Purwadhika
Digital Technology School

# DEVELOPING AN EARLY DIABETES RISK PREDICTOR APP USING MACHINE LEARNING

by Harish Muhammad

# Outlines

- Background/context
- Problem Understanding
- Data Understanding
- Data Analysis
- Modeling
- Hyperparameter tuning
- Confusion matrix
- Learning curve
- LIME
- Recommendations
- Demonstration

# What is diabetes?


WHO: **422 M** diabetic patients
**1.5 M deaths**

Blood glucose too high


Chronic disease

Insulin dysfunction

Complications

# Why an early diabetes predictor App?


Easy & practical


Cost effective

Early warning

# Data Understanding

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 520 entries, 0 to 519
Data columns (total 17 columns):
 #   Column              Non-Null Count  Dtype 
---  ------              --------------  ----- 
 0   Age                 520 non-null    int64 
 1   Gender              520 non-null    object
 2   Polyuria            520 non-null    object
 3   Polydipsia          520 non-null    object
 4   sudden weight loss  520 non-null    object
 5   weakness            520 non-null    object
 6   Polyphagia          520 non-null    object
 7   Genital thrush      520 non-null    object
 8   visual blurring     520 non-null    object
 9   Itching             520 non-null    object
 10  Irritability        520 non-null    object
 11  delayed healing     520 non-null    object
 12  partial paresis     520 non-null    object
 13  muscle stiffness    520 non-null    object
 14  Alopecia            520 non-null    object
 15  Obesity             520 non-null    object
 16  class               520 non-null    object
dtypes: int64(1), object(16)
memory usage: 69.2+ KB
```

1 numerical feature (data type: int64)
15 categorical features (data type: object)

## Dataset Source

**Early Stage Diabetes Risk Prediction**
Donated on 7/11/2020

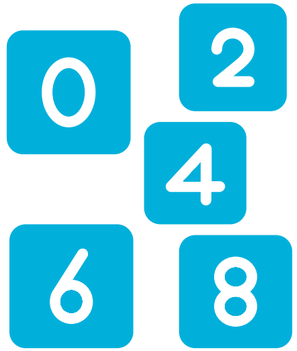
