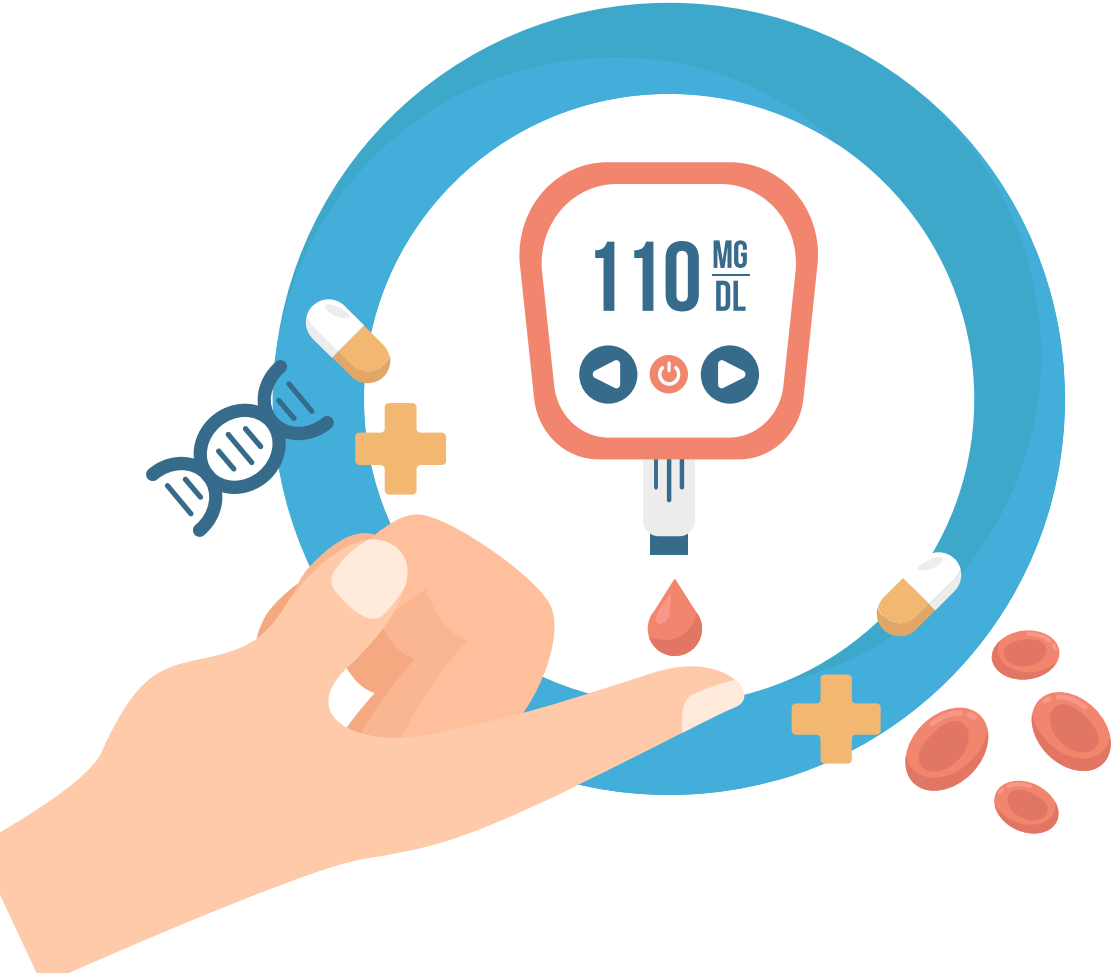


Diabetes Health Indicators

Content

1 Introduction

2 Tools

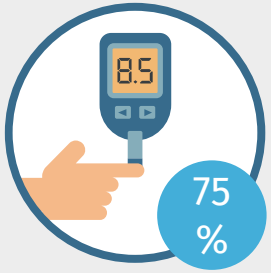
3 EDA

Exploratory data analysis

4 Modeling

5 Conclusion

Introducation



Diabetes

one of the most prevalent chronic diseases in the world, afflicting millions of People each year and imposing a significant financial burden on the economy..

