Early Prediction of Diabetes

Bushra Alzahrani

GENERAL ASSEMBLY

مسك misk

Data Science Immersive course
24 – Oct - 2018

Agenda

Introduction

- What is Diabetes?
- Why this is important?
- Project Objective
- Importance in Health Care sector.

Project Structure

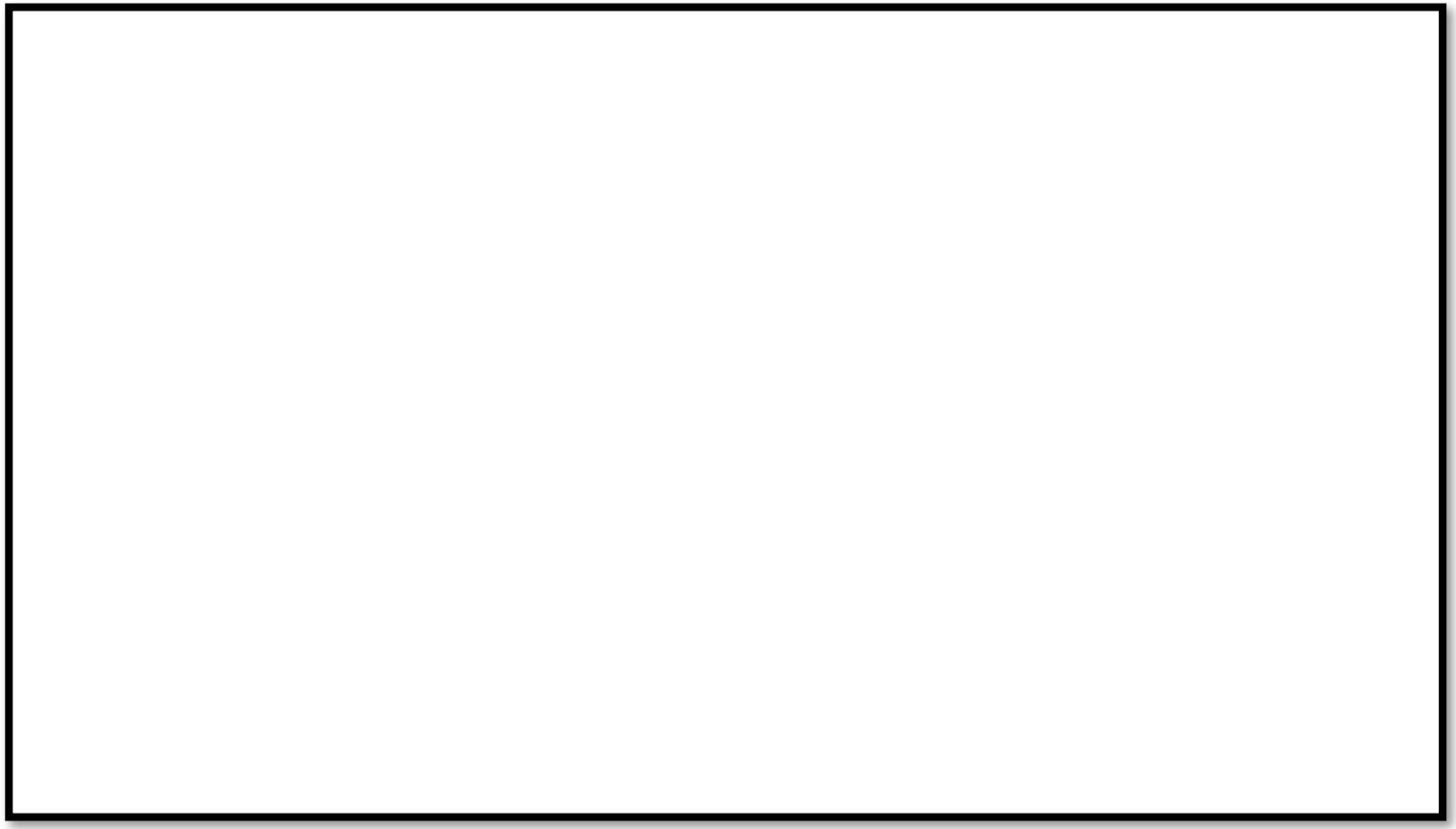
- Collecting data
- Cleaning data and EDA
- Modeling
- Interpretation

Conclusion

- The Usage of the Model
- Comparison between ML model and other tools used to detect diabetes.
- Limitations
- Results\Recommendations
- Future Improvements

Introduction

What is Diabetes?



Diabetes Stats

Based on World Health Organization (WHO) Diabetes affects **382 million** people in the world, and the number of people with type-2 diabetes is **increasing** in every country.

World Health Organization ranked Saudi Arabia as the **7th highest** country in the world with diabetes.