This dataset contains the sign and symptpom data of newly diabetic or would be diabetic patient.

| Dataset Characteristics | Subject Area | Associated Tasks |
|---|---|---|
| Multivariate | Computer Science | Classification |
| **Feature Type** | **# Instances** | **# Features** |
| Categorical, Integer | 520 | 16 |

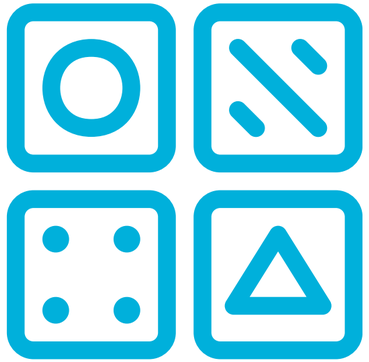
UC Irvine Machine Learning Repository

# Exploratory Data Analysis (EDA)

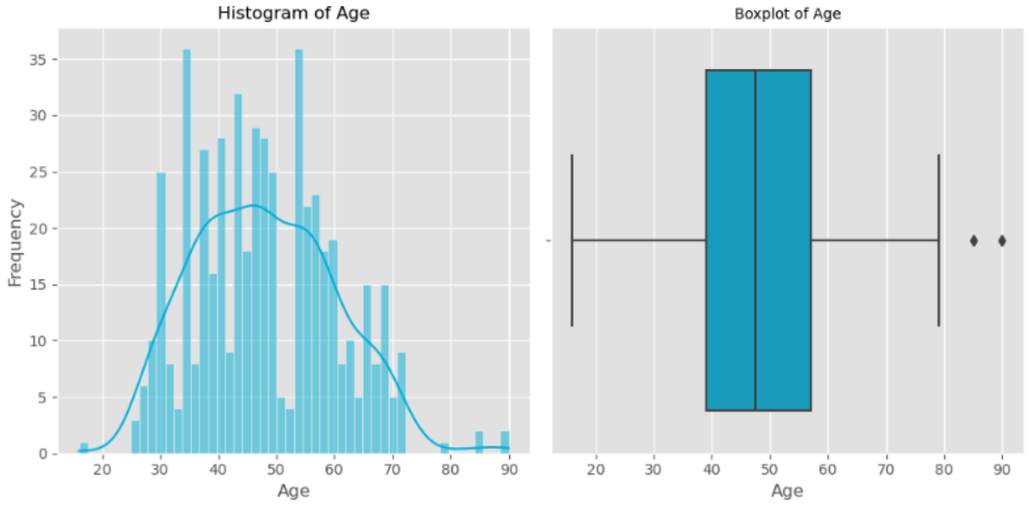

## Numerical vs Target




## Categorical vs Target

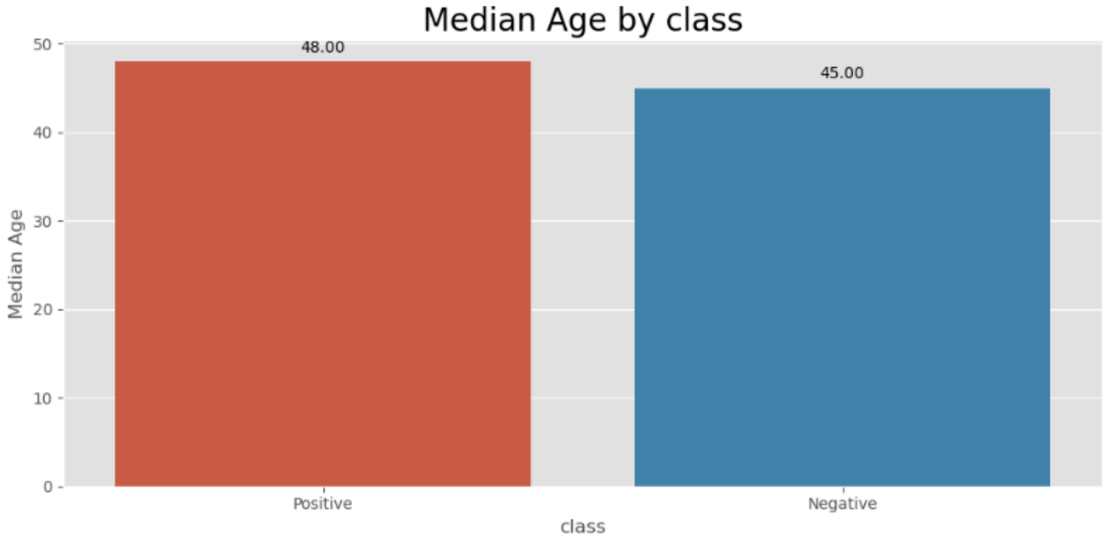


# Numerical vs Target

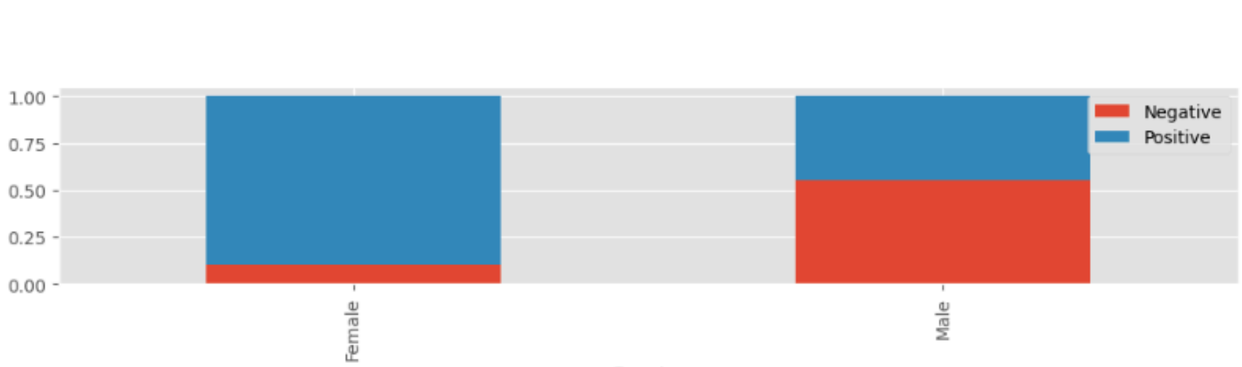


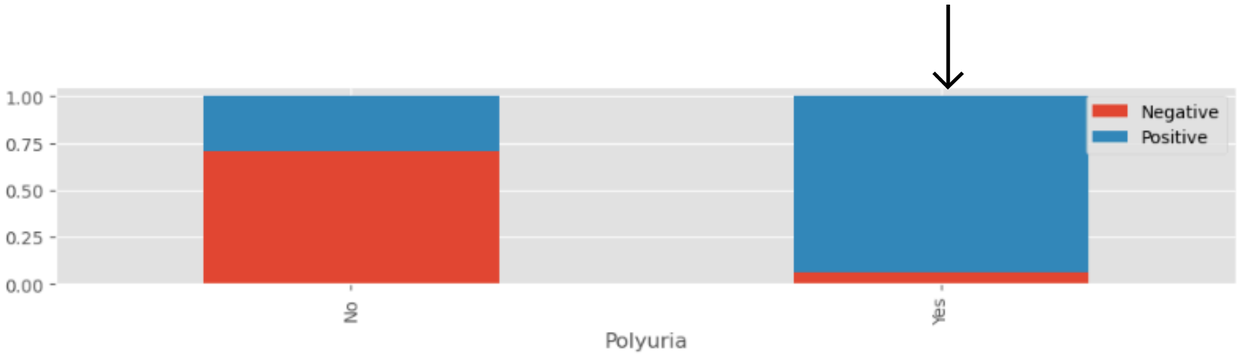


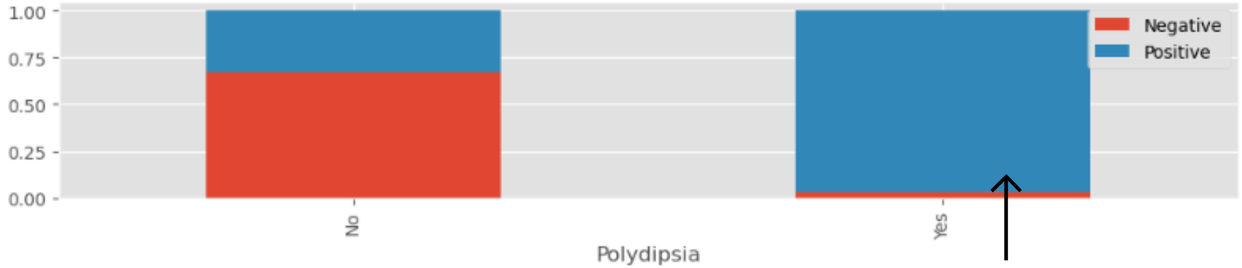


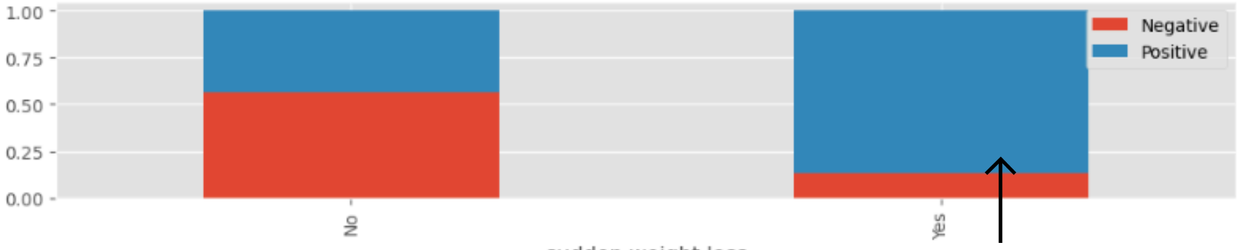


# Categorical vs Target (1)

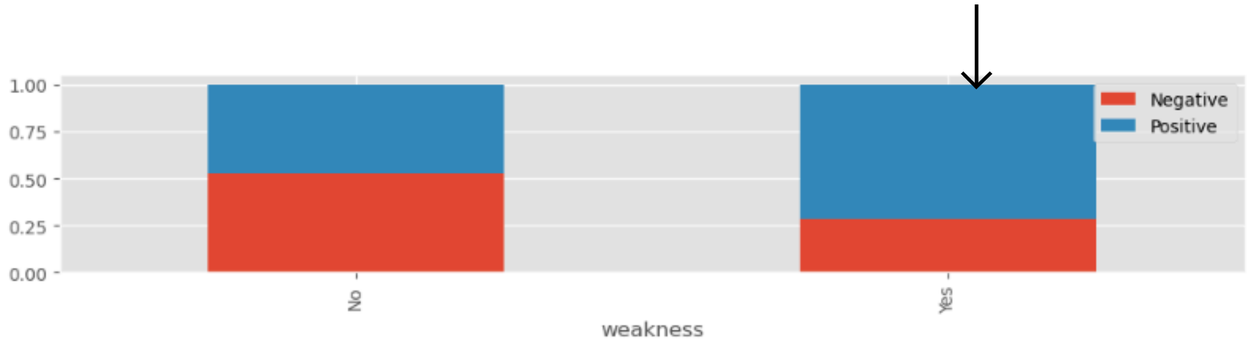


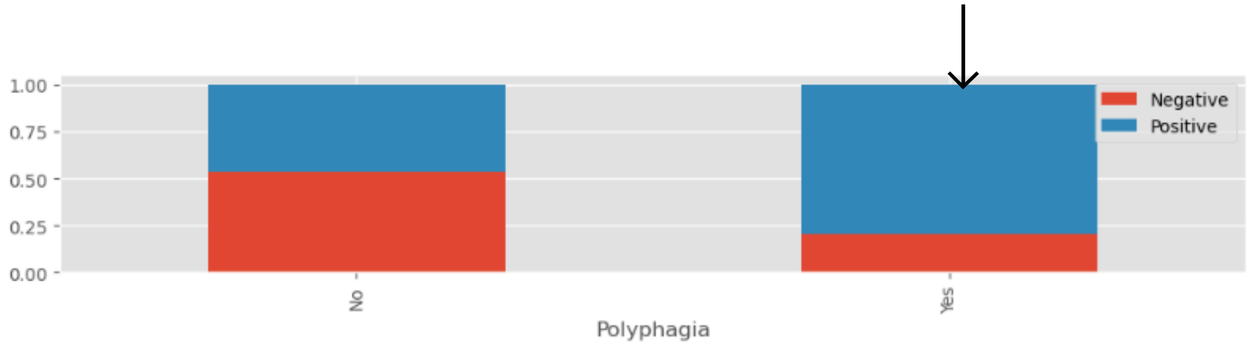



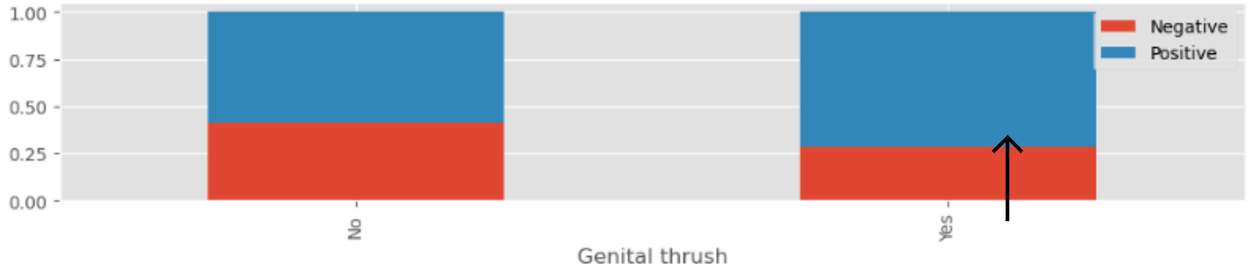



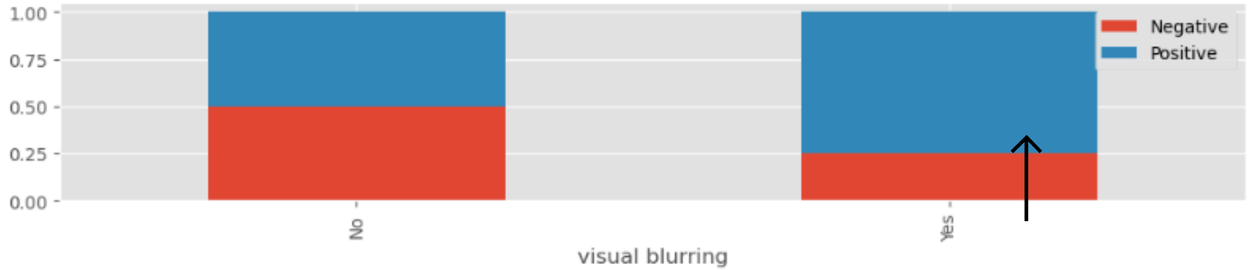