Infection Rate

34.2M

AREAS AFFECTED

- Kidney
- Heart
- Sight
- Ears
- Feet
- Brain

Diabetes Data

Dataset

The dataset is contains **254000 observation** , each has **22 features.**

Columns and Target

High blood

High Cholesterol

Heart Attack

General Health

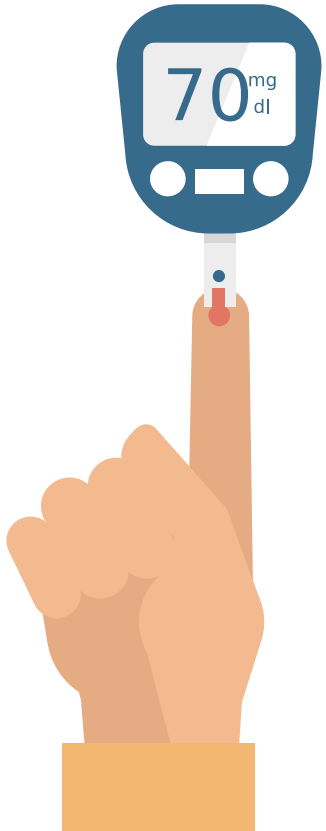
Body mass index (BMI)

Difficulty Walking

Age

Sex

Target : Diabetes Type

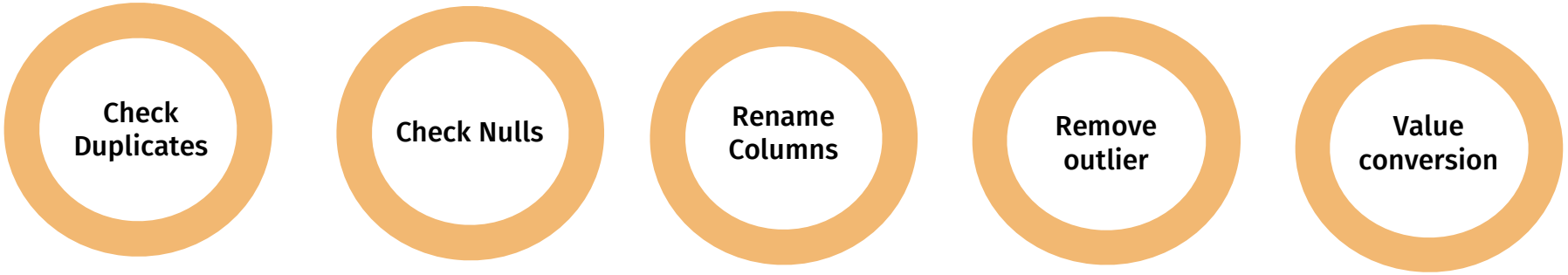


Goal

Classify the people that will infection Diabetes Type is Diabetes or No Diabetes .

