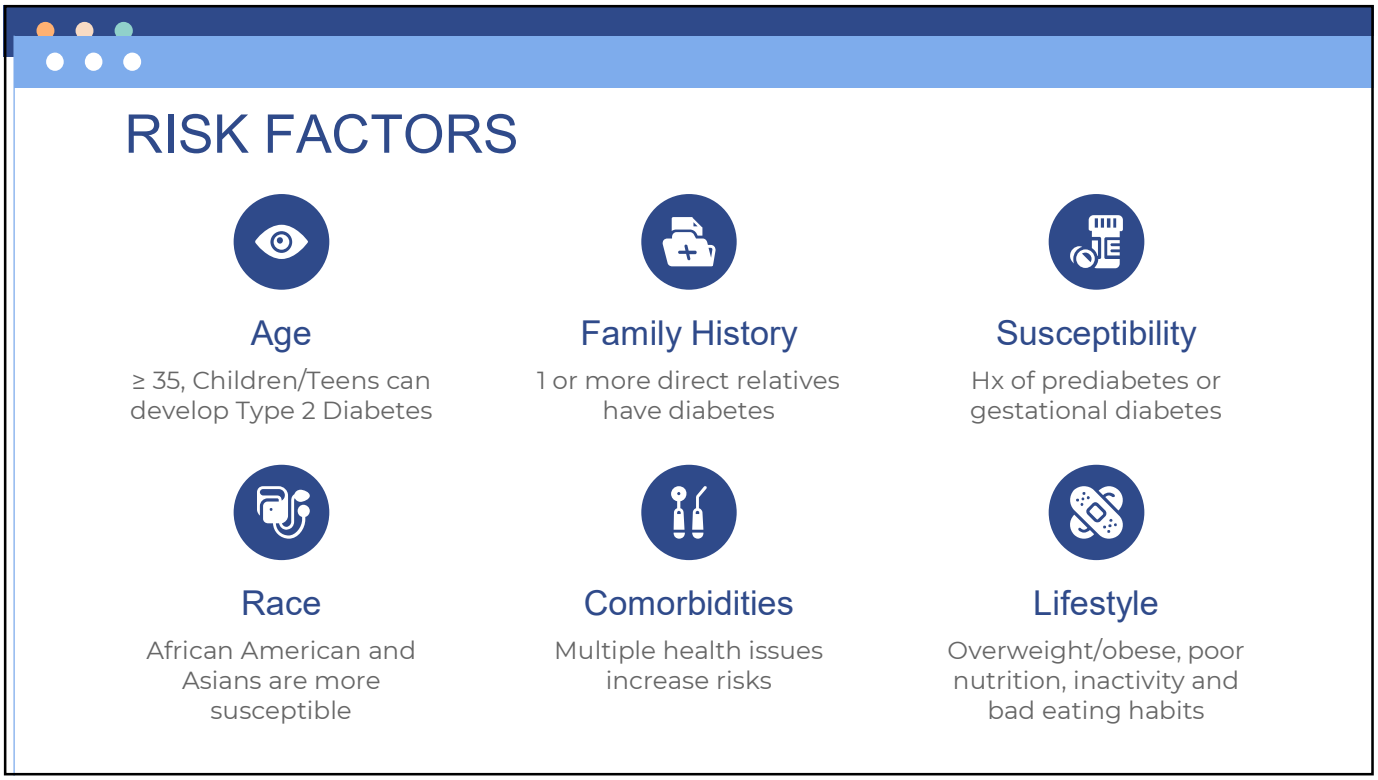
## Type 2

(Lifestyle) The cells of the body do not respond to insulin produced by pancreas

8


Common Symptoms
01. - 03.
Polydipsia
Polyphagia
Polyurea
04.
Blurry Eyesight
05.
Weakness & Fatigue
06.
Unexplained Weight Loss/Gain

9


RISK FACTORS
Age
≥ 35, Children/Teens can develop Type 2 Diabetes
Family History
1 or more direct relatives have diabetes
Susceptibility
Hx of prediabetes or gestational diabetes
Race
African American and Asians are more susceptible
Comorbidities
Multiple health issues increase risks
Lifestyle
Overweight/obese, poor nutrition, inactivity and bad eating habits

10

# KEY STATISTICS

## 37.3 Million Americans

28.7 million are diagnosed, 8.5 million are not

## $327 Billions/Year

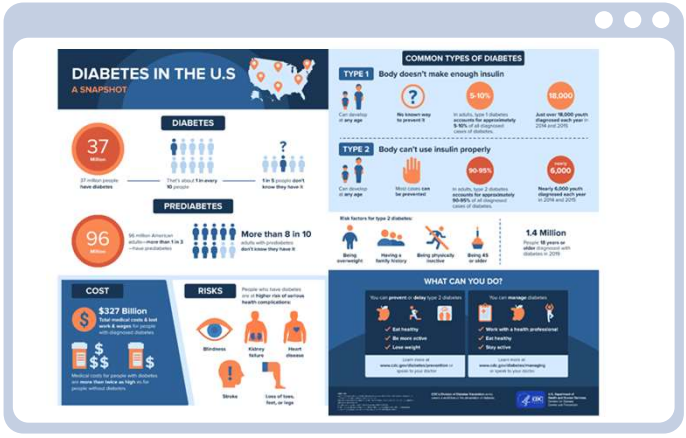
$1 out of every $4 in US health care costs is spent on caring for people with diabetes.