# Categorical vs Target (2)

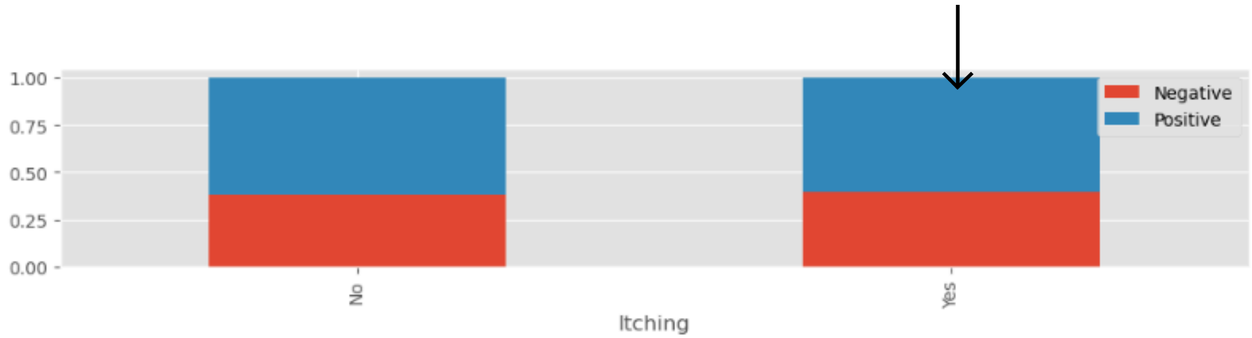


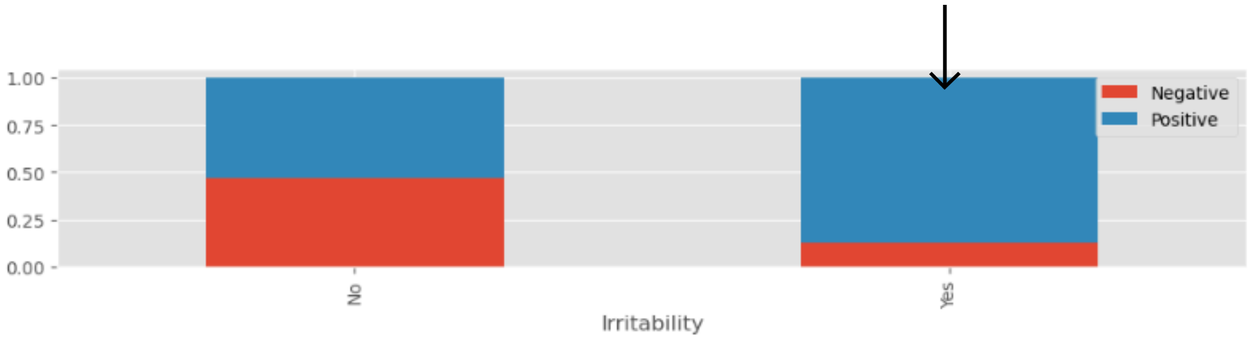



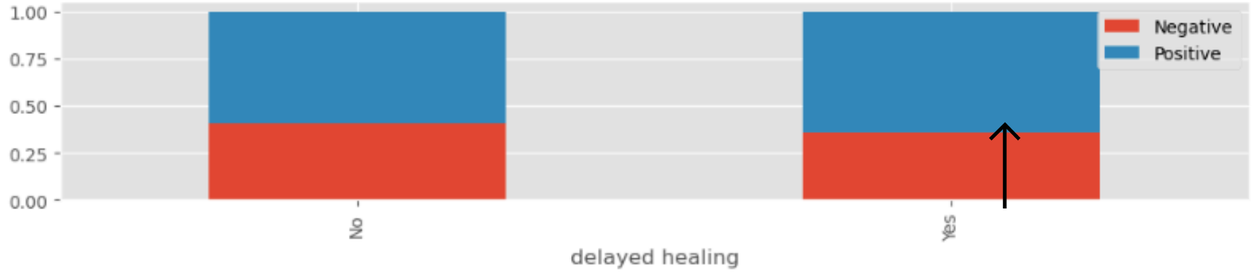



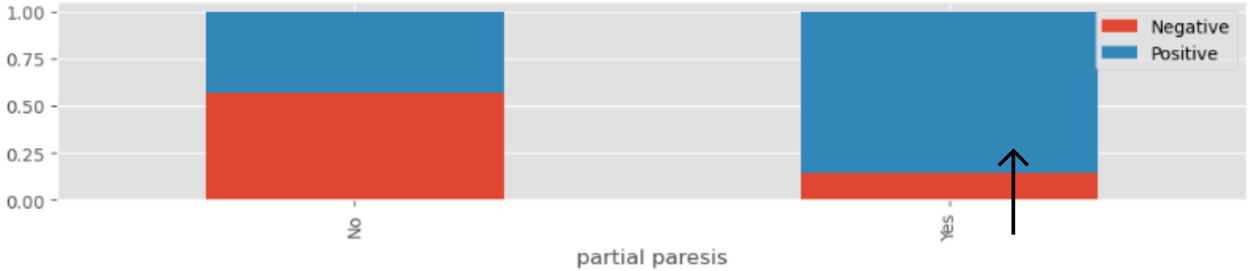



# Categorical vs Target (3)

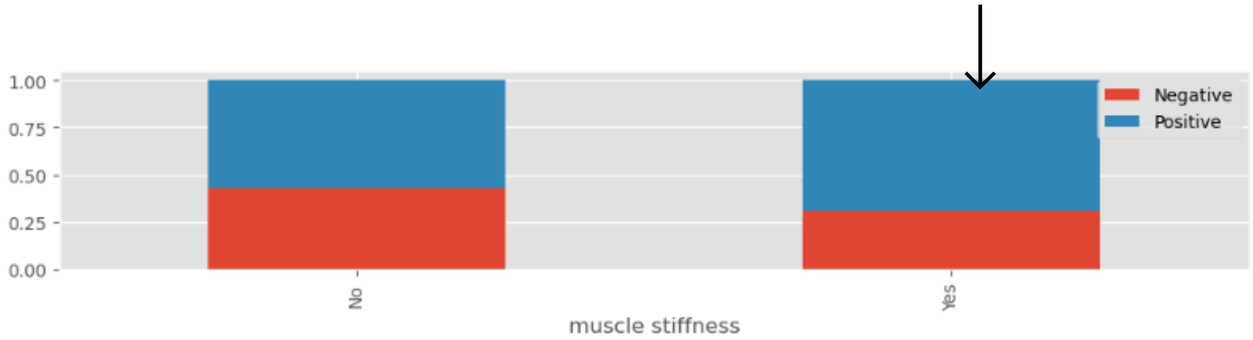


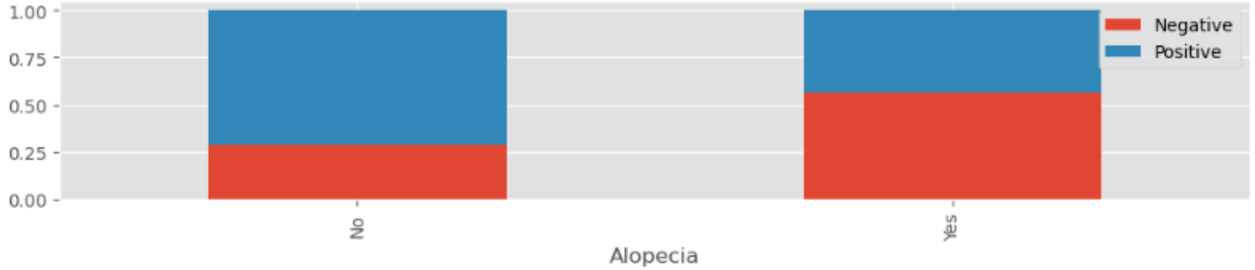



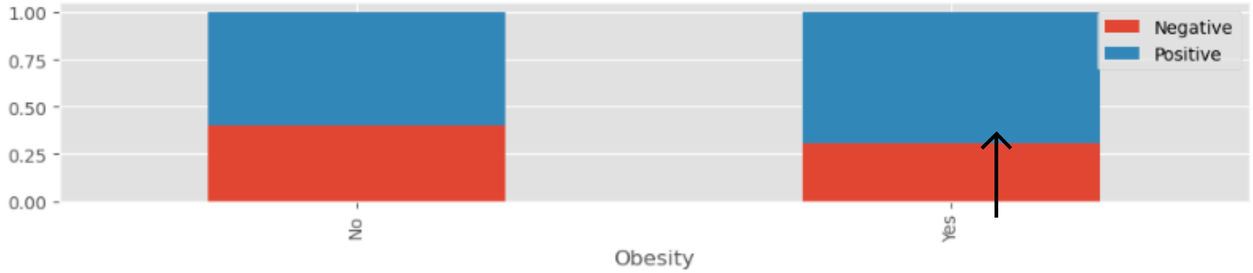



# Categorical vs Target (4)



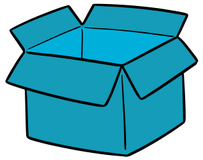



# Preprocessing

## Handling Outliers

### Winsorization



## Encoding

### OneHot Encoder


## Scaling

### MinMaxScaler

## Resampling

### SMOTE & SMOTE NC

# Benchmark Model

01 Logistic Regression

02 KNN Classifier

03 Decision Tree Classifier

04 Random Forest Classifier

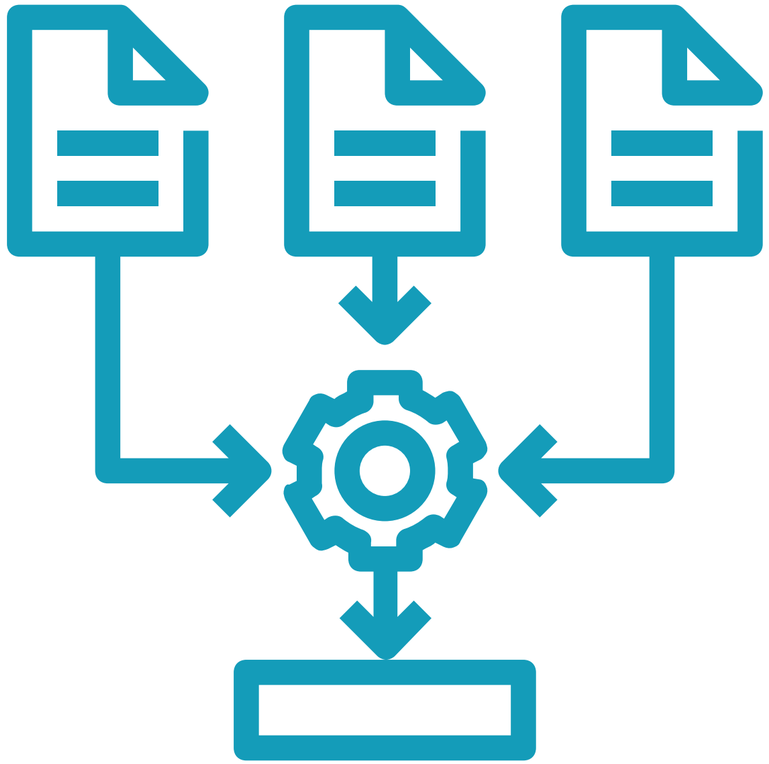
05 Adaptive Booster Classifier

06 Gradient Booster Classifier

07 Categorical Booster Classifier

08 XGBoost Classifier

09 LGBM Classifier



# SMOTE NC vs SMOTE

| model | recall_mean_with_smotenc | recall_mean_with_smote | recall_std_with_smotenc | recall_std_with_smote |
