
# Chronic Disease Detection: A Study on Diabetes Classification

Group 3 – Joost, Paulo and Teja

Monday 13, May 2024

# Content

- Introduction
- Business Understanding
- Literature review
- Exploratory data analysis (EDA)
- Data Preparation
- Modeling, Training & Evaluation
- Conclusion
- Challenges & Future work

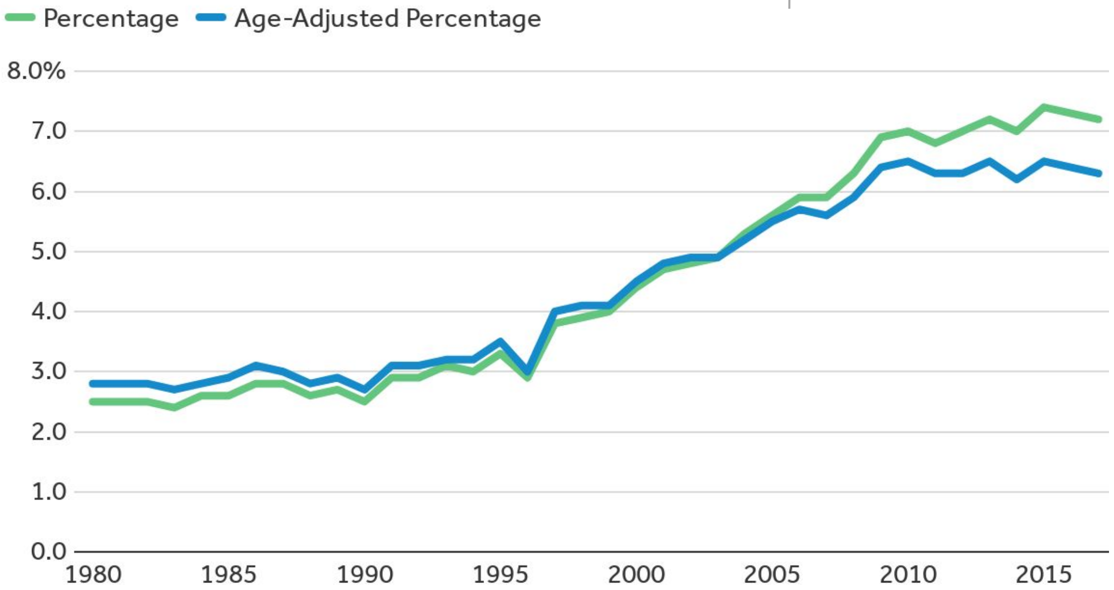
# Introduction: Situation and Problem

- The healthcare sector has a high workload.
- The prevalence of have diabetes is increasing.
- One in five people with diabetes doesn't know they have it.

- Untreated diabetes affects many major organs, including heart, blood vessels, nerves, eyes and kidneys.

- The combination of increasing prevalence of diabetes and the increasing workload for GP´s.

**Creating a machine learning model which classifies if a patient visiting a GP is having diabetes**




# Methodology

*“Target Group” Patients without Diabetes and with glucose exams history*

## Methodology to follow: CRISP-DM

**Main research question:**

“What methodologies and techniques should be used for developing a machine learning model to assist general practitioners in accurately diagnosing diabetes, while simultaneously alleviating their workload, considering key objectives, available data, data preprocessing, choice of algorithms, and hyperparameter tuning?”

# Business Understanding - Patient Registration Flow

➔ First contact with the medical facility

Manage costs and efficient billing

Patient Data

Decision Maker

Medical Facility

Registration Form

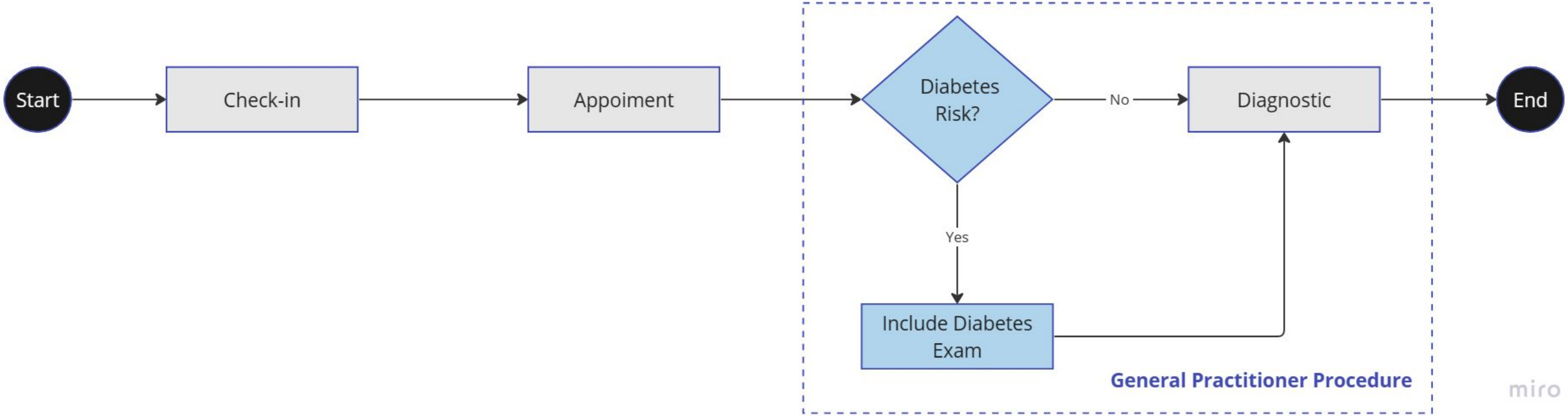
Appointment





Slides courtesy of Prof. Dario Di Nucci

# Business Understanding - Appointment Flow + AI Feature




*"Target Group" Patients without Diabetes and with glucose exams history*

*"Diabetes Risk Predictor" in the General Practioner system showing the Diabetes risk on patient profile.*



Jheronimus Academy of Data Science

Slides courtesy of Prof. Dario Di Nucci

# Literature Review/Related Work

| Author | Year | Project Name | Algorithms used | Accuracy | Adoptions made |
|---|---|---|---|---|---|
| Li, Mingqi, Xiaoyang Fu, and Dongdong Li. | 2020 | Diabetes prediction based on XGBoost algorithm | XGBoost | 80.20% | Gradient Boost |
| Mahabub, Atik. | 2019 | A robust voting approach for diabetes prediction using traditional machine learning techniques. | AdaBoost, gradient boost, XGBoost, random forest, etc. | 84.42% | |
| Mushtaq, Zaigham, Muhammad Farhan Ramzan, Sikandar Ali, Samad Baseer, Ali Samad, and Mujtaba Husnain. | 2022 | Voting classification-based diabetes mellitus prediction using hypertuned machine-learning techniques. | Voting Classifier (includes Random Forest, logistic regression, Support Vector Machine, KNN, Naive Bayes Theorem, and Gradient Boosting Classifier | 81.50% | Voting Classifier |
| Shahid Mohammad Ganie | 2023 | An ensemble learning approach for diabetes prediction using boosting techniques | Gradient boosting algorithm | 96.00% | Gradient boosting |
| Lai, Hang, Huaxiong Huang, Karim Keshavjee, Aziz Guergachi, and Xin Gao. | 2019 | Predictive models for diabetes mellitus using machine learning techniques | Logistic Regression | 88.00% | Logistic Regression |