Institute for Health Metrics and Evaluation (IHME) estimated the **costs in 2014 were 17 billion SR**, expected to **increase to 27 billion SR** Incase **undiagnosed people are documented**, and **43 billion SR** if **pre-diabetics become diabetics**.

Prevalence of type-2 Diabetes in Saudi Arabia is **32.8%**, expected to **reach 45.36%** in the **year 2030**.

Prevention and early detection is the way to face these challenges

Project Objective

The objective of this project is to build a machine learning classification model to predict whether or not the patient has diabetes based on certain diagnostic measurements.

Importance in Health Care?

More understanding of diseases

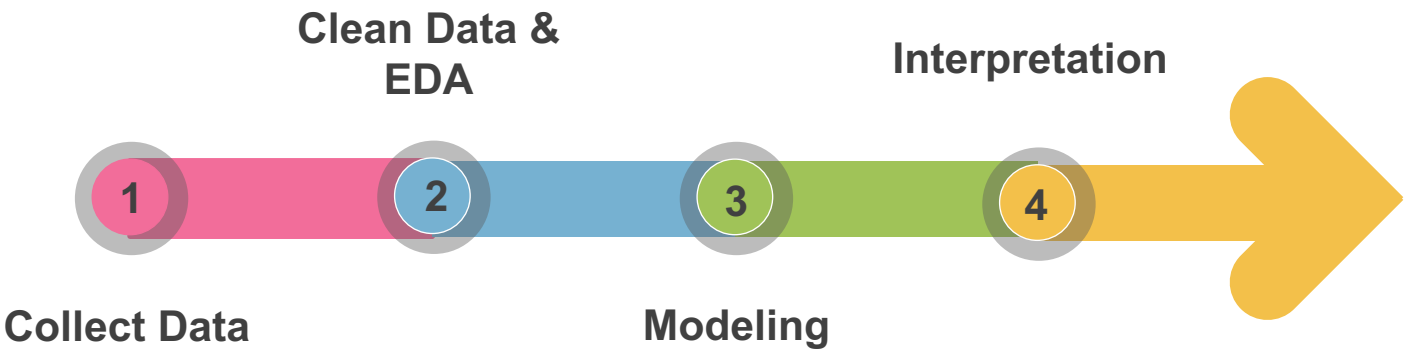
Early identifying patients

management the disease

Decreasing complications and costs

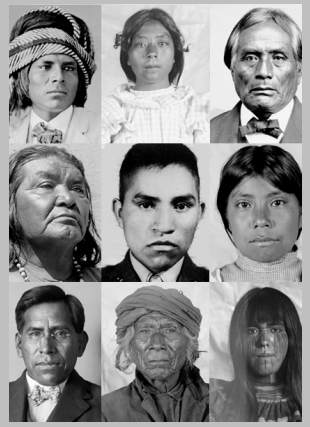
Project Structure

Project Structure



Stage 1

Collect Data



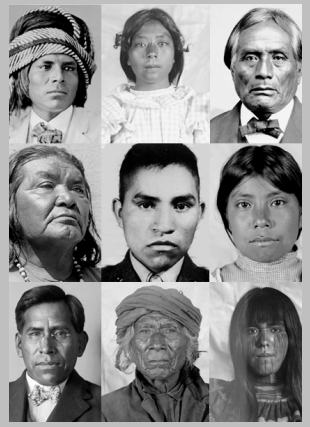
- The population for this study was the **Pima Indian population** near **Phoenix, Arizona**.
- The population has been under continuous study **since 1965** by the National Institute of Diabetes and Digestive and Kidney Diseases (**NIDDK**) because of its high incidence rate of diabetes.
- This dataset is originally from the **NIDDK**, and it is credited to **UCI Machine Learning Repository**.
- All patients in this dataset are **Females** of **at least 21 years** of age.

UCI Machine Learning Repository

NIH National Institute of Diabetes and Digestive and Kidney Diseases

Stage 1

Collect Data



Dataset contains 768 instances and the following 9 attributes:

- **(Pregnancies)** Number of times pregnant
- **(BloodPressure)** Diastolic blood pressure in mm Hg
- **(SkinThickness)** Triceps skin fold thickness in mm
- **(Insulin)** 2-Hour serum insulin in mu U/ml
- **(Glucose)** Plasma glucose concentration a 2 hours in an oral glucose tolerance test.
- **(BMI)** Body mass index measured as (weight in kg/(height in m)^2)
- **(DiabetesPedigreeFunction)** Diabetes pedigree function
- **(Age)** in years
- **(Outcome)** Class Variable (0 or 1) - **Target**

Stage 1

Collect Data

Challenge

Stage 2

Clean Data

EDA

- Data types
- Studying the target variable.
- Examine the distribution of the data.
- Missing data.
- Identify outliers.
- Find Correlation between features.
- Get the importance of features.
- Standardization.