Tools

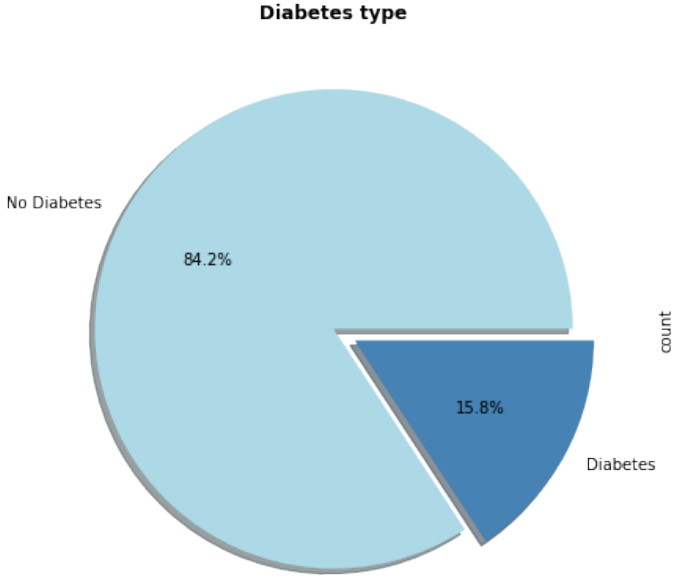
Cleaning Process

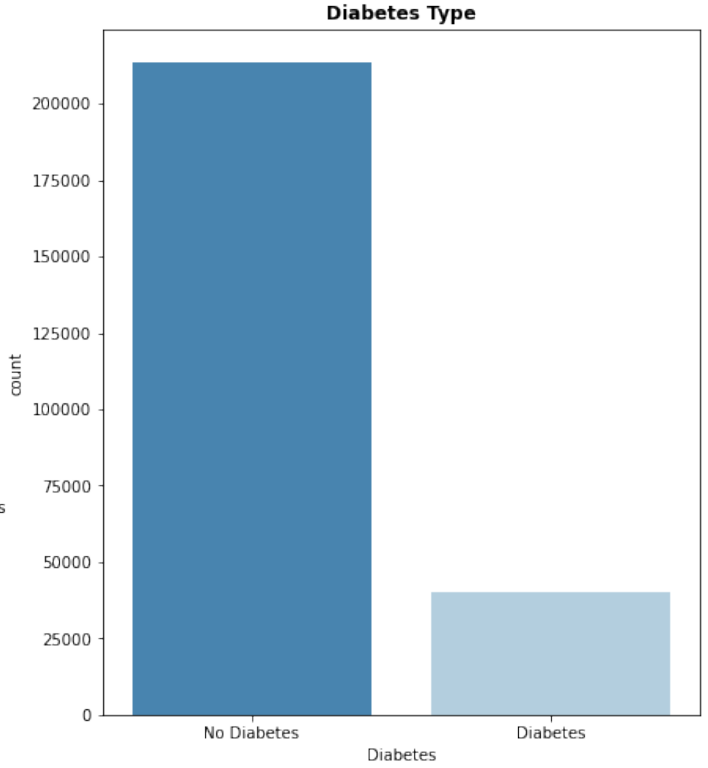


Note : After cleaning , the data is clean and there are no missing values

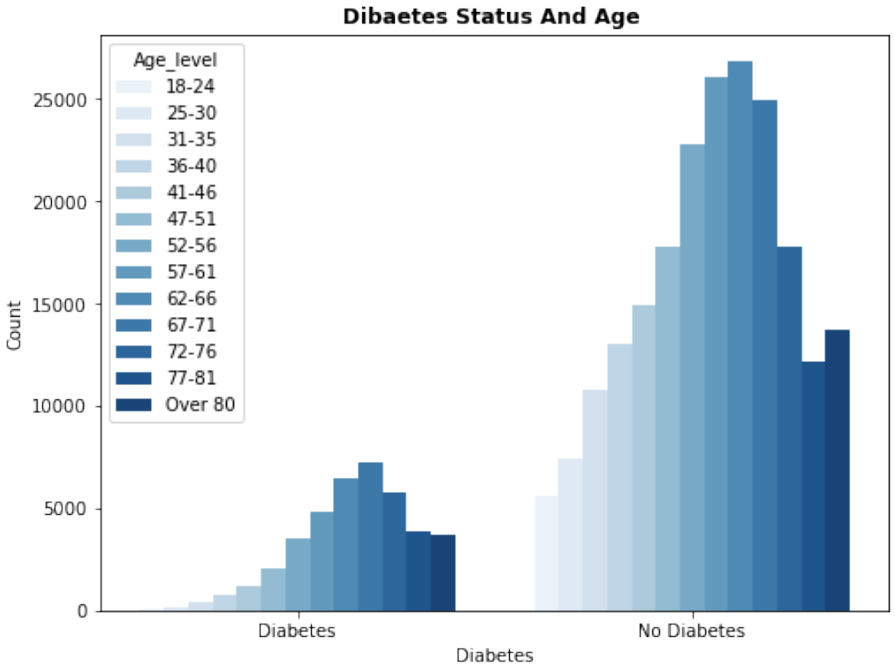
Visualization

Diabetes Type