# Data Understanding

- ‘Diabetes prediction dataset’ sourced from Kaggle repository
- Contains  100,000 records with 9 features
- The data is a collection of medical and demographic data from patients
- Categorical variables: Gender, smoking history
- Numerical variables: Age, hypertension, heart disease, smoking history, BMI, HbA1c level, diabetes
- Dependent/Predicted Variable: Diabetes status (binary classification: 1 or 0).
- Independent/Predictor Variables: Age, gender, BMI, hypertension, heart disease, smoking history, HbA1c level, and blood glucose levels.

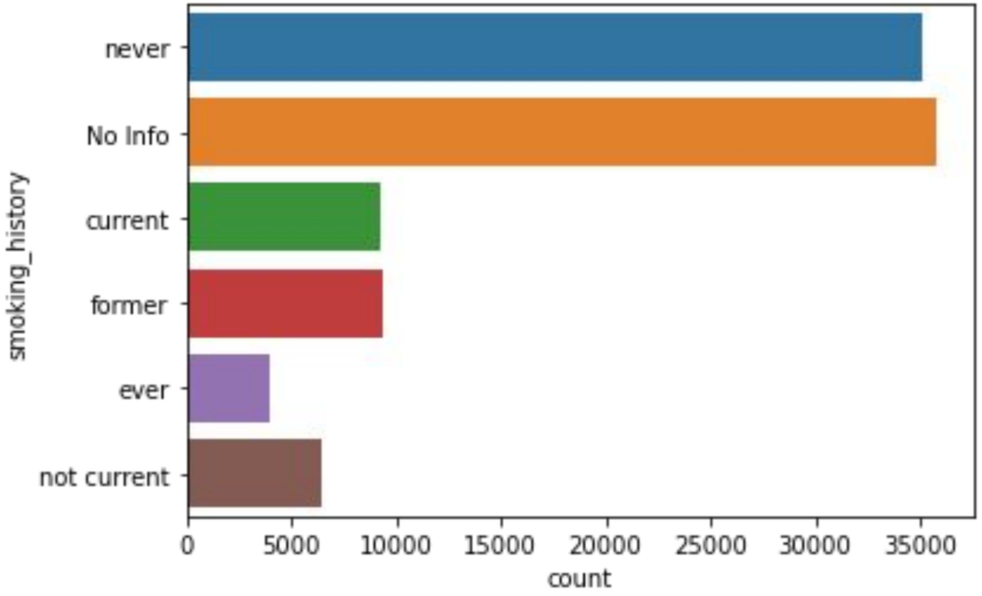
| | gender | age | hypertension | heart_disease | smoking_history | bmi | HbA1c_level | blood_glucose_level | diabetes |
|---|---|---|---|---|---|---|---|---|---|
| 0 | Female | 80.00 | 0 | 1 | never | 25.19 | 6.60 | 140 | 0 |
| 1 | Female | 54.00 | 0 | 0 | No Info | 27.32 | 6.60 | 80 | 0 |
| 2 | Male | 28.00 | 0 | 0 | never | 27.32 | 5.70 | 158 | 0 |
| 3 | Female | 36.00 | 0 | 0 | current | 23.45 | 5.00 | 155 | 0 |
| 4 | Male | 76.00 | 1 | 1 | current | 20.14 | 4.80 | 155 | 0 |