Stage 2

Clean Data

EDA

- Data types

```
1  diabetes.info()
```

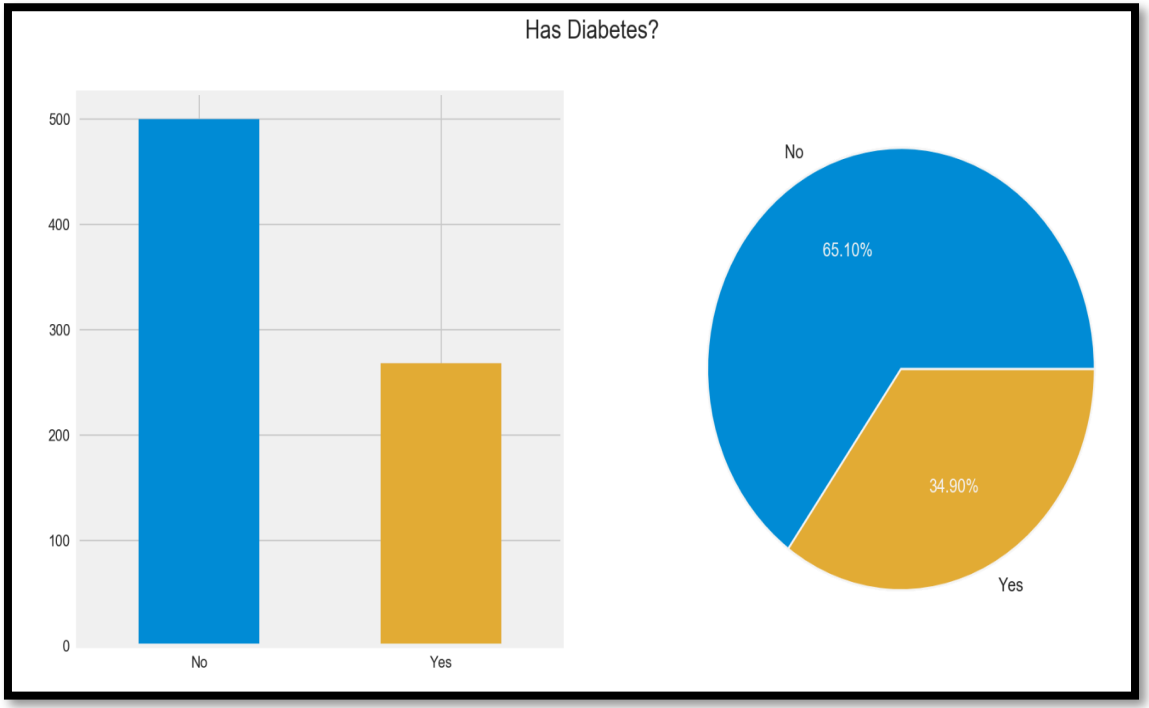
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
Pregnancies                 768 non-null int64
Glucose                     768 non-null int64
BloodPressure               768 non-null int64
SkinThickness               768 non-null int64
Insulin                     768 non-null int64
BMI                         768 non-null float64
DiabetesPedigreeFunction    768 non-null float64
Age                         768 non-null int64
Outcome                     768 non-null int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

Stage 2

Clean Data

EDA

- Studying the target variable.