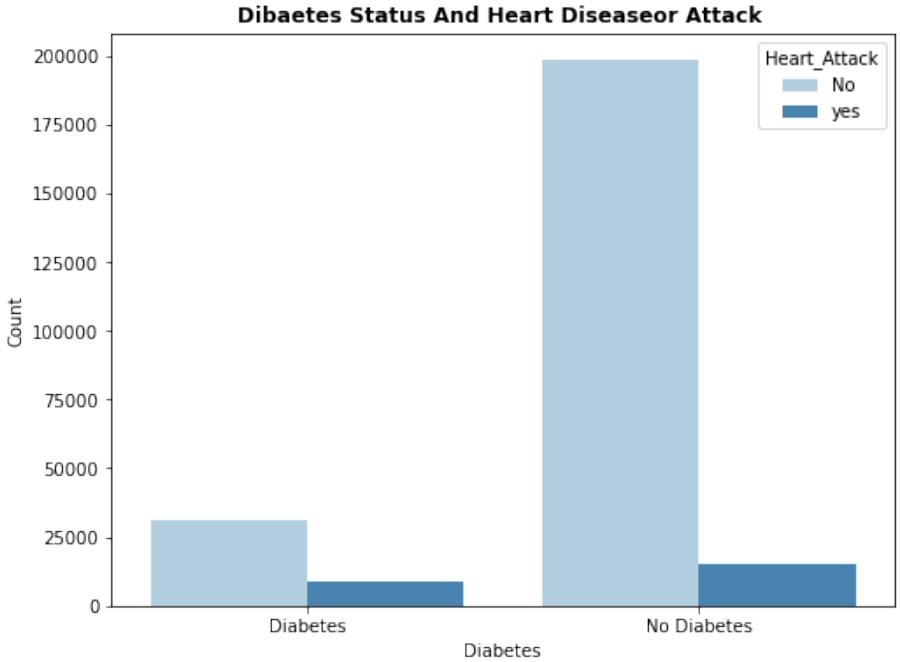


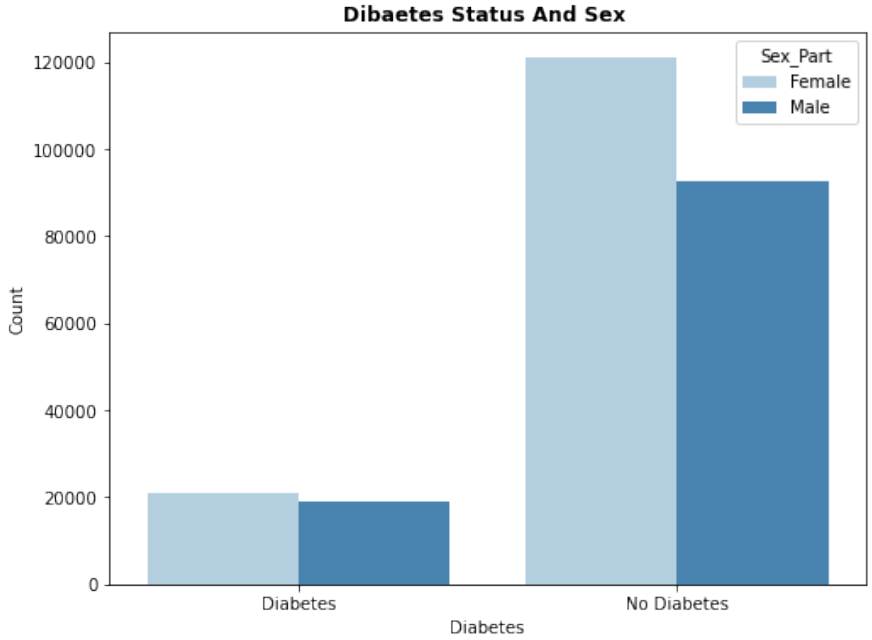


Distribution Age

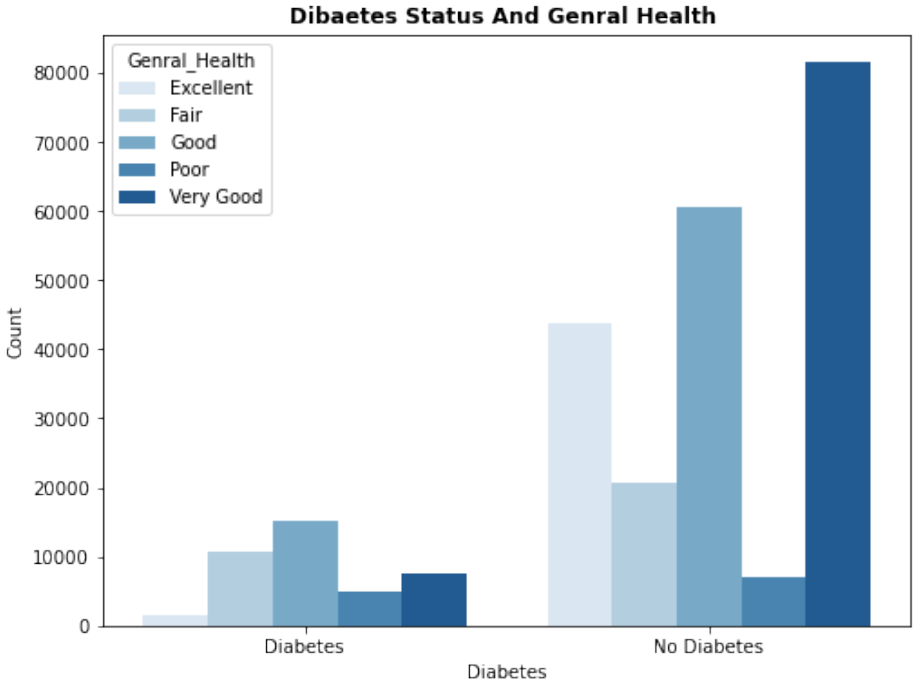


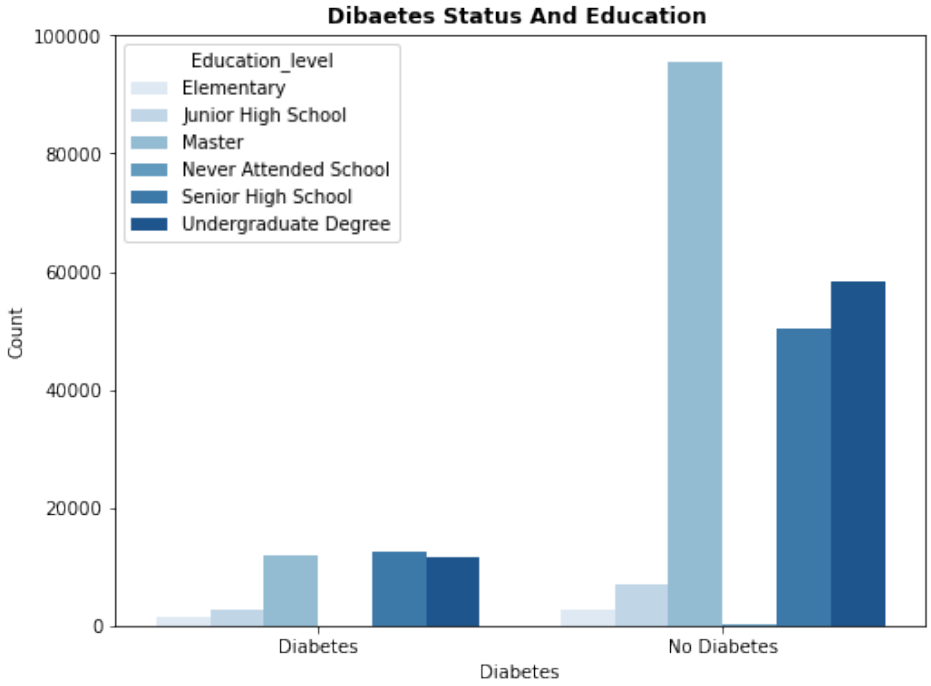
Diseaseor Attack & Sex





General Health & Education





Modeling

- K Neighbors Classifier
- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- Ensembling with Voting
- Hard , Soft , Average

Class imbalance

Before model training:

Resampling strategies
Oversampling: Random over Sampler , SMOTE
Undersampling: Random Under Sampler .