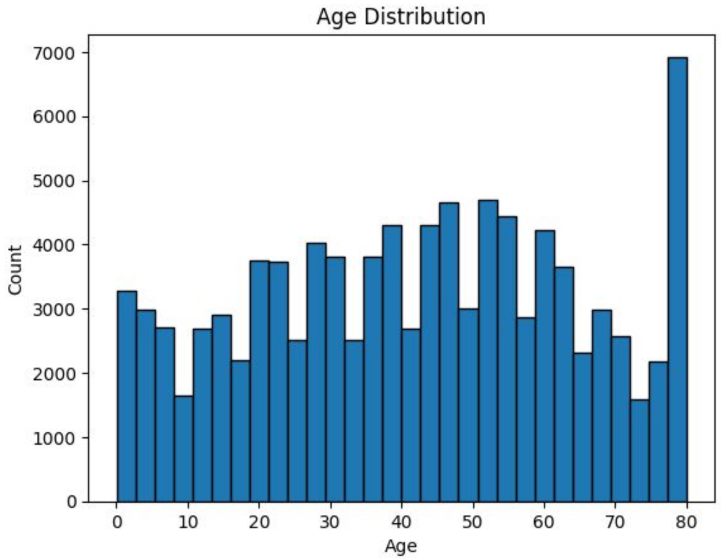


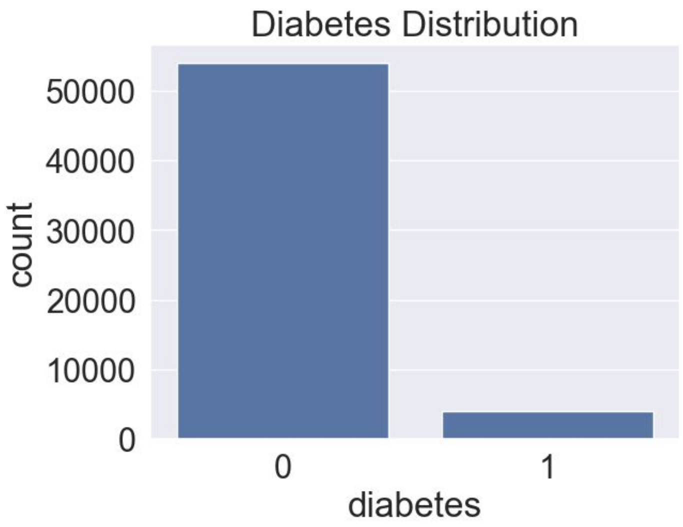
Jheronimus Academy of Data Science

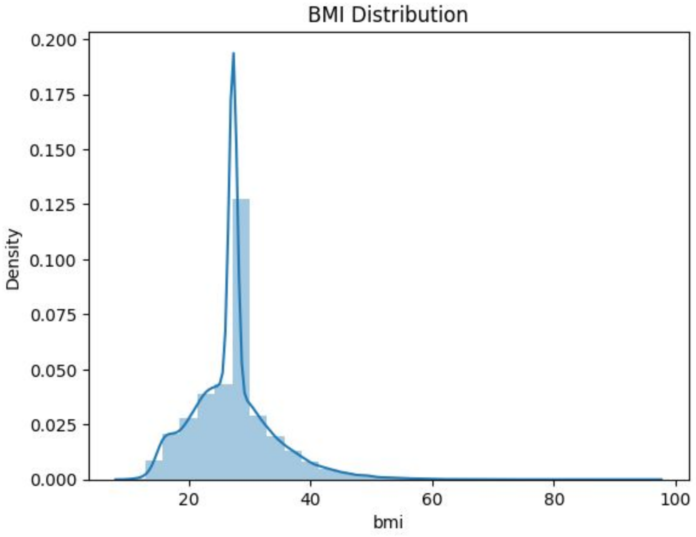
Slides courtesy of Prof. Dario Di Nucci

# EDA Summary



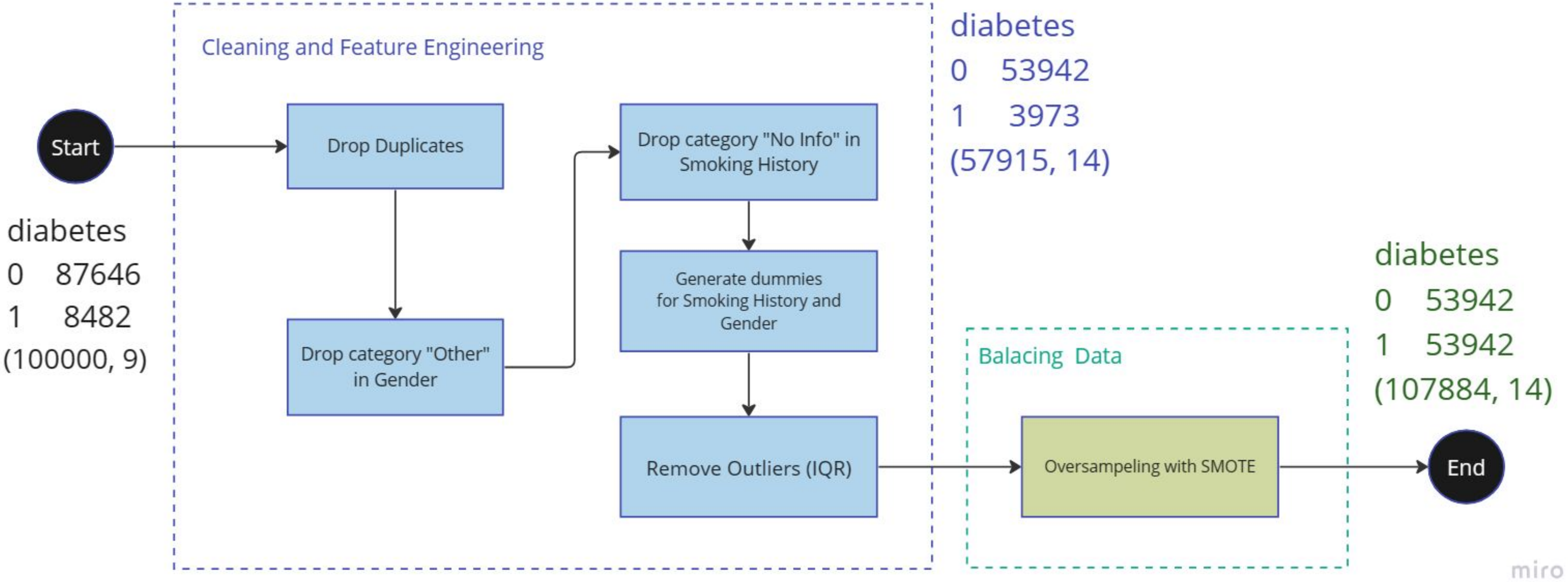






Slides courtesy of Prof. Dario Di Nucci

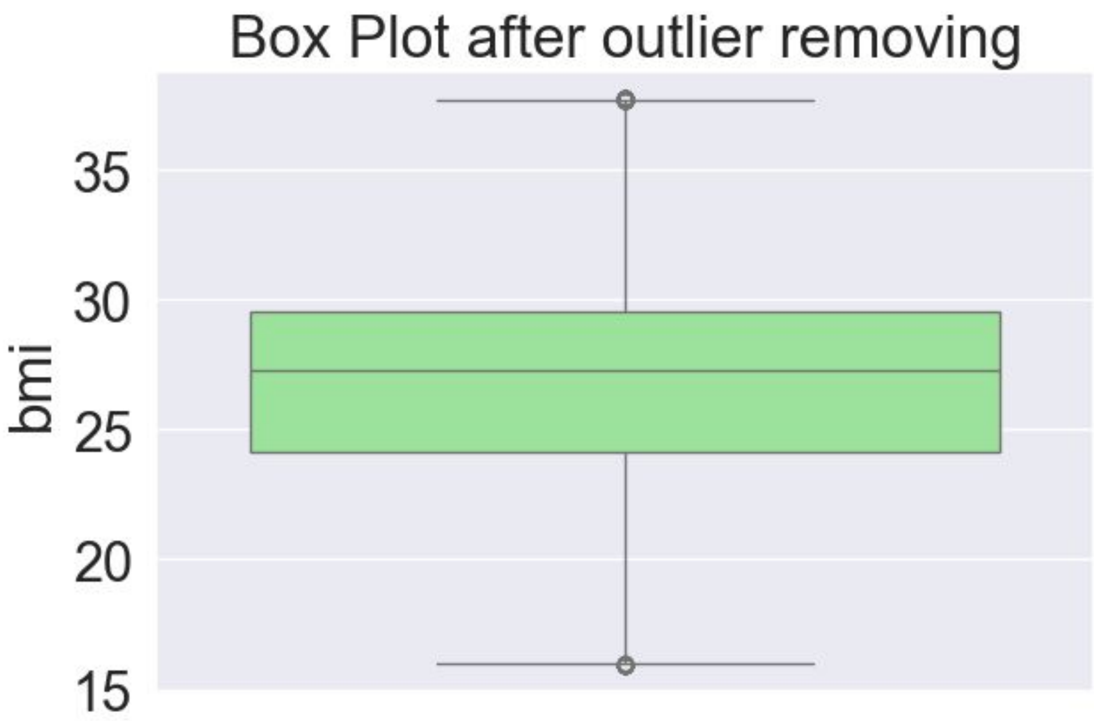
# Data Cleaning and Feature Engineering



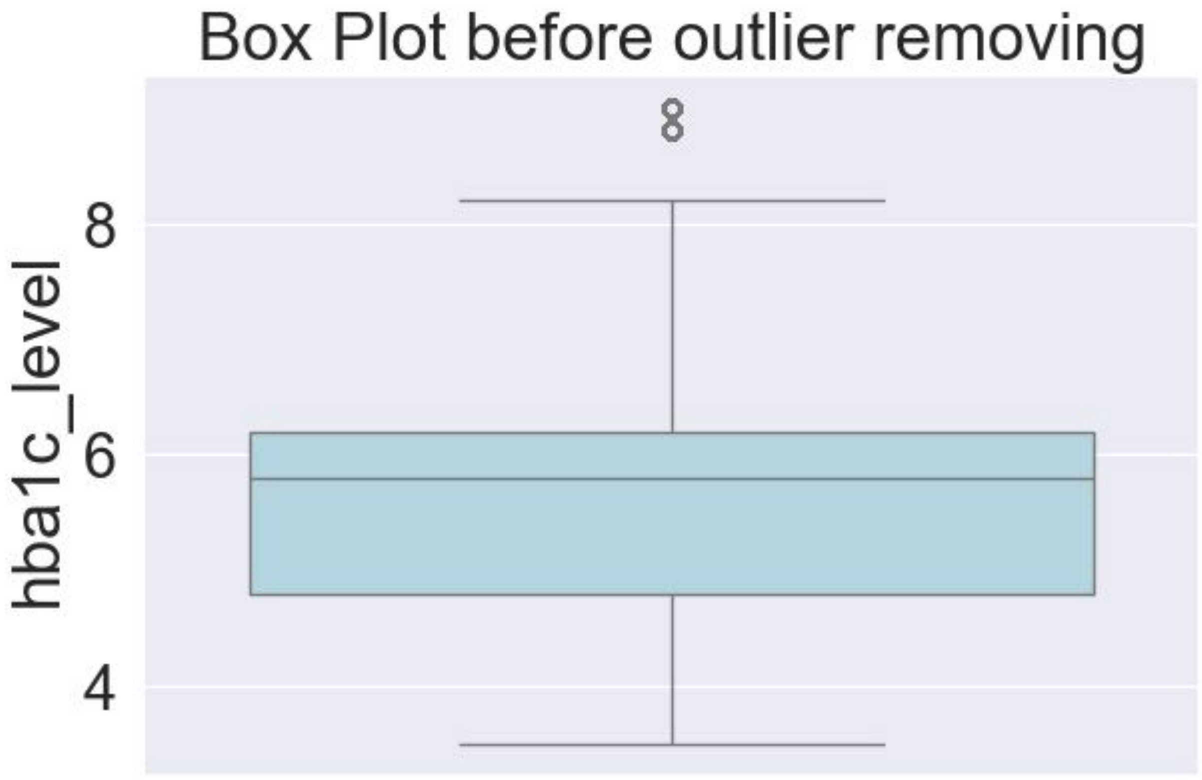