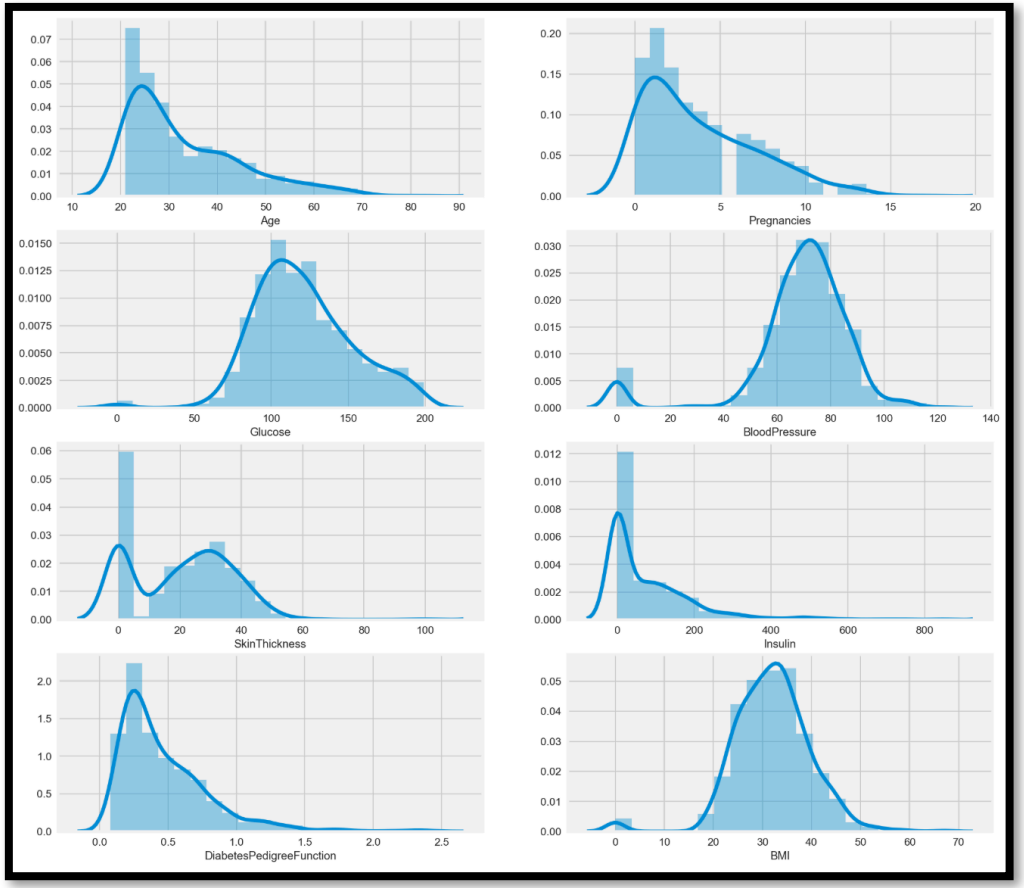
using parameter (class_weight="balanced")

Stage 2

Clean Data

EDA

- Examine the distribution of the data.



Stage 2

Clean Data

EDA

- Missing data

```
1 diabetes.describe().T
```

	count	mean	std	min	25%	50%	75%	max
Pregnancies	768.0	3.845052	3.369578	0.000	1.00000	3.0000	6.00000	17.00
Glucose	768.0	120.894531	31.972618	0.000	99.00000	117.0000	140.25000	199.00
BloodPressure	768.0	69.105469	19.355807	0.000	62.00000	72.0000	80.00000	122.00
SkinThickness	768.0	20.536458	15.952218	0.000	0.00000	23.0000	32.00000	99.00
Insulin	768.0	79.799479	115.244002	0.000	0.00000	30.5000	127.25000	846.00
BMI	768.0	31.992578	7.884160	0.000	27.30000	32.0000	36.60000	67.10
DiabetesPedigreeFunction	768.0	0.471876	0.331329	0.078	0.24375	0.3725	0.62625	2.42
Age	768.0	33.240885	11.760232	21.000	24.00000	29.0000	41.00000	81.00
has_diabetes	768.0	0.348958	0.476951	0.000	0.00000	0.0000	1.00000	1.00