|---|---|---|---|---|
| Random Forest | 0.977576 | 0.968586 | 0.014376 | 0.023141 |
| LightGBM | 0.964242 | 0.950909 | 0.022745 | 0.021736 |
| CatBoost | 0.955354 | 0.955354 | 0.028112 | 0.028112 |
| GradienBoost | 0.955253 | 0.964242 | 0.020221 | 0.022745 |
| XGBoost | 0.946364 | 0.946364 | 0.022848 | 0.033386 |
| Decision Tree | 0.932929 | 0.941919 | 0.020339 | 0.017939 |
| Logistic Regression | 0.919596 | 0.919596 | 0.017997 | 0.022835 |
| AdaBoost | 0.915152 | 0.906263 | 0.016757 | 0.032615 |
| KNN | 0.910505 | 0.888182 | 0.032295 | 0.025641 |

# Model Performance

| Model | Mean Recall Train | Mean Recall Test |
|---|---|---|
| Random Forest | 0.977576 | 0.989583 |
| XGBoost | 0.946364 | 0.989583 |
| Gradient Boosting | 0.955253 | 0.968750 |
| CatBoost | 0.955354 | 0.968750 |
| LightGBM | 0.964242 | 0.968750 |
| Decision Tree | 0.932929 | 0.958333 |
| Logistic Regression | 0.919596 | 0.947917 |