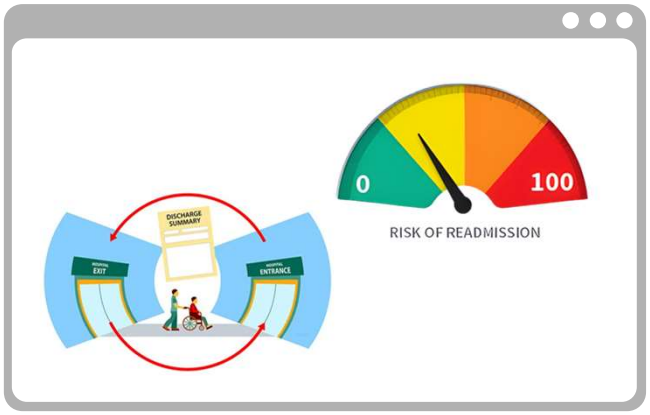
## 22 -26% RR

General Hospital Readmission Rates: 8.5-13.5%

11

# Significance




12

# Diabetic Data

*From : UCI ML Repository*

10 Years ( 1999 - 2008) of Clinical Diabetic Data from 130 US Hospitals. To be used to construct (test, train, and deploy) a ML Model for 30-Day Diabetic Based Readmissions Prediction

13

# Diabetic Data

*From : UCI ML Repository*

Includes patient clinical details such as :

- Demographics
- Diagnoses
- Test Results
- Medications
- Hospice Information

14

## Key Metrics (Demographics)

Caucasian, African American, Asian, Hispanic & Other

**Race**

Groups of 10 from 0 - 80 Years of Age

**Age**

Male Vs. Female

**Gender (Sex)**

Groups of 25 from 0 - 200, and ≥ 200

**Weight**

15

## Key Metrics (Hospice)

Admission Designation i.e., Emergency, Urgent, Newborn

**Admission Type**

Specialty Designation i.e., Endocrinology, OBGyn, Pediatrics

**Medical Specialty**

Discharge Designation i.e., Home, Acute Care, Another Facility

**Discharge Type**

Medical Diagnoses (ICD-9) i.e., [240-279] Endocrine and Immunity Disorders

**Diagnoses**

16

## Key Metrics (Medication)

Insulin Medication “Rate” i.e., Steady, Down, Up,

INSULIN

Metformin Medication “Rate” i.e., Steady, Down, Up,

METFORMIN

Blood Sugar Level (mg/dl) i.e., Normal [ 150], Prediabetic [≥ 200], Diabetic [≥ 300]

MAX GLU SERUM

Hemoglobin A1C Level i.e., Normal [ 150], Prediabetic [≥ 7], Diabetic [≥ 8]