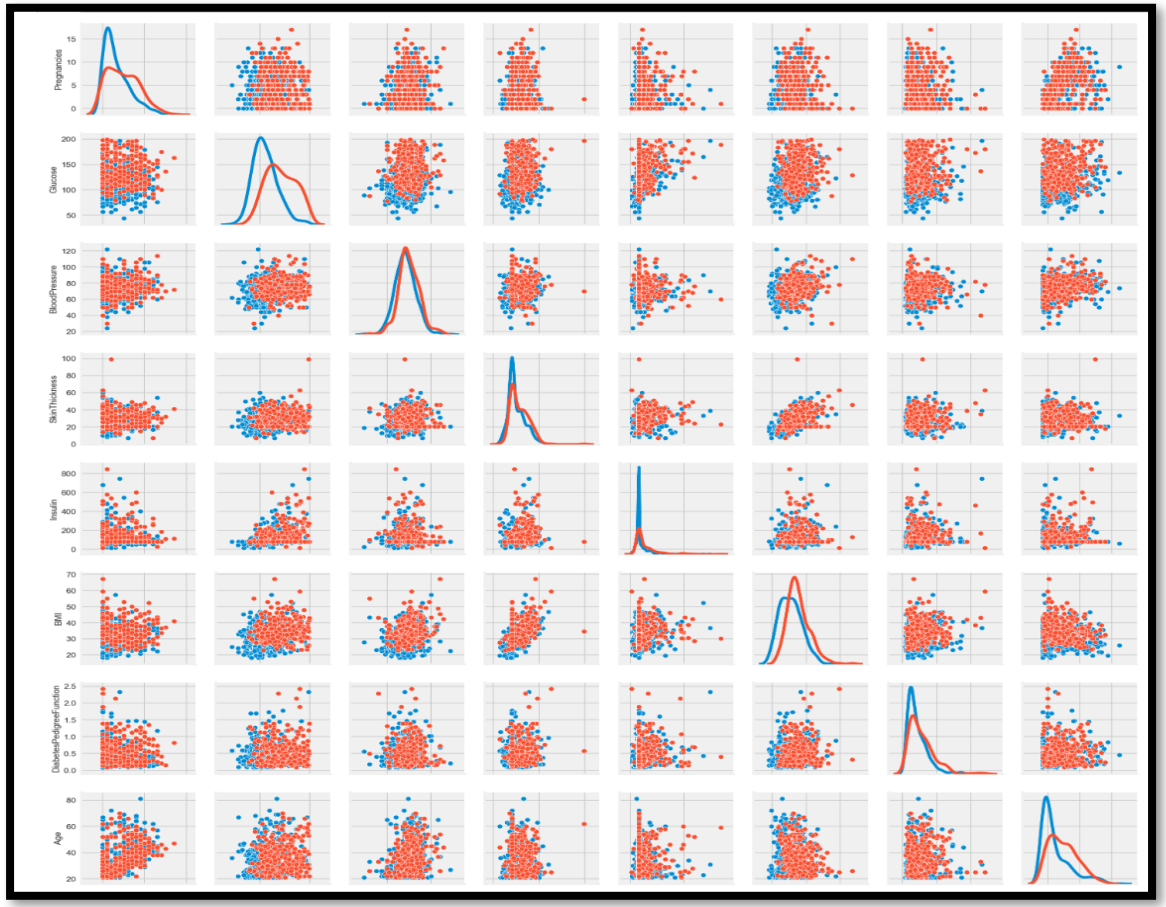
Replace zeroes with **mean**

Stage 2

Clean Data

EDA

- Identify outliers

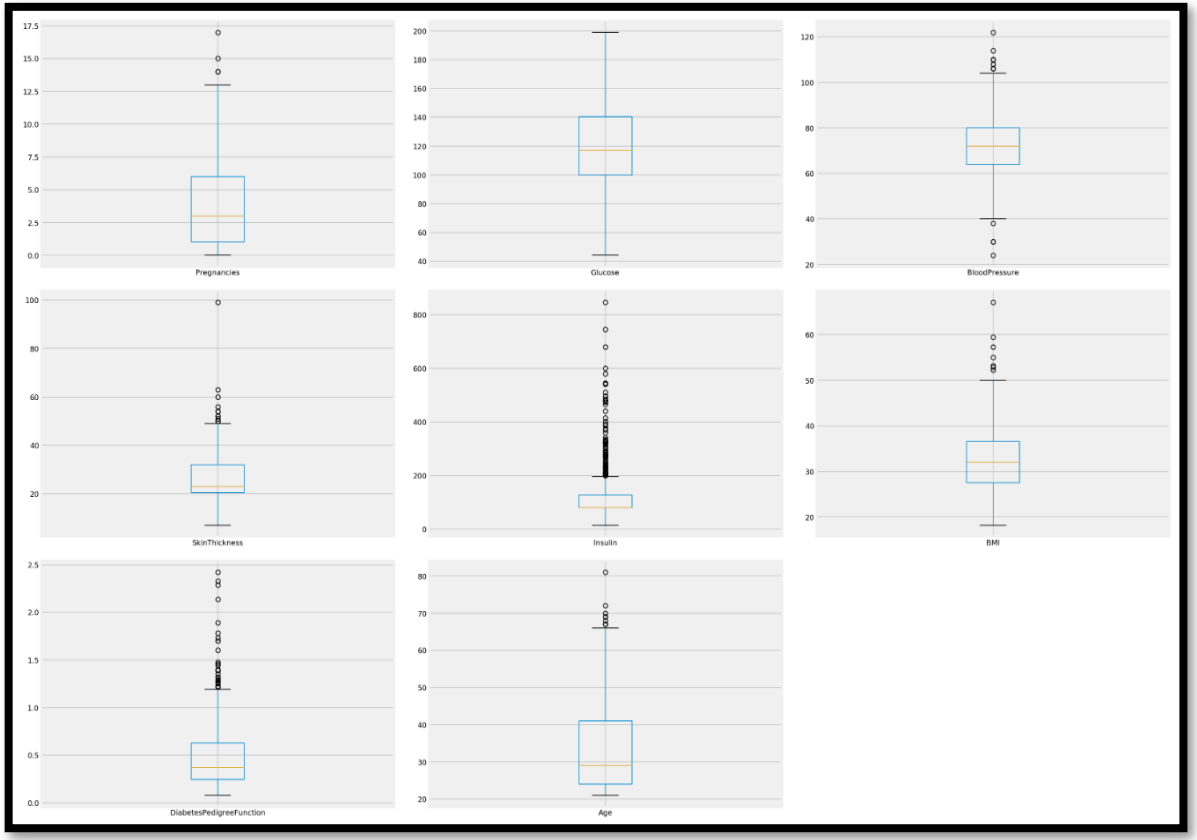


Stage 2

Clean Data

EDA

- Identify outliers



Two Datasets

Stage 2

Clean Data

EDA

- Find Correlation between features

Stage 2

Clean Data

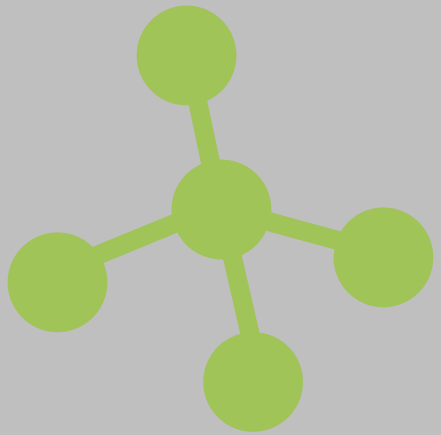
EDA

- Get the importance of features

Glucose	0.298491
BMI	0.153658
DiabetesPedigreeFunction	0.127618
Age	0.123754
Pregnancies	0.090990
BloodPressure	0.079216
Insulin	0.064997