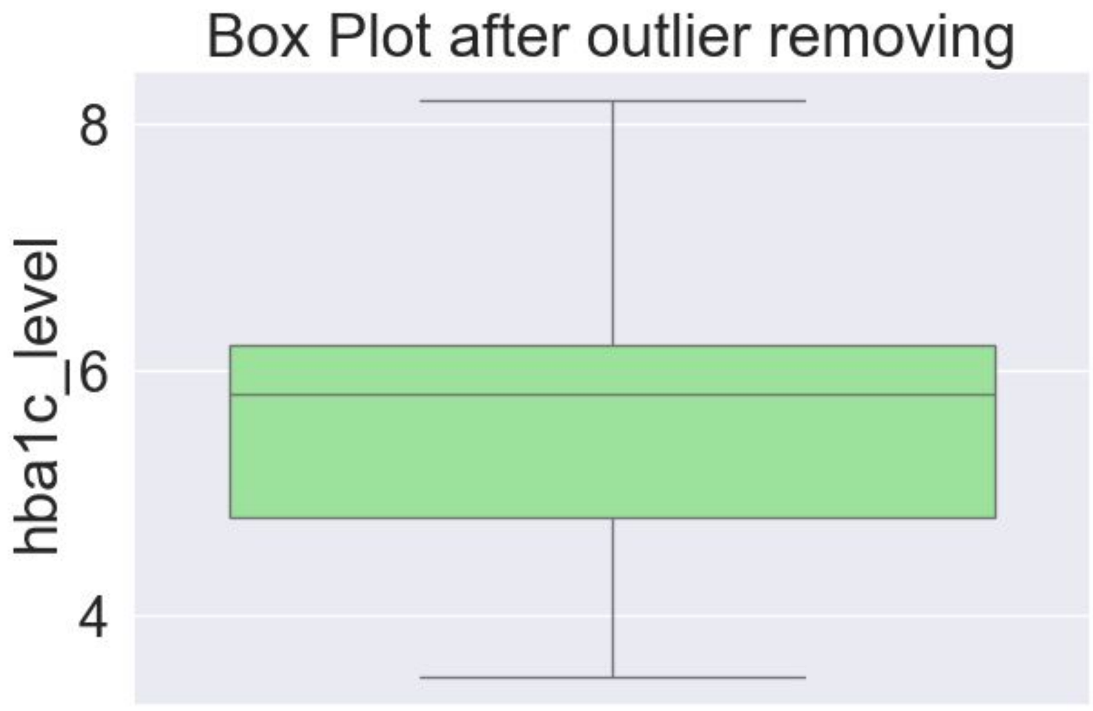

Slides courtesy of Prof. Dario Di Nucci

# Outliers - BMI

# Outliers - HbA1c Level



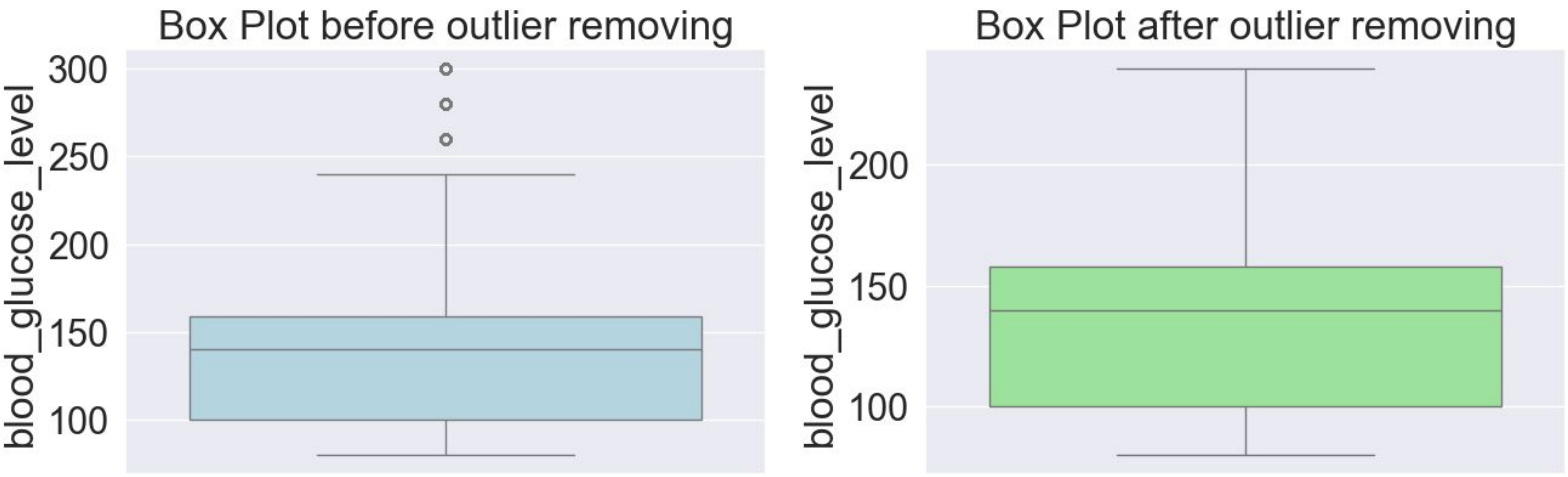








Slides courtesy of Prof. Dario Di Nucci

# Outliers - Blood glucose level






JADS Jheronimus Academy of Data Science

[Slides courtesy of Prof. Dario Di Nucci]

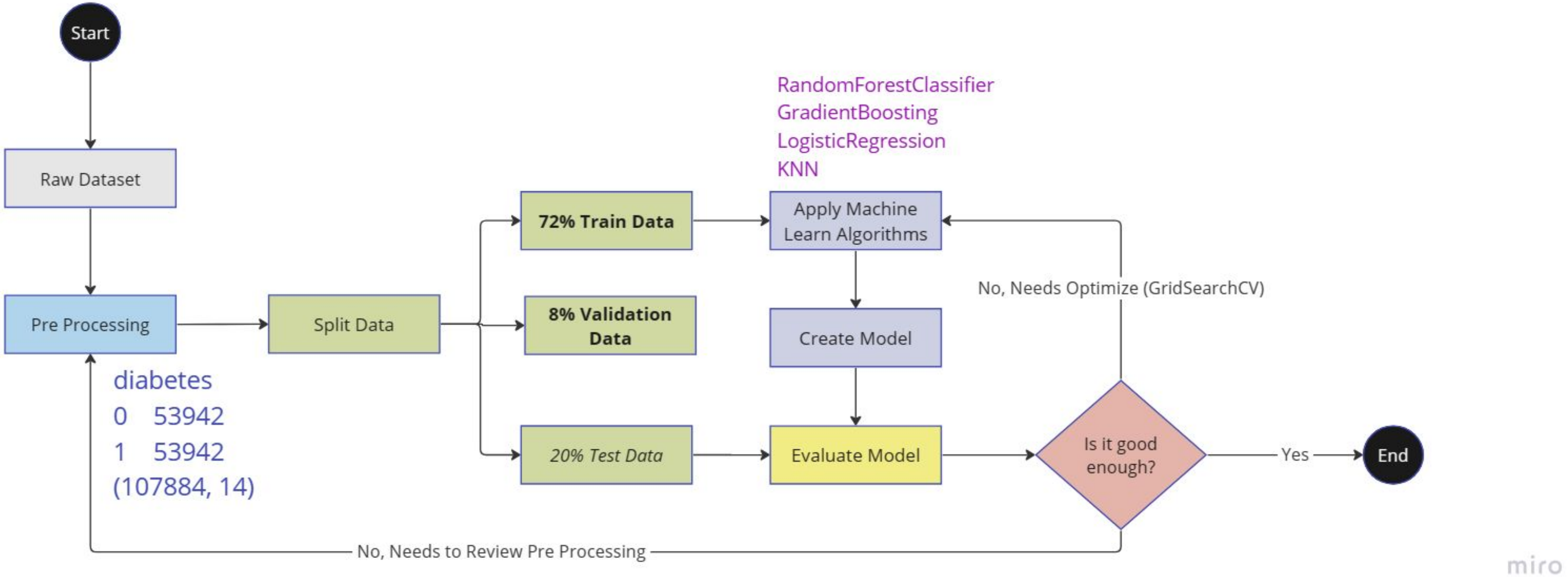
# Correlation Matrix

Features Correlating with Diabetes

| | diabetes |
|---|---|
| diabetes | 1 |
| HbA1c_level | 0.3 |
| blood_glucose_level | 0.25 |
| age | 0.22 |
| hypertension | 0.16 |
| bmi | 0.15 |
| heart_disease | 0.14 |
| smoking_history_former | 0.068 |
| gender_Male | 0.054 |
| smoking_history_ever | 0.0068 |
| smoking_history_not current | -0.0063 |
| smoking_history_current | -0.011 |
| smoking_history_never | -0.039 |
| gender_Female | -0.054 |