| AdaBoost | 0.915152 | 0.937500 |
| KNN | 0.910505 | 0.906250 |


Random Forest


XGBoost

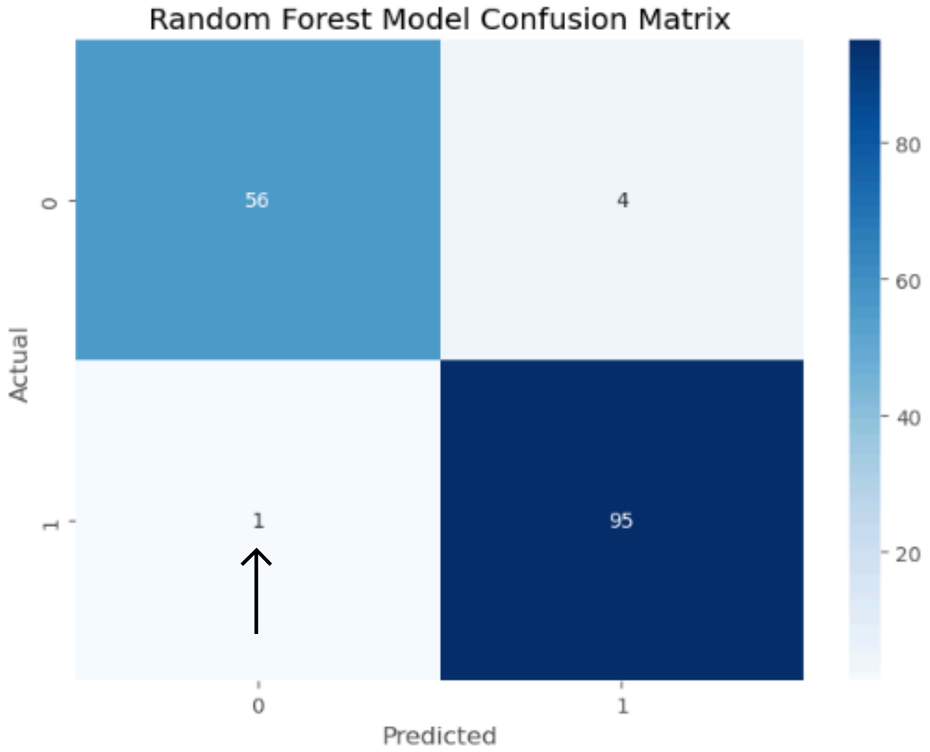
# Hyperparameter Tuning

## Model Performance Before & After Tuning

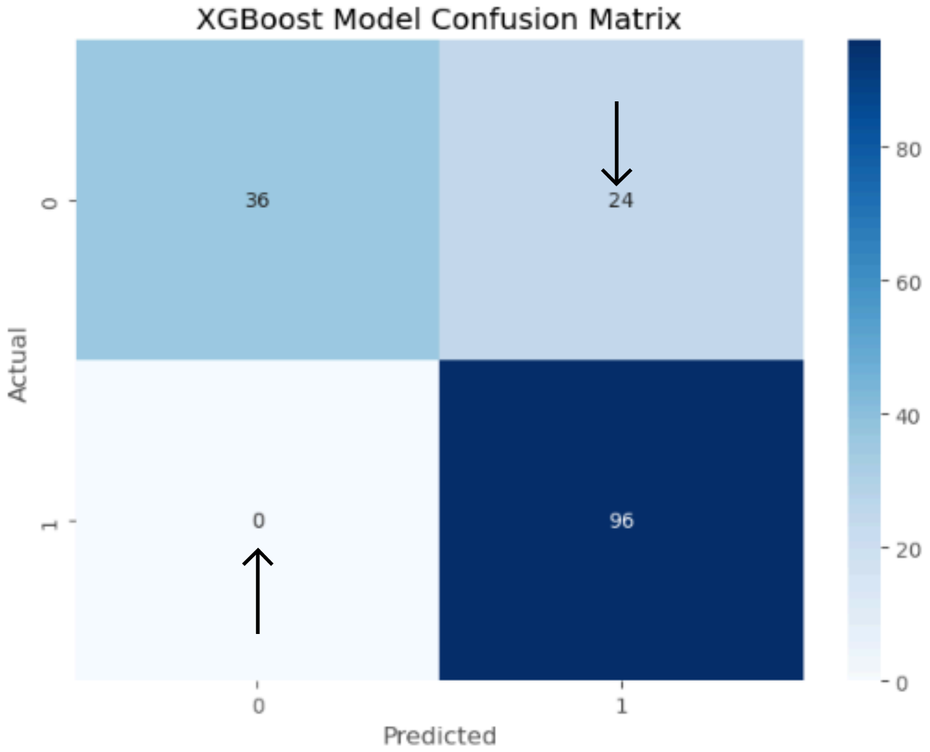
| Model | Conditions | Train score | Test score | |
|---|---|---|---|---|
| Random Forest | Before Tuning | 0.978 | 0.989 | |
| Random Forest | After Tuning | 0.978 | 0.989 | ← |
| XGBoost Classifier | Before Tuning | 0.946 | 0.989 | |
| XGBoost Classifier | After Tuning | 0.991 | 1.0 | ← |

# Confusion matrix

## Model Performance Before & After Tuning

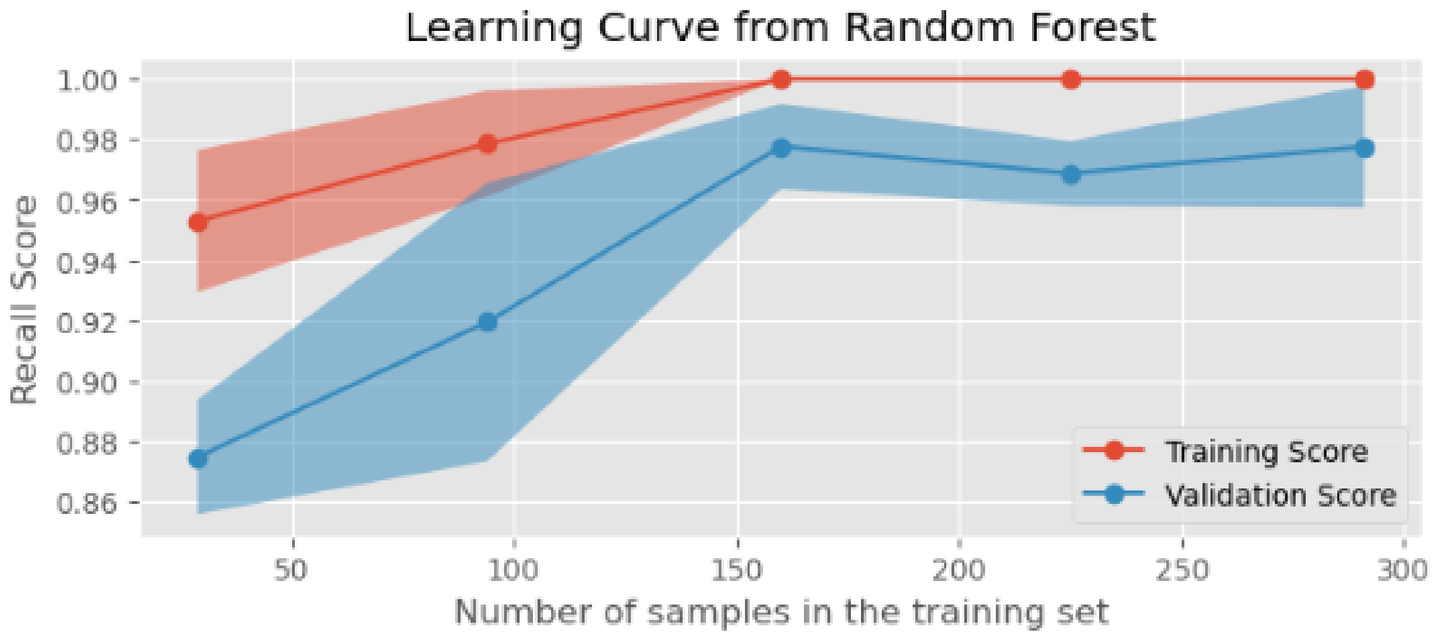



Random Forest Model Classification Report

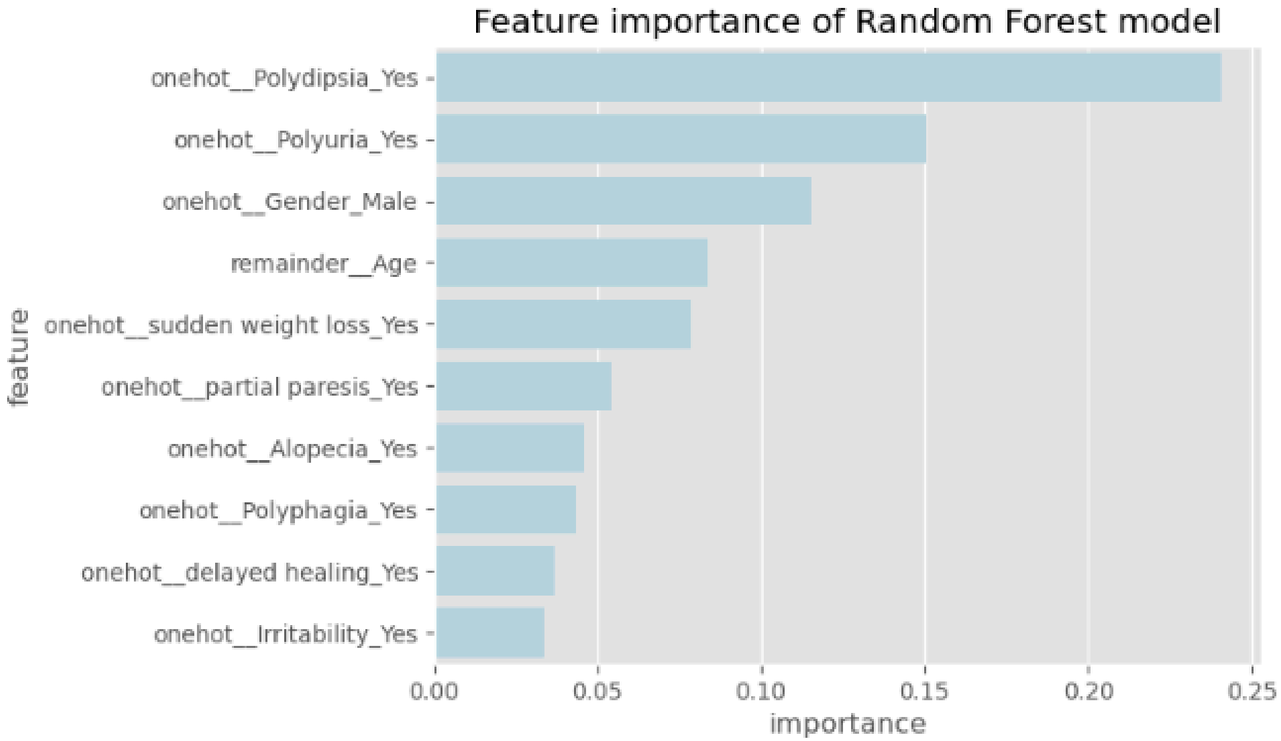


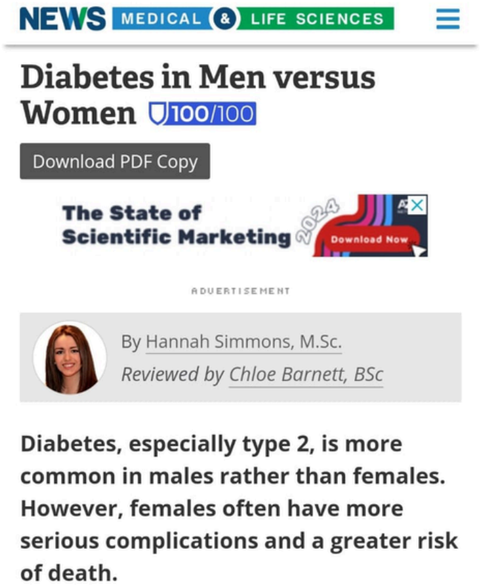

XGBoost Model Classification Report

# Learning curve



# Feature importance

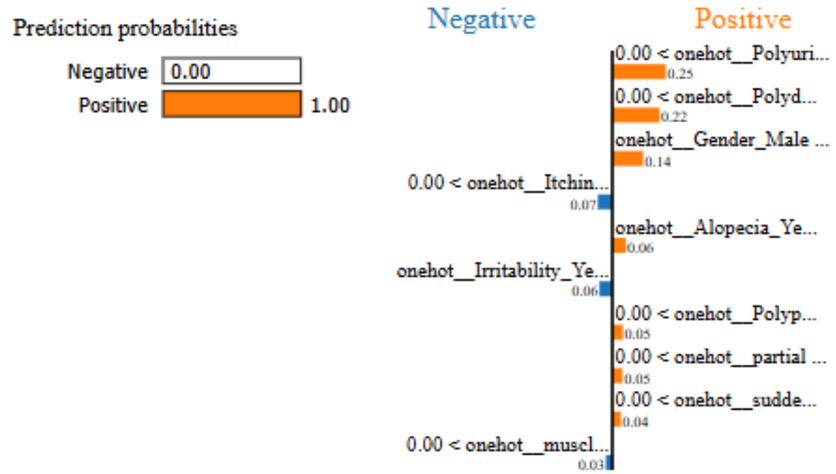


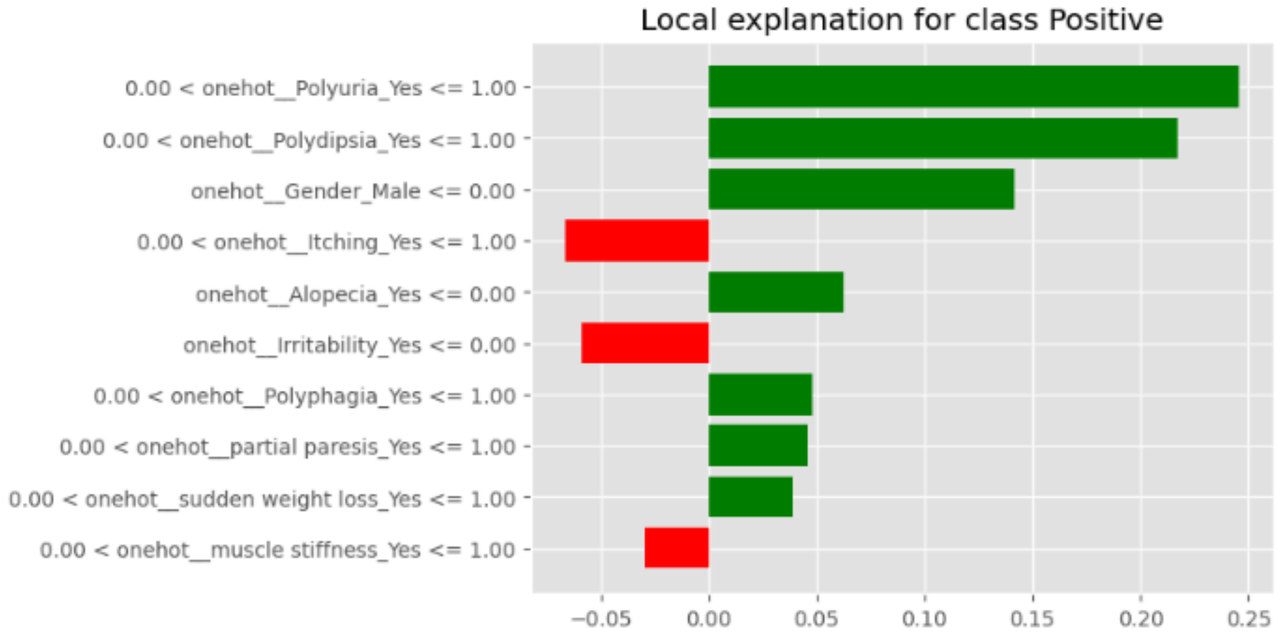

# LIME

## Patient [70] – Positive Diabetes