During model training:
Training with adjusted class weights After model

Class imbalance

Logistic Regression

Oversampling:

Random over Sampler :

Training Score after balance the labels 0.7392441690536038

Validation Score after balance the labels 0.7257138633619946

SMOTE :

Training Score after balance the labels (Smote): 0.7411220552990004

Validation Score after balance the labels (Smote): 0.7249254724186356

Class Weights

Training Score after Balanced class weights Logistic Regression 0.7280712019956269

Validation Score after Balanced class weights Logistic Regression: 0.7256399517110547

Undersampling:

Random Under Sampler :

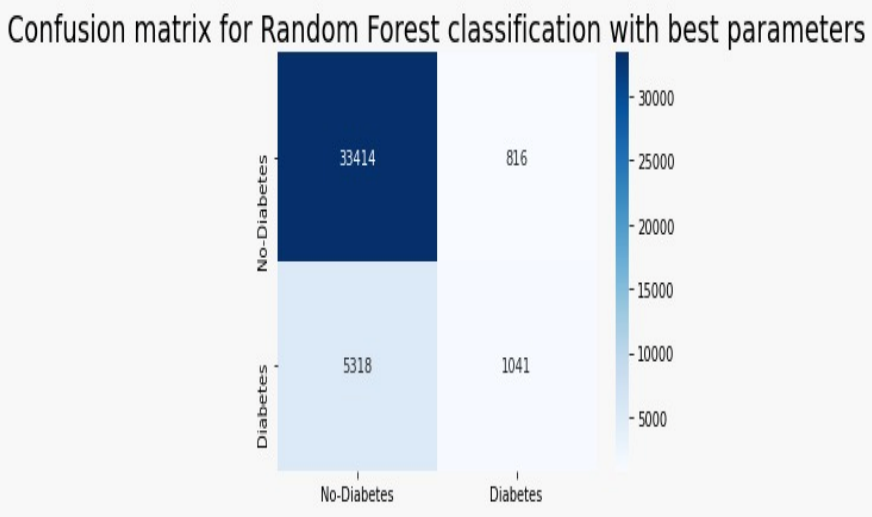
Training Score after balance the labels :0.7389505470802776

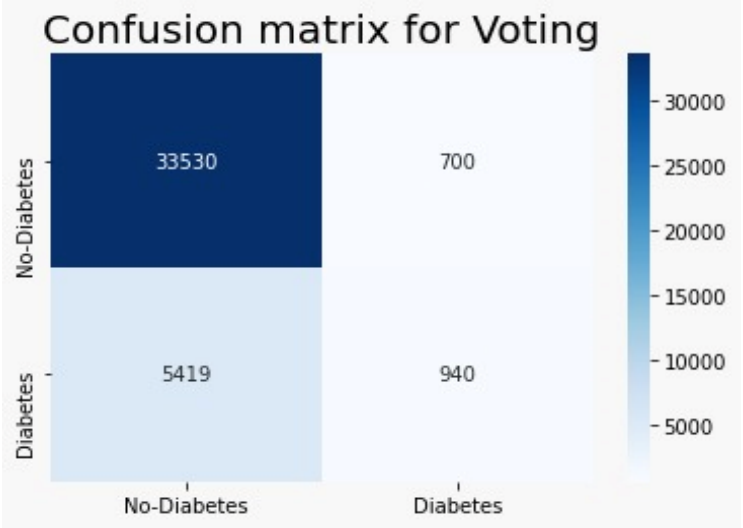
Validation Score after balance the labels: 0.7259848727487743

Experiments – Results

	Model	Accuracy	Precision	Recall	F1 score
0	KNN after balance our target's labels	0.742663	0.313909	0.551049	0.399971
1	LogisticRegression with Smote	0.724925	0.332380	0.749332	0.460498
2	Decision Tree Classification	0.847791	0.549535	0.157886	0.245297
3	RandomForestClassifier_best parameters	0.848875	0.560582	0.163705	0.253408
4	RandomForestClassifier_best parameters	0.848875	0.560582	0.163705	0.253408
5	VotingClassifier-Hard	0.849245	0.573171	0.147822	0.235029
6	VotingClassifier-Average Voting	0.847767	0.543520	0.176757	0.266762
7	VotingClassifier-Weighted Voting	0.846535	0.528114	0.192011	0.281628

Confusion matrix of Best Model





THANKS