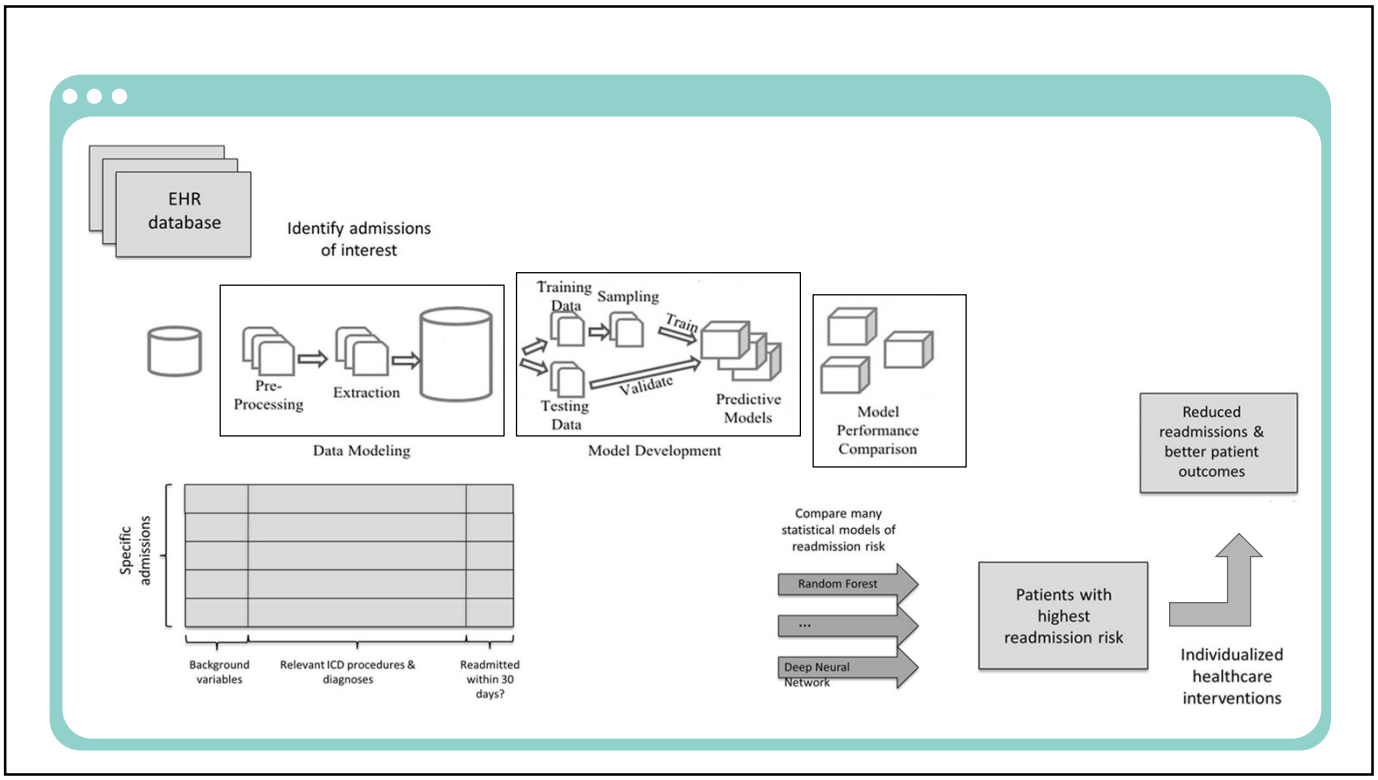
A1CRESULT

17

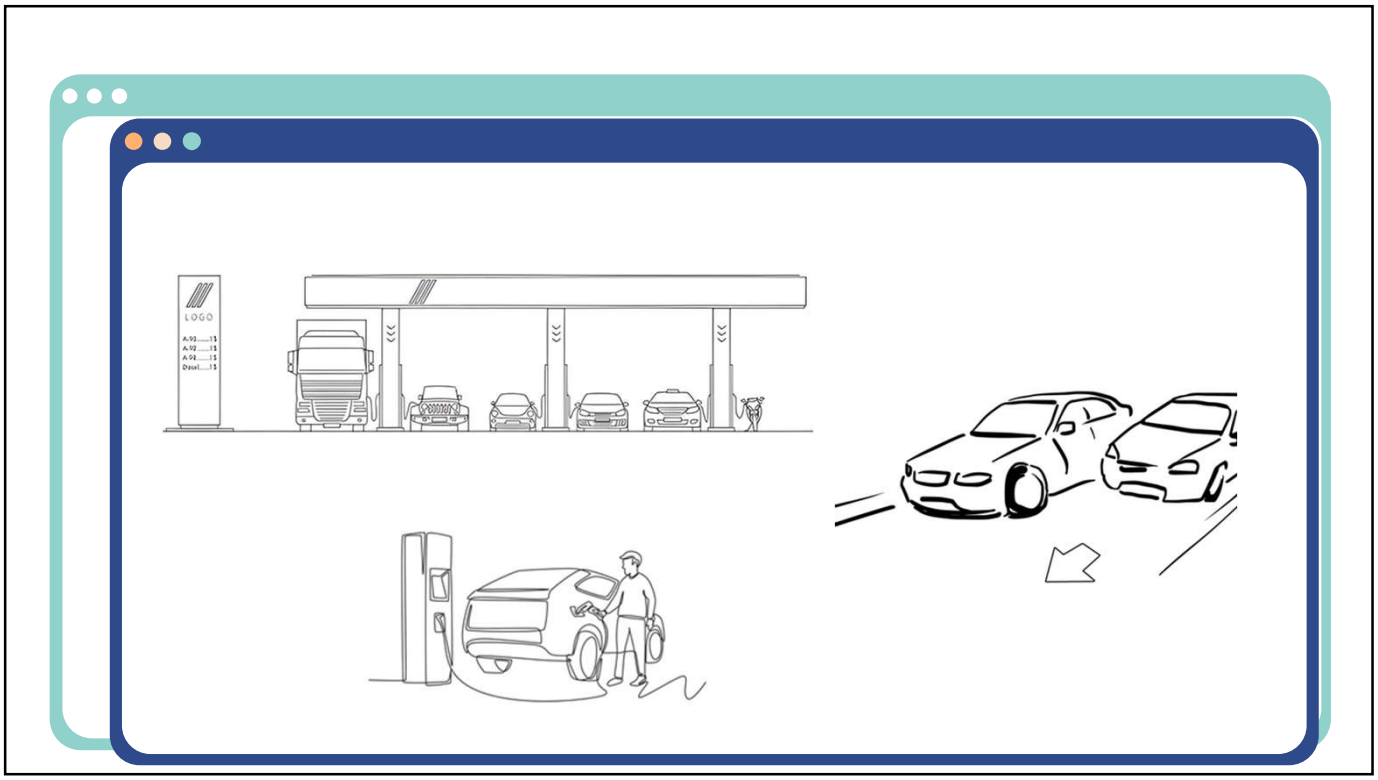
# 03.

## Infrastructure Project

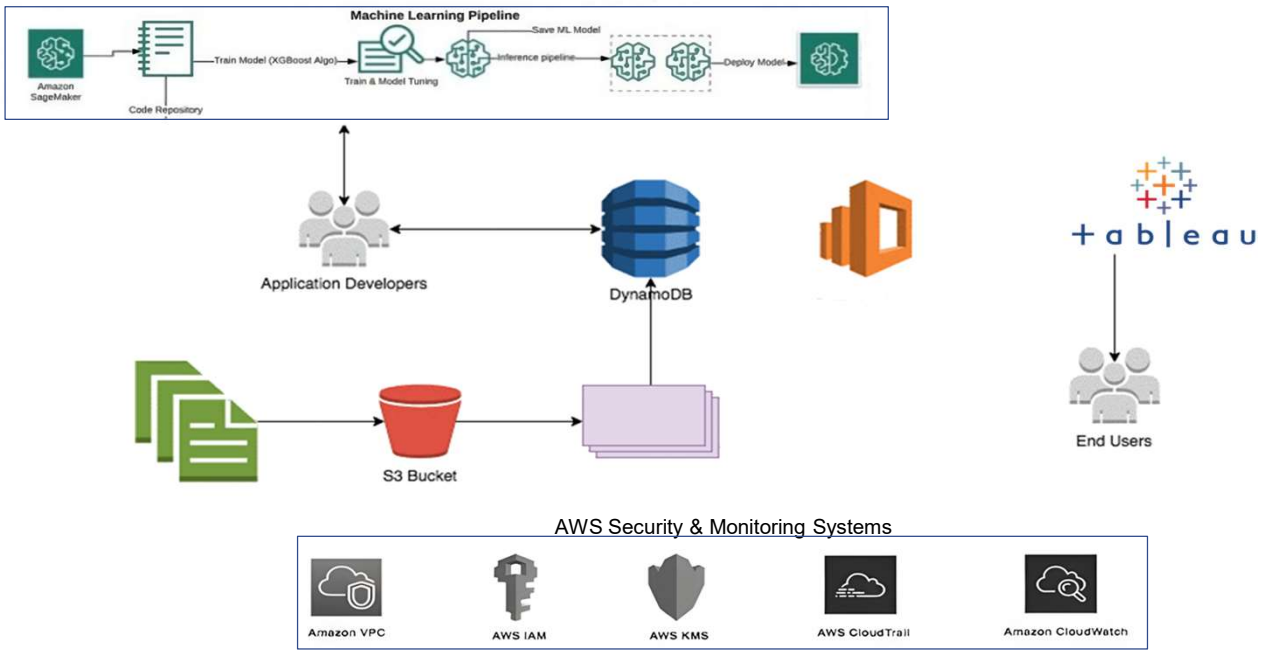
18


EHR database
Identify admissions of interest
Pre-Processing
Extraction
Data Modeling
Training Data
Sampling
Train
Testing Data
Validate
Predictive Models
Model Development
Model Performance Comparison
Reduced readmissions & better patient outcomes
Specific admissions
Background variables
Relevant ICD procedures & diagnoses
Readmitted within 30 days?
Compare many statistical models of readmission risk
Random Forest
...
Deep Neural Network
Patients with highest readmission risk
Individualized healthcare interventions

19



20


Machine Learning Pipeline
Amazon SageMaker
Code Repository
Train Model (XGBoost Algo)
Train & Model Tuning
Save ML Model
Inference pipeline
Deploy Model
Application Developers
DynamoDB
tableau
End Users
S3 Bucket
AWS Security & Monitoring Systems
Amazon VPC
AWS IAM
AWS KMS
AWS CloudTrail
Amazon CloudWatch

21


04.
Python
Project
HUMAN ORGAN

22

# Python Project Overview

### Data Cleaning

Grouping Data, Missing Data, etc.