SkinThickness	0.061276

Stage 3

Modeling

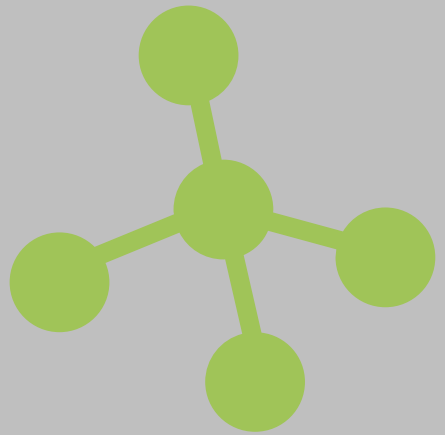


Consedirations before Modeling

- Outliers -> Use two datasets
- No enough data to create a validation set → k-fold cross validation to pick the best model.
- Imbalanced dataset -> set the input parameter **class_weight="balanced"**
- Features selection.

Stage 3

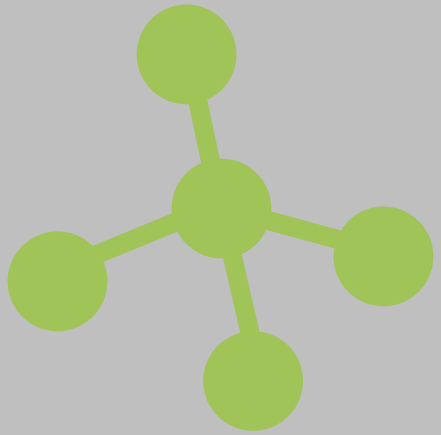
Modeling



- **Logistic Regression**
- **Random Forest Classifier**
- **Decision Tree**
- **Support Vector Machine**