```
The prediction must predict this patient is: Positive diabetes from y_test
```




| Feature | Value |
|---|---|
| onehot__Polyuria_Yes | 1.00 |
| onehot__Polydipsia_Yes | 1.00 |
| onehot__Gender_Male | 0.00 |
| onehot__Itching_Yes | 1.00 |
| onehot__Alopecia_Yes | 0.00 |
| onehot__Irritability_Yes | 0.00 |
| onehot__Polyphagia_Yes | 1.00 |
| onehot__partial paresis_Yes | 1.00 |
| onehot__sudden weight loss_Yes | 1.00 |

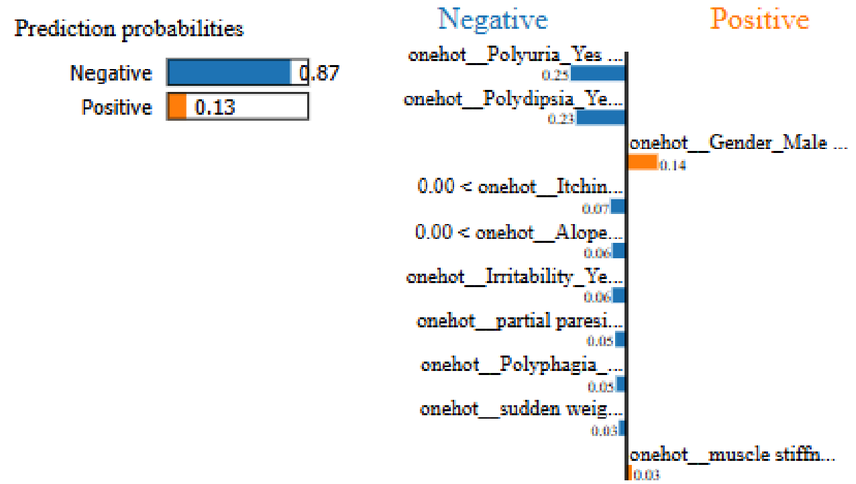



Positive

# LIME

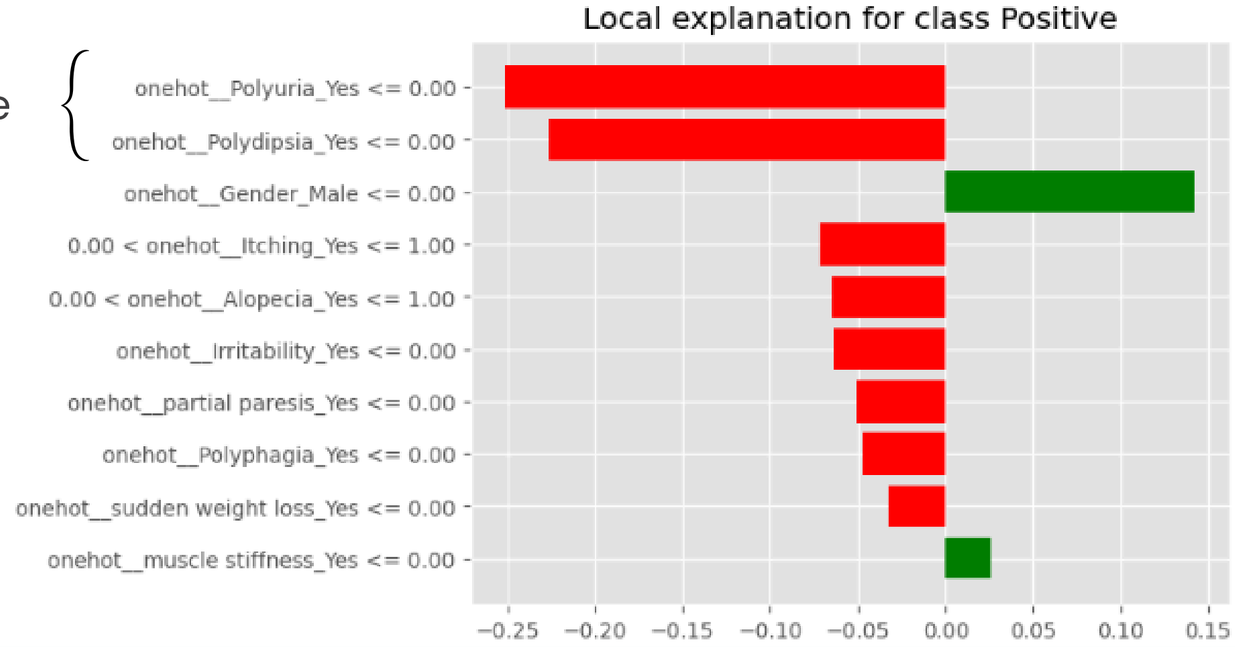
## Patient [71] – Negative Diabetes