### Feature Selection

Feature Importance, Correlation Heat Map, Recursive Feature Elimination

### Exploratory Analysis

Review Missing Data, Categorical Data, Numerical Data

### ML + Hyperparameter Tuning

Logistic Regression, Random Forest, etc.

23

# Data Cleaning

### Re-Admitted

Final Variables: 'YES' and 'NO'

### Admission Source ID

'Physician Referral' 'Emergency Room' 'Clinic Referral' 'Transfer from a hospital' 'Transfer from a Skilled Nursing Facility (SNF)', 'Transfer from another health care facility' nan 'HMO Referral' 'Court/Law Enforcement' 'childbirth', 'Transfer from critical access hospital' 'Transfer from hospital inpt/same fac reslt in a sep claim', 'Transfer from Ambulatory Surgery Center'

### Admission Type ID

Final Variables: NULL 'Emergency' 'Urgent' 'Elective' 'Newborn' 'Trauma Center'

24

# Data Cleaning

**discharge disposition ID**

nan
'home'
'acute_care'
'expired'
'left_AMA'
'hospice'
'other_healthcare_facility'

**Medical Specialities**

'surgery'
'pediatrics'
'orthopedics'
'ob_gyn'
'psychiatry'
'primary_care'
'cardiology'
'oncology'
'endocrinology'
'neurology'
'anesthesiology'
'radiology'

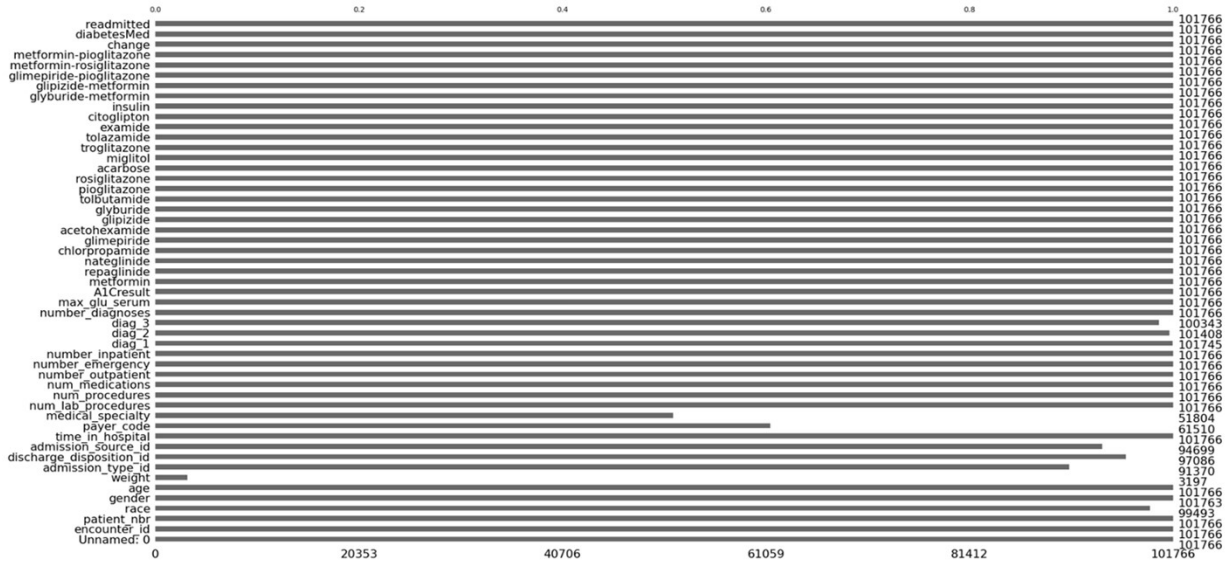
25

# Data Cleaning - Diagnostic 1, 2, 3 Columns

'ENDOCRINE, NUTRITIONAL AND METABOLIC DISEASES, AND IMMUNITY DISORDERS'
'COMPLICATIONS OF PREGNANCY, CHILDBIRTH, AND THE PUERPERIUM'
'INFECTIOUS AND PARASITIC DISEASES' 'NEOPLASMS'
'DISEASES OF THE CIRCULATORY SYSTEM' 'DISEASES OF THE RESPIRATORY SYSTEM'
'INJURY AND POISONING' 'DISEASES OF THE SKIN AND SUBCUTANEOUS TISSUE'
'DISEASES OF THE MUSCULOSKELETAL SYSTEM AND CONNECTIVE TISSUE'
'DISEASES OF THE DIGESTIVE SYSTEM'
'SUPPLEMENTARY CLASSIFICATION OF FACTORS INFLUENCING HEALTH STATUS AND CONTACT WITH HEALTH SERVICES'
'SYMPTOMS, SIGNS, AND ILL-DEFINED CONDITIONS'
'DISEASES OF THE GENITOURINARY SYSTEM'
'MENTAL, BEHAVIORAL AND NEURODEVELOPMENTAL DISORDERS'
'DISEASES OF THE NERVOUS SYSTEM AND SENSE ORGANS'
'DISEASES OF THE BLOOD AND BLOOD-FORMING ORGANS' nan
'CONGENITAL ANOMALIES'
'SUPPLEMENTARYCLASSIFICATION OF EXTERNAL CAUSES OF INJURY AND POISONING'