Stage 3

Modeling

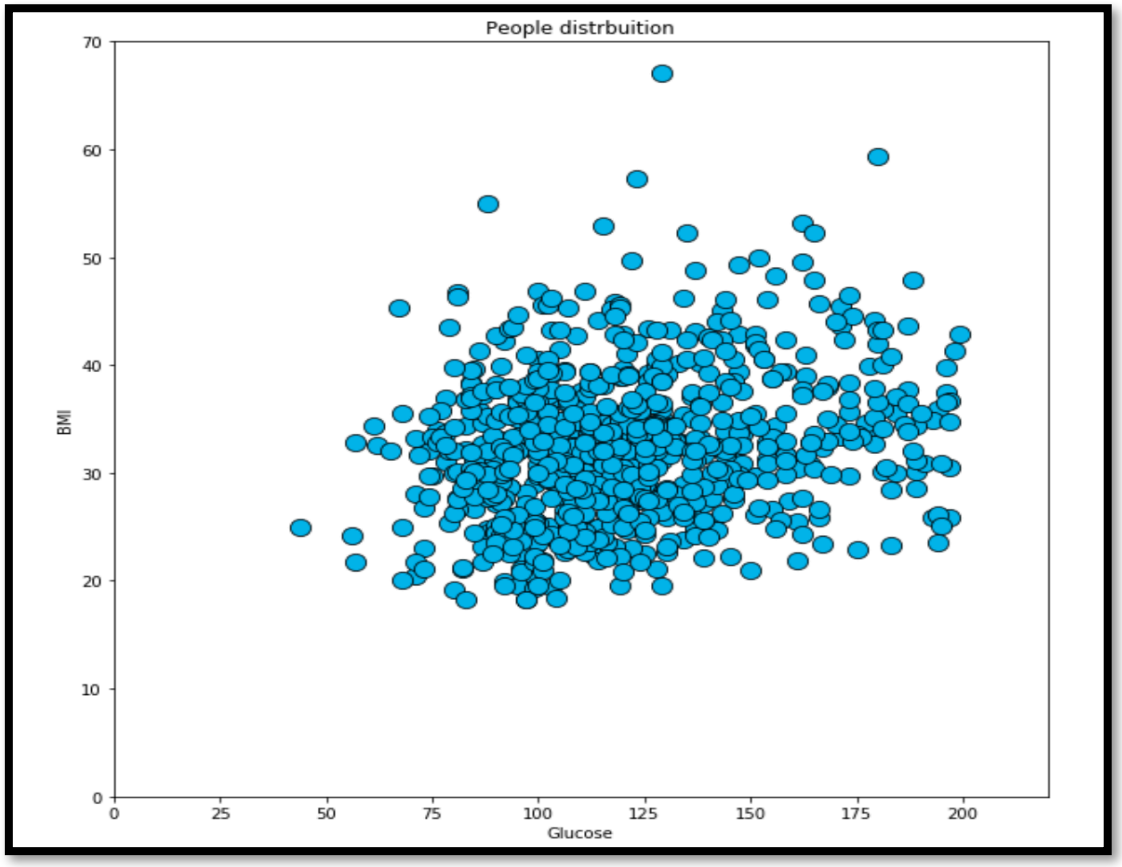


First Dataset (with Outliers)			Second Dataset (without Outliers)	
Model Name	Features	Accuracy	Features	Accuracy
Logistic Regression	All	77.0%	All	79.0%
	Selected	77.0%	Selected	78.0%
Random Forest	All	75.0%	All	75.0%
	Selected	74.0%	Selected	75.0%
Decision Tree	All	71.0%	All	71.0%
	Selected	70.0%	Selected	72.0%
SVM	All	77.0%	All	79.0%
	Selected	77.0%	Selected	78.0%