| | age | hypertension | heart_disease | bmi | HbA1c_level | blood_glucose_level | diabetes | smoking_history_current | smoking_history_ever | smoking_history_former | smoking_history_never | smoking_history_not current | gender_Female | gender_Male |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| age | 1 | 0.25 | 0.23 | 0.2 | 0.068 | 0.053 | 0.22 | -0.048 | 0.038 | 0.22 | -0.16 | 0.022 | -0.02 | 0.02 |
| hypertension | 0.25 | 1 | 0.12 | 0.12 | 0.05 | 0.043 | 0.16 | -0.0075 | 0.0082 | 0.062 | -0.026 | -0.027 | -0.03 | 0.03 |
| heart_disease | 0.23 | 0.12 | 1 | 0.045 | 0.04 | 0.036 | 0.14 | -0.0067 | 0.038 | 0.095 | -0.079 | -0.0021 | -0.092 | 0.092 |
| bmi | 0.2 | 0.12 | 0.045 | 1 | 0.041 | 0.04 | 0.15 | 0.0079 | 0.019 | 0.082 | -0.067 | -0.01 | -0.062 | 0.062 |
| HbA1c_level | 0.068 | 0.05 | 0.04 | 0.041 | 1 | 0.071 | 0.3 | -0.0062 | 0.0029 | 0.019 | -0.011 | -9.4e-06 | -0.021 | 0.021 |
| blood_glucose_level | 0.053 | 0.043 | 0.036 | 0.04 | 0.071 | 1 | 0.25 | -0.0006 | -0.0018 | 0.016 | -0.0068 | -0.0055 | -0.018 | 0.018 |
| diabetes | 0.22 | 0.16 | 0.14 | 0.15 | 0.3 | 0.25 | 1 | -0.011 | 0.0068 | 0.068 | -0.039 | -0.0063 | -0.054 | 0.054 |
| smoking_history_current | -0.048 | -0.0075 | -0.0067 | 0.0079 | -0.0062 | -0.0006 | -0.011 | 1 | -0.11 | -0.17 | -0.45 | -0.14 | -0.05 | 0.05 |
| smoking_history_ever | 0.038 | 0.0082 | 0.038 | 0.019 | 0.0029 | -0.0018 | 0.0068 | -0.11 | 1 | -0.11 | -0.29 | -0.087 | -0.021 | 0.021 |
| smoking_history_former | 0.22 | 0.062 | 0.095 | 0.082 | 0.019 | 0.016 | 0.068 | -0.17 | -0.11 | 1 | -0.45 | -0.14 | -0.077 | 0.077 |
| smoking_history_never | -0.16 | -0.026 | -0.079 | -0.067 | -0.011 | -0.0068 | -0.039 | -0.45 | -0.29 | -0.45 | 1 | -0.37 | 0.1 | -0.1 |
| smoking_history_not current | 0.022 | -0.027 | -0.0021 | -0.01 | -9.4e-06 | -0.0055 | -0.0063 | -0.14 | -0.087 | -0.14 | -0.37 | 1 | 0.00038 | -0.00038 |
| gender_Female | -0.02 | -0.03 | -0.092 | -0.062 | -0.021 | -0.018 | -0.054 | -0.05 | -0.021 | -0.077 | 0.1 | 0.00038 | 1 | -1 |
| gender_Male | 0.02 | 0.03 | 0.092 | 0.062 | 0.021 | 0.018 | 0.054 | 0.05 | 0.021 | 0.077 | -0.1 | -0.00038 | -1 | 1 |

# ML Flow overview






Jheronimus Academy of Data Science

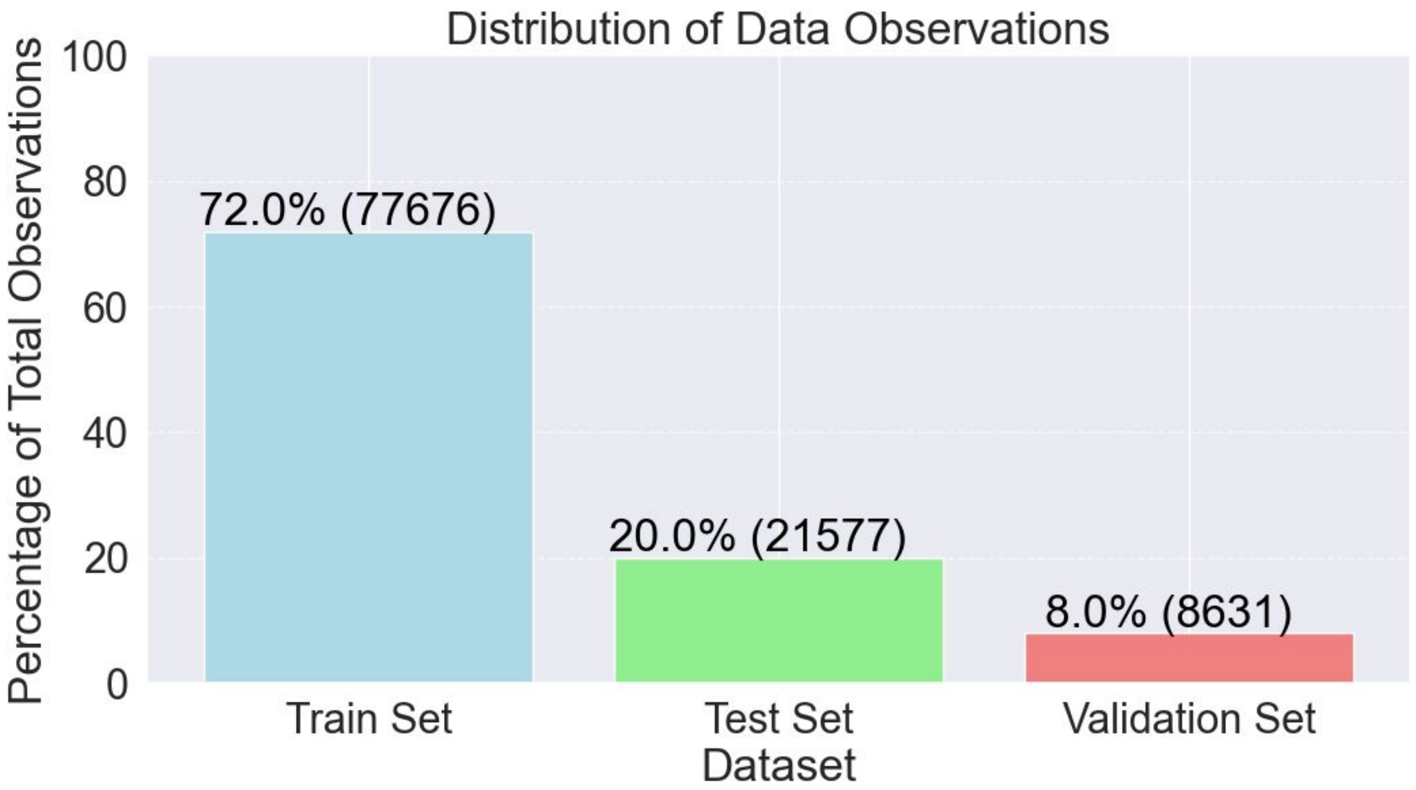
Slides courtesy of Prof. Dario Di Nucci

# Splitting Train, Test and validation Set

```python
from sklearn.model_selection import train_test_split

# Split the dataset into train and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, train_size=0.8)

# Further split the train set into train and validation sets
X_train, X_valid, y_train, y_valid = train_test_split(X_train, y_train, train_size=0.9)
```