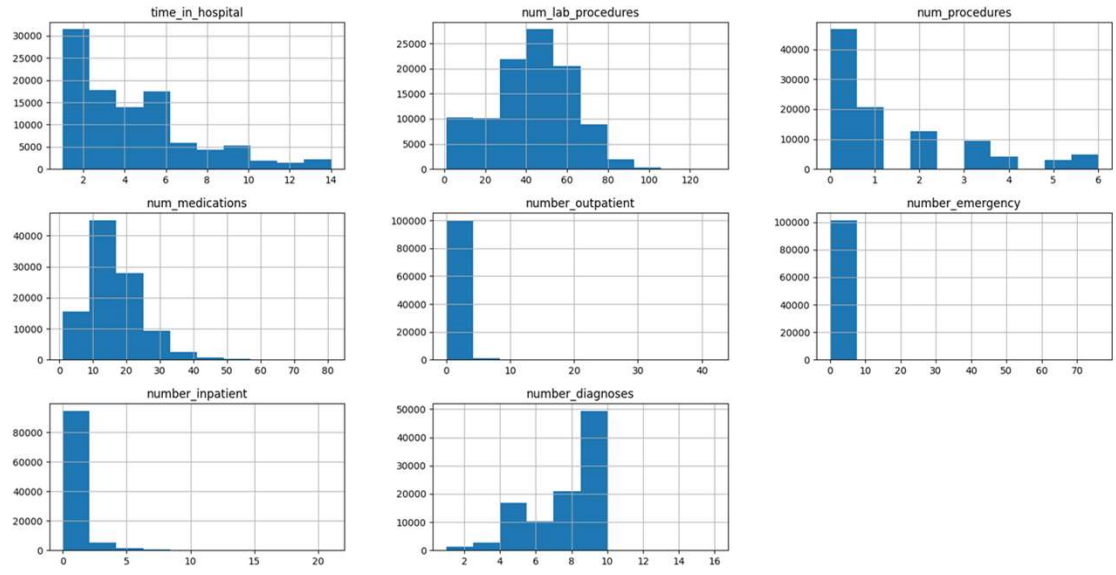
26

# Exploratory Analysis - Missing Values



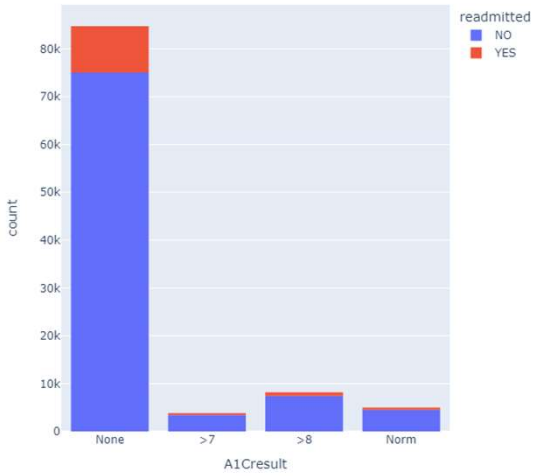

27

# Exploratory Analysis - Numerical Features




28

# Exploratory Analysis - Bivariate Analysis




An A1C result of >7 has a probability of 10.05 % to be readmitted

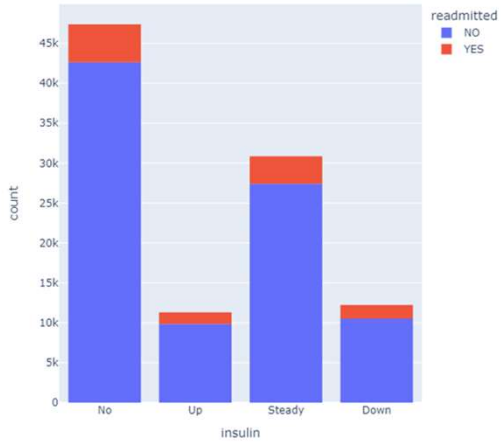
An A1C result of >8 has a probability of 9.87 % to be readmitted

An A1C result of Norm has a probability of 9.66 % to be readmitted

An A1C result of None has a probability of 11.42 % to be readmitted

29

# Exploratory Analysis - Bivariate Analysis cont.




An Insulin result of No has a probability of 10.04 % to be readmitted

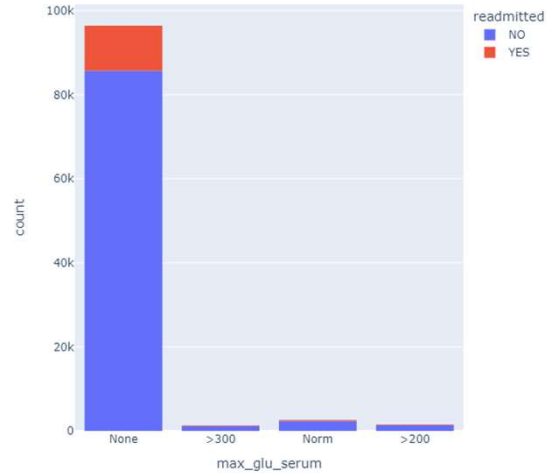
An Insulin result of Up has a probability of 12.99 % to be readmitted

An Insulin result of Steady has a probability of 11.13 % to be readmitted

An Insulin result of Down has a probability of 13.9 % to be readmitted

30

# Exploratory Analysis - Bivariate Analysis cont.




A max glucose serum result of None has a probability of 11.09 % to be readmitted

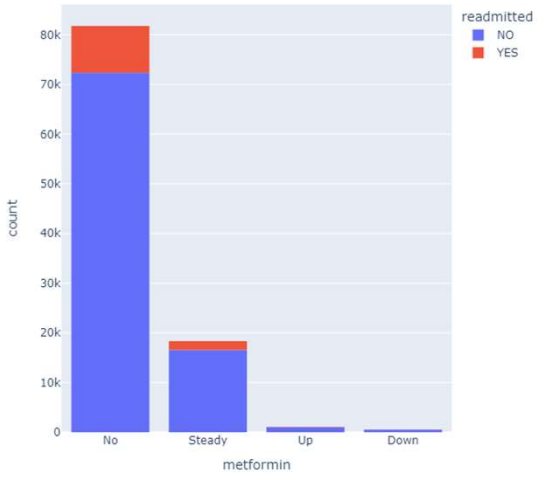
A max glucose serum result of >300 has a probability of 14.32 % to be readmitted

A max glucose serum result of Norm has a probability of 11.36 % to be readmitted

A max glucose serum result of >200 has a probability of 12.46 % to be readmitted

31

# Exploratory Analysis - Bivariate Analysis cont.




A metformin result of No has a probability of 11.52 % to be readmitted

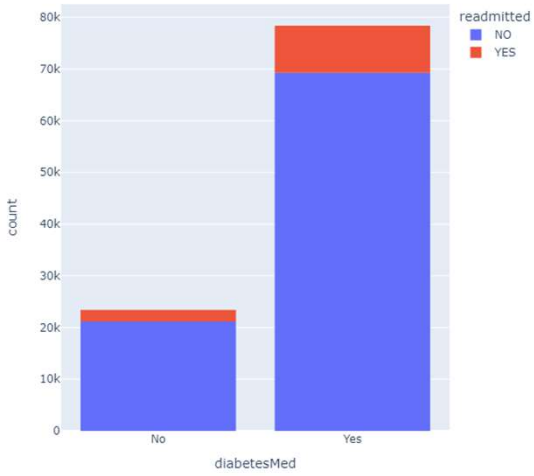
A metformin result of Steady has a probability of 9.71 % to be readmitted