Stage 3

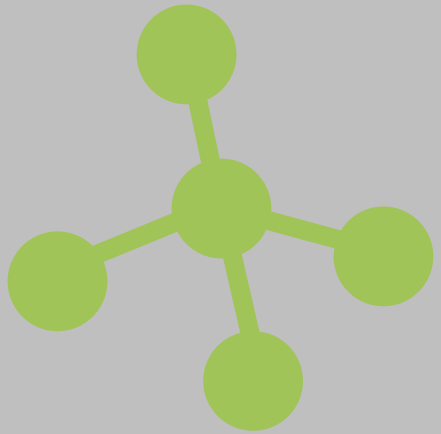
Modeling

- **DBSCAN (Glucose + BMI)**

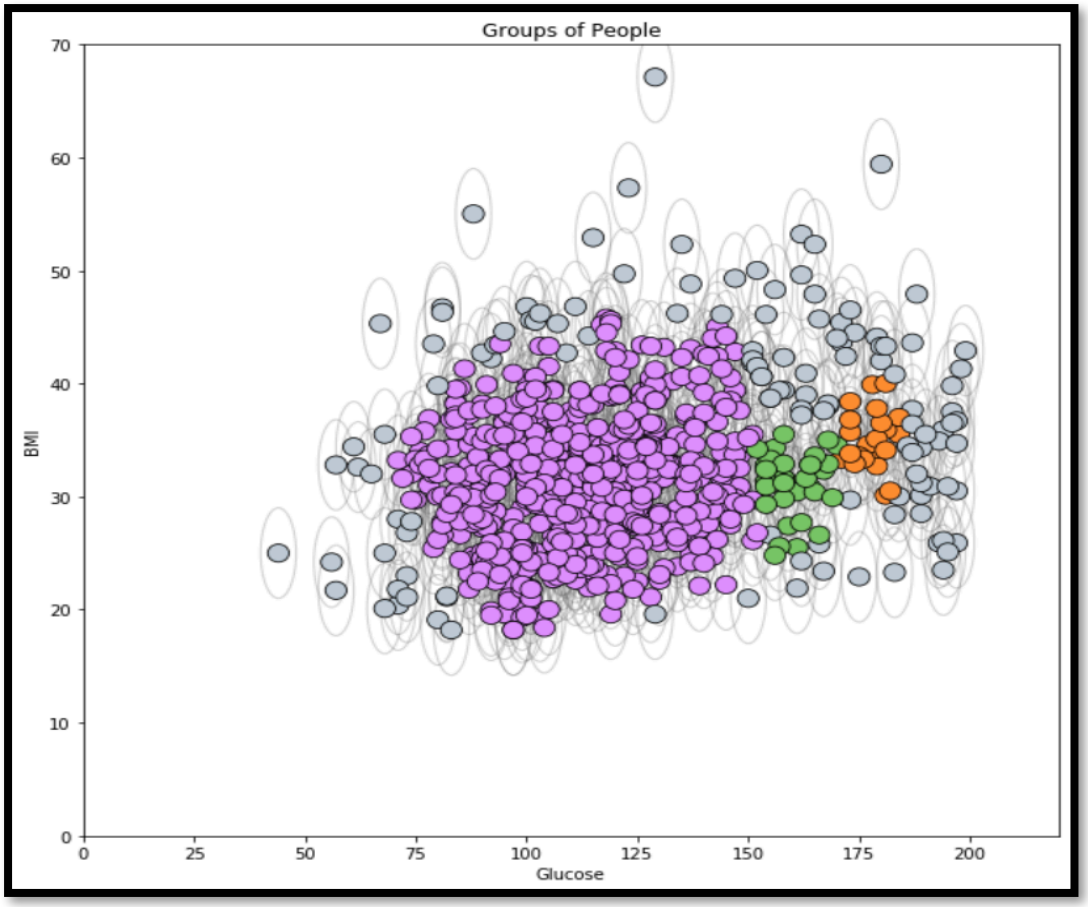


Stage 3

Modeling

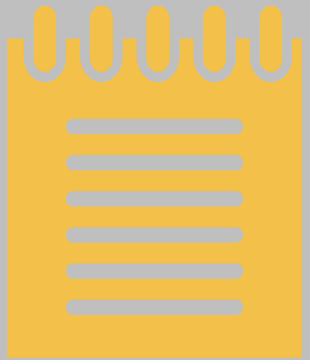


- **Three groups** (eps=4, min_samples=10)



Stage 4

Interpretation



Group 1

• **75% of people are clustered as Group 1**

•**Glucose**: **71 to 152** mg/dl

•**BMI**: **18 to 45**

•**26%** probability of getting diabetic for each person in this group

Group 2

• **4% of people are clustered as Group 2**

•**Glucose**: **152 to 170** mg/dl

•**BMI**: **24 to 35.5**

• **66%** probability of getting diabetic for each person in this group

Group 3

• **3% of people are clustered as Group 3**

•**Glucose**: **171 to 184** mg/dl

•**BMI**: **30 to 40**

• **88%** probability of getting diabetic for each person in this group

Conclusion

The Usage of the Model

1. **Timely** prediction from new information
2. **Decrease** in missed and undiagnosed patients
3. **Early** identification and management
4. **Decrease** in disease complications and costs

Achieving GOALS of VISION 2030

The Usage of the Model

Comparison between ML prediction Model and other tools

A study in Brazil for Type 2 diabetes, **22 million capillary glucose tests** were performed in individuals aged **40 years and older**, concluded that the screening program will yield a large health benefit but higher costs compared to no screening.

	Capillary glucose tests	Risk calculators	ML Model
Man power	13,000 health care provider	Physician or nurse or educated personnel to fill the form	Automated, self learning, timely
Time	7 March – 4 April 2001	10 min per patient	Less than 1 min to run the mod el
Number of screened patients	1 month =22.1 million 1 day =736,666 1 hour = 30,694 1 min = 511 patients	48 patient/ 8 hours	Less than 1 min = 768+ patient s
Costs	$76 per case diagnosed (345,000 newly diagnosed = $26,220,000)	Salary, papers or internet access, electricity, patient and physicians time, patient emotions	Computer, electricity bill

Limitations

- Lack of open data in Saudi Arabia.
- Missing data.
- More attributes:
  - urine test
  - hemoglobin A1c test
- Small Dataset.

Results & Recommendations

- Results were acceptable but could be improved (accuracy 70 - 79%).
- Results may be limited to Pima Indians, but it gives us a good start on how to begin diagnosing other populations with diabetes.
- The model could be used in:
    1. Prevention and Health monitoring programs.
    2. Timely mass screening tool with low cost compared to traditional methods.
- Encourage governmental sectors and hospitals to share their data.

Future Improvements

- Use more features.
- Use other imputing methods for missing data.
- Try different features for clustering.
- Try another ways to solve Imbalanced problem:

Synthetic Minority Oversampling Technique

Thank you

References

- Dataset:
  - UCI machine Learning Repository
    https://archive.ics.uci.edu/ml/datasets.html?format=&task=cla&att=num&area=life&numAtt=&numIns=&type=mvar&sort=typeUp&view=list
- Introduction video:
  - https://www.youtube.com/watch?v=wZAjVQWbMIE
- Diabetes Stats:
  - **World Health Organization**: http://www.who.int/
  - **Institute for Health Metrics and Evaluation:** http://www.healthdata.org/saudi-arabia
  - https://machinelearningmastery.com/case-study-predicting-the-onset-of-diabetes-within-five-years-part-1-of-3/
- Brazil mass screening Research:
  - https://www.ncbi.nlm.nih.gov/pubmed/26523154