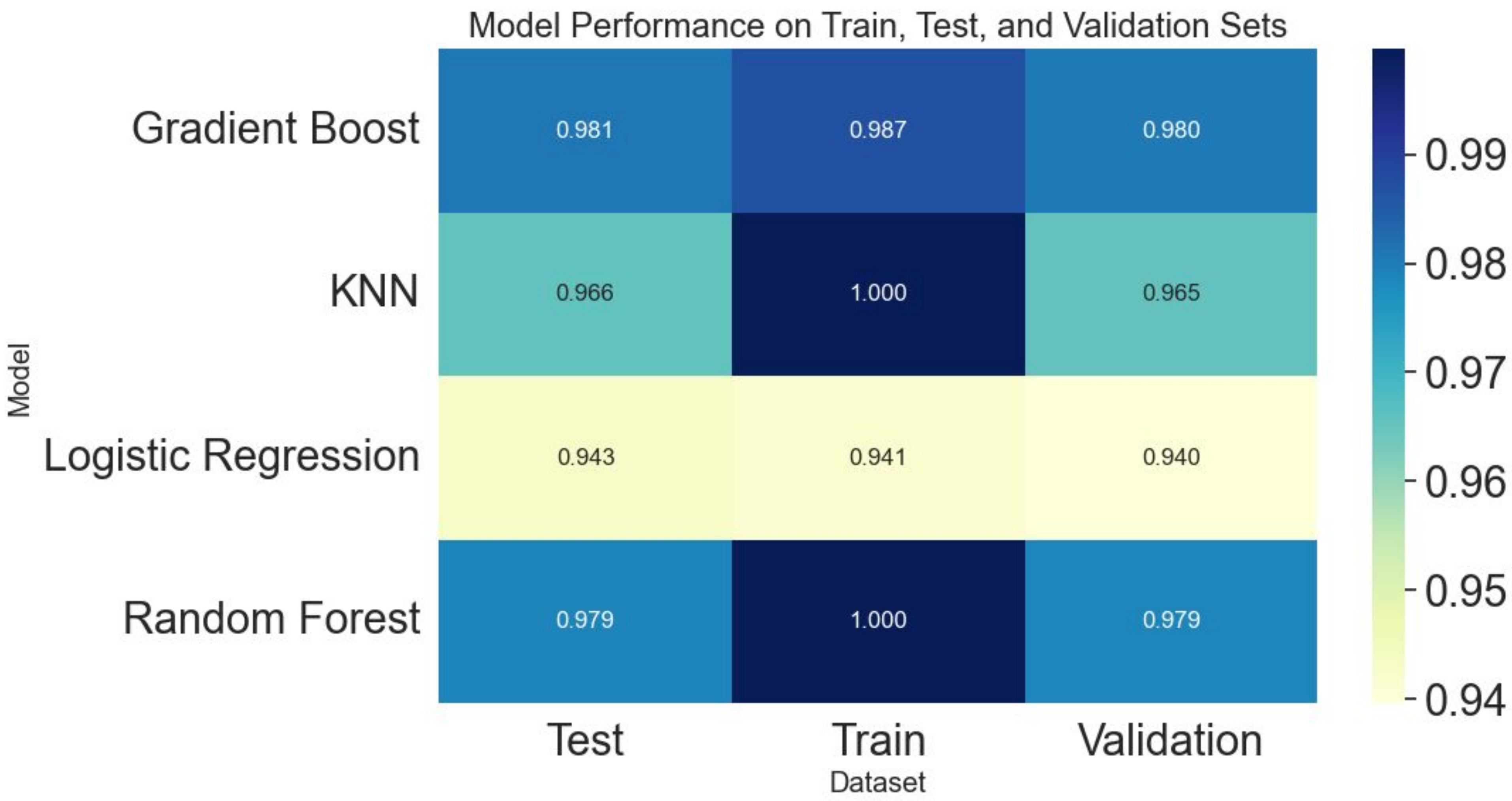





JADS Jheronimus Academy of Data Science

Slides courtesy of Prof. Dario Di Nucci

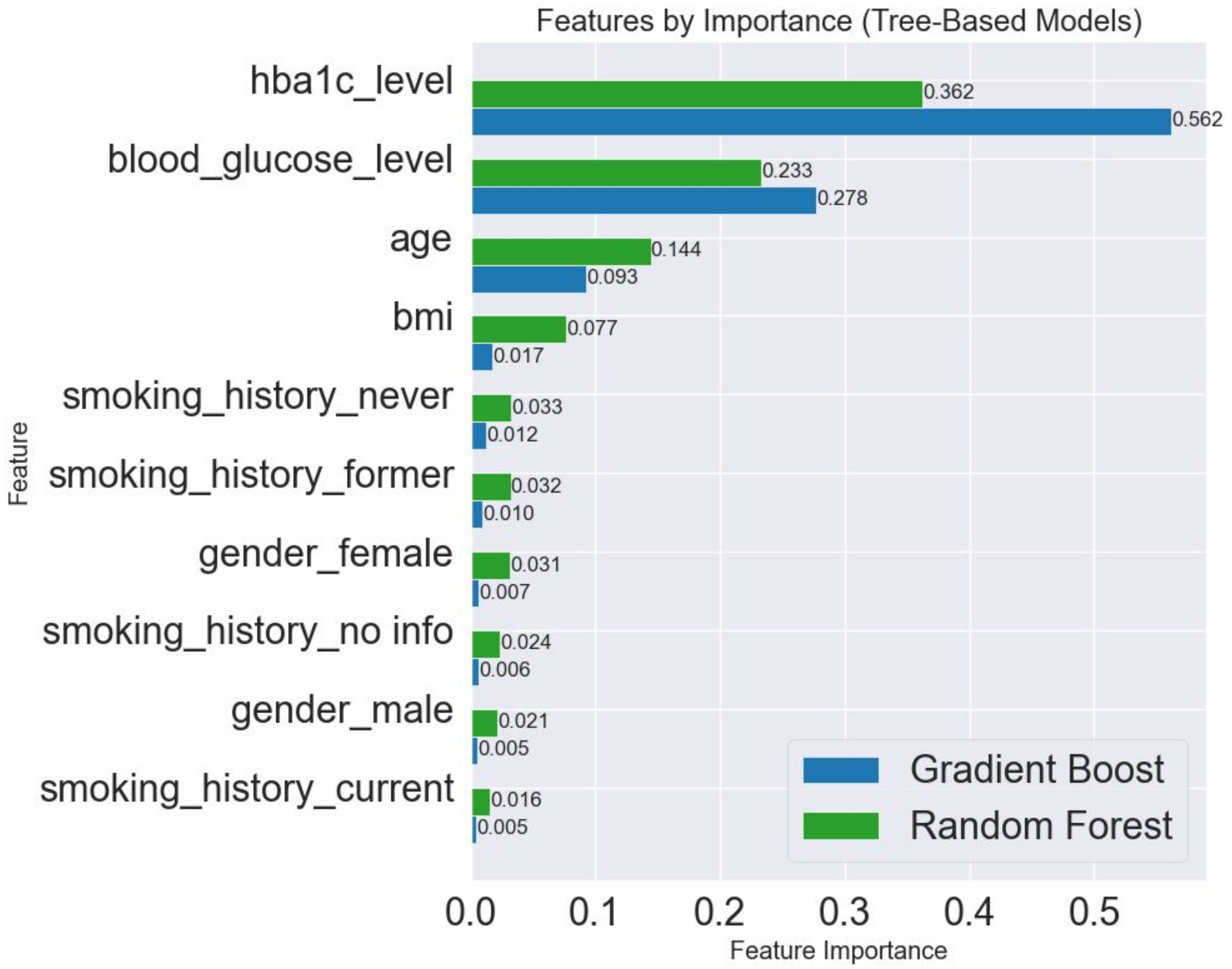
# Modeling Training Results - Accuracy

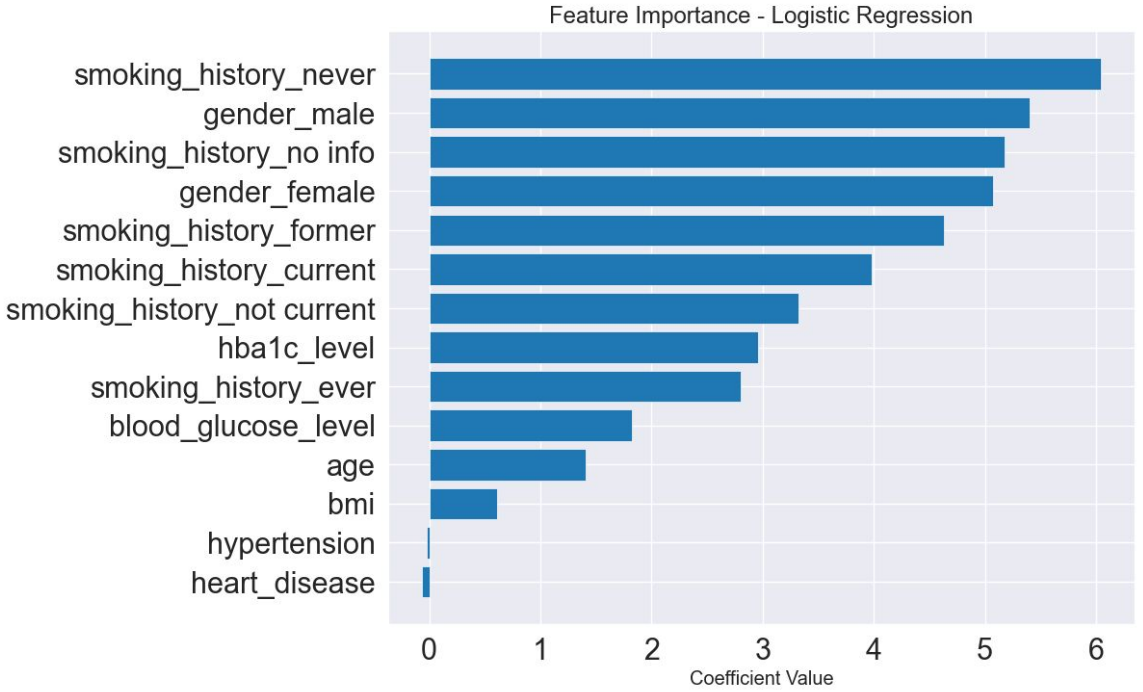