A metformin result of Up has a probability of 8.25 % to be readmitted

A metformin result of Down has a probability of 12.0 % to be readmitted

32

# Exploratory Analysis - Bivariate Analysis cont.

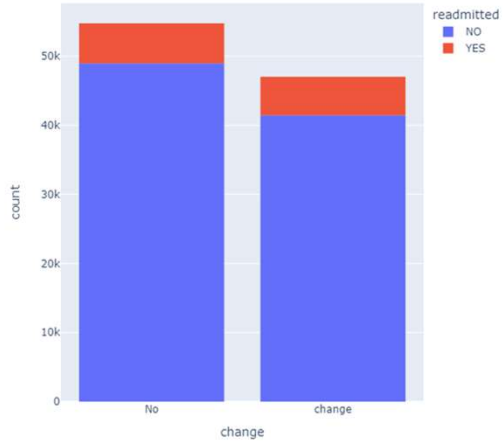



A diabetesMed value of No has a probability of 9.6 % to be readmitted

A diabetesMed value of Yes has a probability of 11.63 % to be readmitted

33

# Exploratory Analysis - Bivariate Analysis cont.

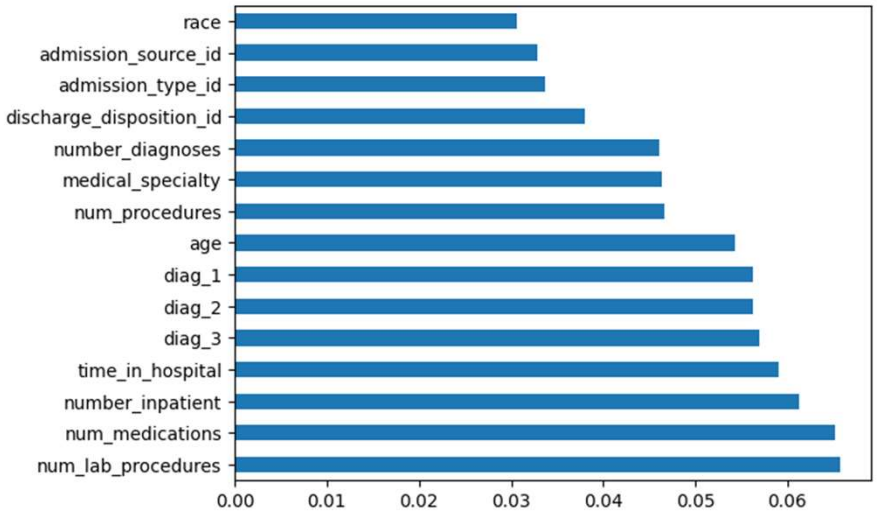



A change value of No has a probability of 10.59 % to be readmitted

A change value of change (yes) has a probability of 11.82 % to be readmitted

34

# Feature Selection - Mutual Importance




35

# Feature Selection - Recursive Feature Elimination

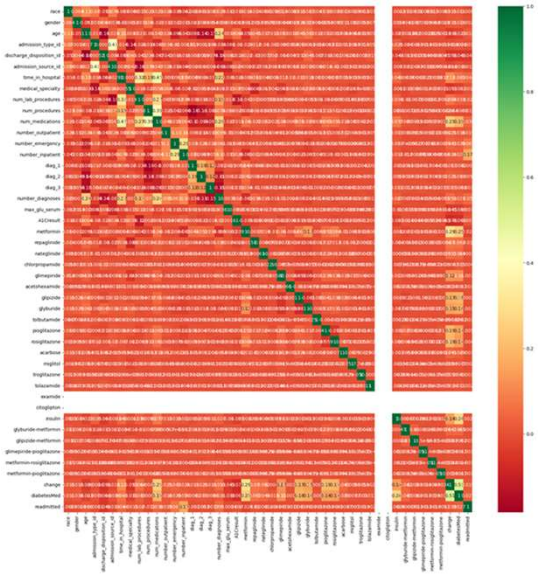
```
y = y_train
X = X_train

from sklearn.feature_selection import RFE
from sklearn.linear_model import LogisticRegression

logreg = LogisticRegression()
rfe = RFE(estimator=LogisticRegression(), n_features_to_select=15)
rfe = rfe.fit(X, y)
print(rfe.support_)
print(rfe.ranking_)
```

```
[False False False False  True False  True False False False False False
 False  True False False False  True  True False  True  True  True  True
  True False False False  True False False  True  True False  True False
 False False False False False False False False  True]
[ 8 18  7 15  1 12  1 19 23 10  9 17  3  1 14 20 21  1  1  4  1  1  1  1
  1 26  2 11  1  6 16  1  1 31  1 25 28  5 13 24 29 30 27 22  1]
```

**Important Columns:**
'diabetesMed'
'tolazamide'
'miglitol'
'acarbose'
'tolbutamide'
'glimepiride'
'chlorpropamide'
'nateglinide'
'repaglinide'
'metformin
','max_glu_serum'
'number_diagnoses'
'number_inpatient'
'time_in_hospital'
'discharge_disposition_id'