```
The prediction must predict this patient is: Negative diabetes from y_test
```

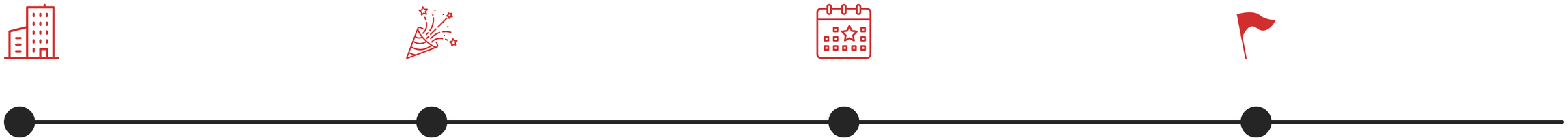



| Feature | Value |
|---|---|
| onehot__Polyuria_Yes | 0.00 |
| onehot__Polydipsia_Yes | 0.00 |
| onehot__Gender_Male | 0.00 |
| onehot__Itching_Yes | 1.00 |
| onehot__Alopecia_Yes | 1.00 |
| onehot__Irritability_Yes | 0.00 |
| onehot__partial paresis_Yes | 0.00 |
| onehot__Polyphagia_Yes | 0.00 |
| onehot__sudden weight loss_Yes | 0.00 |

Negative



# Conlusion



## Best Model

Tuned - Random Forest

## Performance

Recall train score 0.978
Recall test score 0.989

## Confusion matrix

- Small False Negative
- Small False Positive

## Feature importance & LIME

- Polyuria
- Polydipsia
- Gender - Male

# RECOMMENDATION


Health Monitoring: Incorporate regular health check-ups and symptom tracking

Further Research: Explore additional data sources and refine models

Further Research: Explore other features

# App Demonstration

## Diabetes risk predictor apps

### This apps predicts early symptoms of diabetes melitus

The dataset for this prediction was obtained from Diabetes symptoms dataset by UC Irvine ML repository.


THANK YOU


DIABETES