36

# Feature Selection - Correlation Matrix



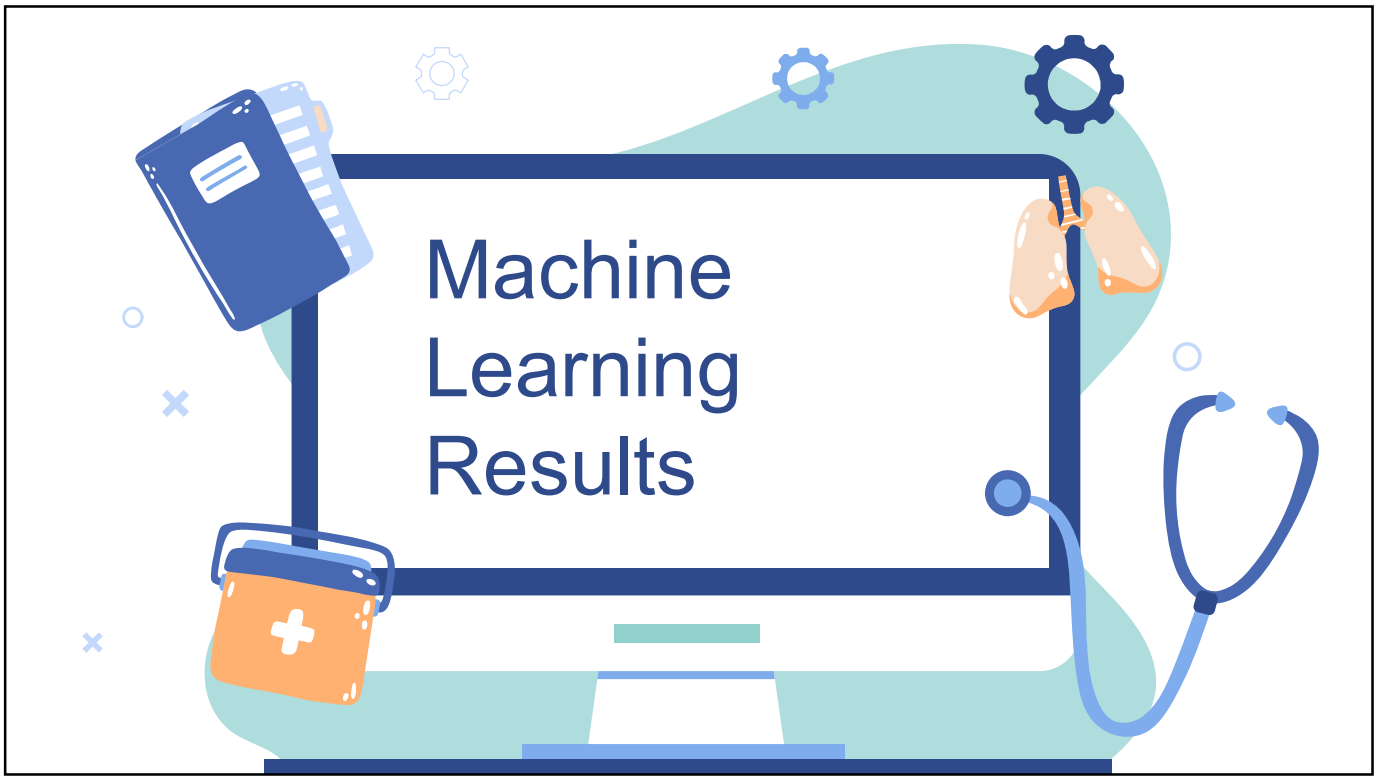
**Important Columns:**
'number_inpatient'
'diabetes_med'
'glimepiride'
'glipizide'
'insulin'
'change
'diabetes_med'
'glyburide'
'number_diagnoses'
'admission_type_id'
'time_in_hospital'
'admission_source_id'
'num_lab_procedures'
'num_procedures'
'num_medications'

37

# Final Columns

'time_in_hospital'
'num_lab_procedures'
'num_procedures'
'num_medications'
'number_outpatient',
'number_emergency'
'number_inpatient'
'number_diagnoses'
'discharge_disposition_id'
'diabetesMed'
'max_glu_serum'
'change','metformin',
'diag_1',
'diag_2',
'diag_3'

38




39

## Results - Without Feature Selection

| | AUC | Accuracy | Recall | Precision | Specificity | Prevalence |
|---|---|---|---|---|---|---|
| Logistic Regression | 0.663 | 0.646 | 0.573 | 0.172 | 0.655 | 0.111 |
| KNN | 0.644 | 0.646 | 0.544 | 0.166 | 0.619 | 0.111 |
| Stochastic Gradient Descent | 0.654 | 0.655 | 0.556 | 0.173 | 0.667 | 0.111 |
| Decision Tree | 0.632 | 0.628 | 0.594 | 0.168 | 0.631 | 0.111 |
| Random Forest | **0.669** | 0.640 | **0.618** | **0.178** | 0.643 | 0.111 |
| Linear SVC | 0.662 | **0.879** | 0.094 | 0.340 | 0.977 | 0.111 |
| Gradient Boosting | 0.649 | 0.609 | 0.614 | 0.164 | 0.609 | 0.111 |
| XG Boost | 0.657 | 0.613 | 0.617 | 0.166 | 0.613 | 0.111 |
| Cat Boost | 0.648 | 0.602 | 0.610 | 0.161 | 0.601 | 0.111 |

40

## Results - Feature Selection

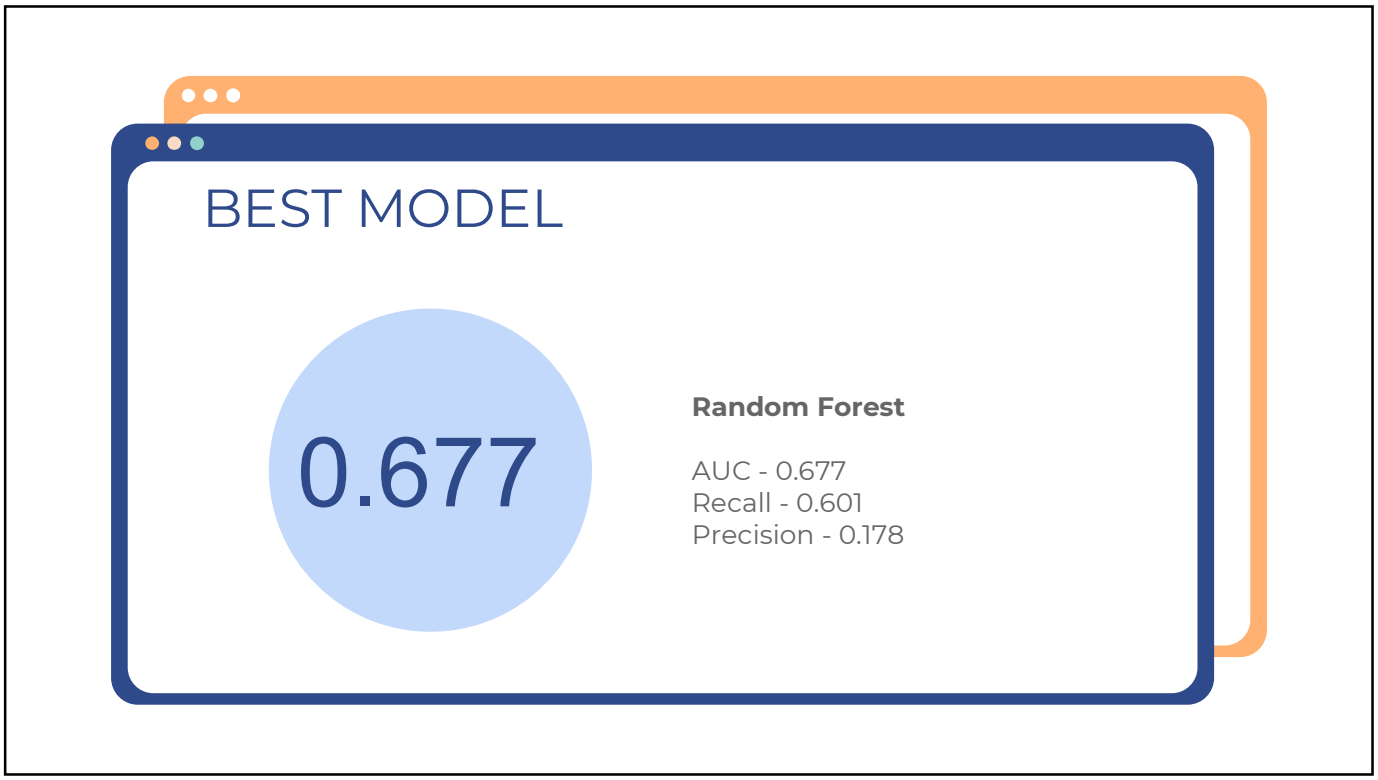
| | AUC | Accuracy | Recall | Precision | Specificity | Prevalence |
|---|---|---|---|---|---|---|
| Logistic Regression | 0.663 | 0.660 | 0.569 | 0.178 | 0.671 | 0.111 |
| KNN | 0.652 | 0.658 | 0.535 | 0.170 | 0.640 | 0.111 |
| Stochastic Gradient Descent | 0.653 | 0.661 | 0.546 | 0.174 | 0.671 | 0.111 |
| Decision Tree | 0.633 | 0.628 | 0.584 | 0.166 | 0.613 | 0.111 |
| Random Forest | **0.673** | 0.650 | 0.591 | 0.179 | 0.661 | 0.111 |
| Linear SVC | 0.664 | **0.880** | 0.077 | **0.335** | 0.981 | 0.111 |
| Gradient Boosting | 0.646 | 0.611 | 0.590 | 0.161 | 0.614 | 0.111 |
| XG Boost | 0.650 | 0.612 | **0.612** | 0.165 | 0.612 | 0.111 |
| Cat Boost | 0.646 | 0.605 | 0.608 | 0.161 | 0.605 | 0.111 |

41

## Hyper-Parameter Tuning

| | AUC | RECALL | PRECISION |
|---|---|---|---|
| DECISION TREE | 0.62901127267603 | 0.6301531213191991 | 0.15853211009174312 |
| RANDOM FOREST | **0.6773528020383** | **0.6095406360424** | **0.178171802375624** |
| XGBOOST | 0.6402066158071 | 0.598939929328622 | 0.1611216730038023 |

42




43

```
rf_grid = {'n_estimators':range(200,1000,200), # number of trees
           'max_features':['auto','sqrt'], # maximum number of features to use at each split
           'max_depth':range(1,11,1), # maximum depth of the tree
           'min_samples_split':range(2,10,2), # minimum number of samples to split a node
           'criterion':['gini','entropy']} # criterion for evaluating a split

rf_random = RandomizedSearchCV(estimator = rf_clf, param_distributions = rf_grid,
                               n_iter = 20, cv = 2, scoring=recall_score,
                               verbose = 1, random_state = 111)

rf_random.fit(X_train, y_train)

rf_random.best_params_

rf_hp_preds = rf_random.best_estimator_.predict(X_valid)
rf_hp_preds_proba = rf_random.best_estimator_.predict_proba(X_valid)[:,1]
roc_auc_score(y_valid, rf_hp_preds_proba)
```

```
Fitting 2 folds for each of 20 candidates, totalling 40 fits
0.6773528020382722
```

```
recall_score(y_valid, rf_hp_preds)
```

```
0.6095406360424028
```

```
precision_score(y_valid, rf_hp_preds)
```

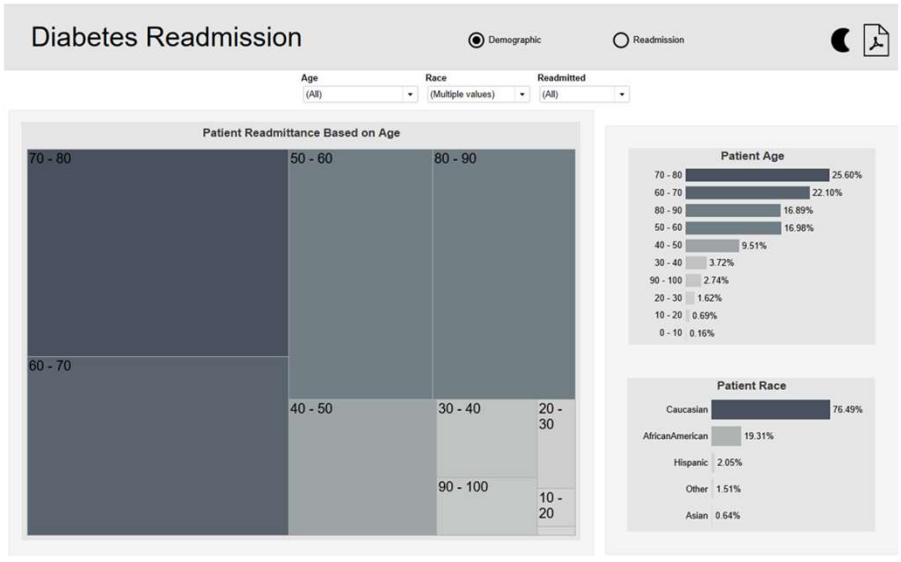
```
0.17817180237562402
```

44

# 05.
## Tableau Visualization

45

## Let's Begin the Demo




46

# THANKS!

Do you have any questions?



47