JADS Jheronimus Academy of Data Science

Slides courtesy of Prof. Dario Di Nucci

# Feature Importance

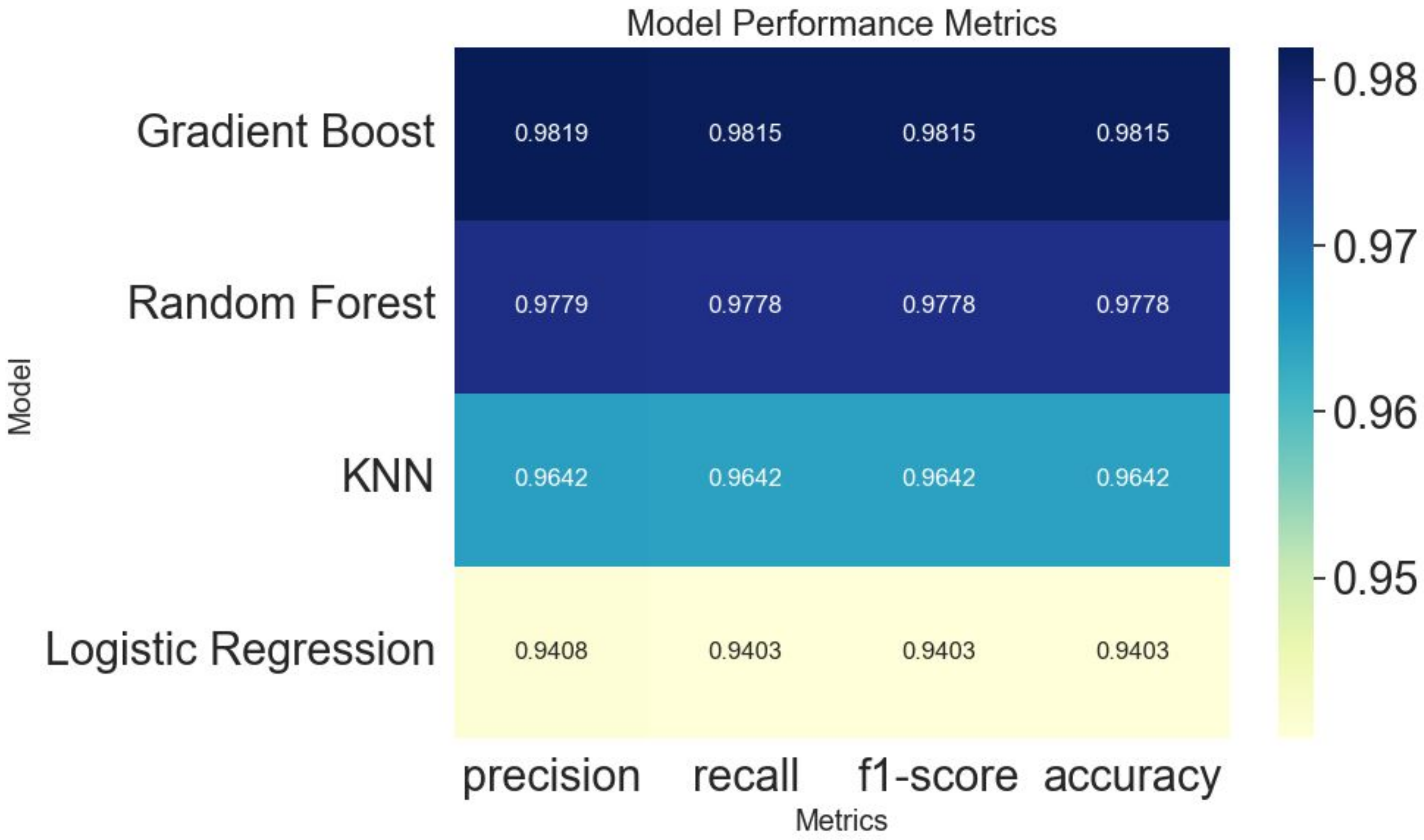



*Gender and smoking history more important for Logistic Regression*

Slides courtesy of Prof. Dario Di Nucci

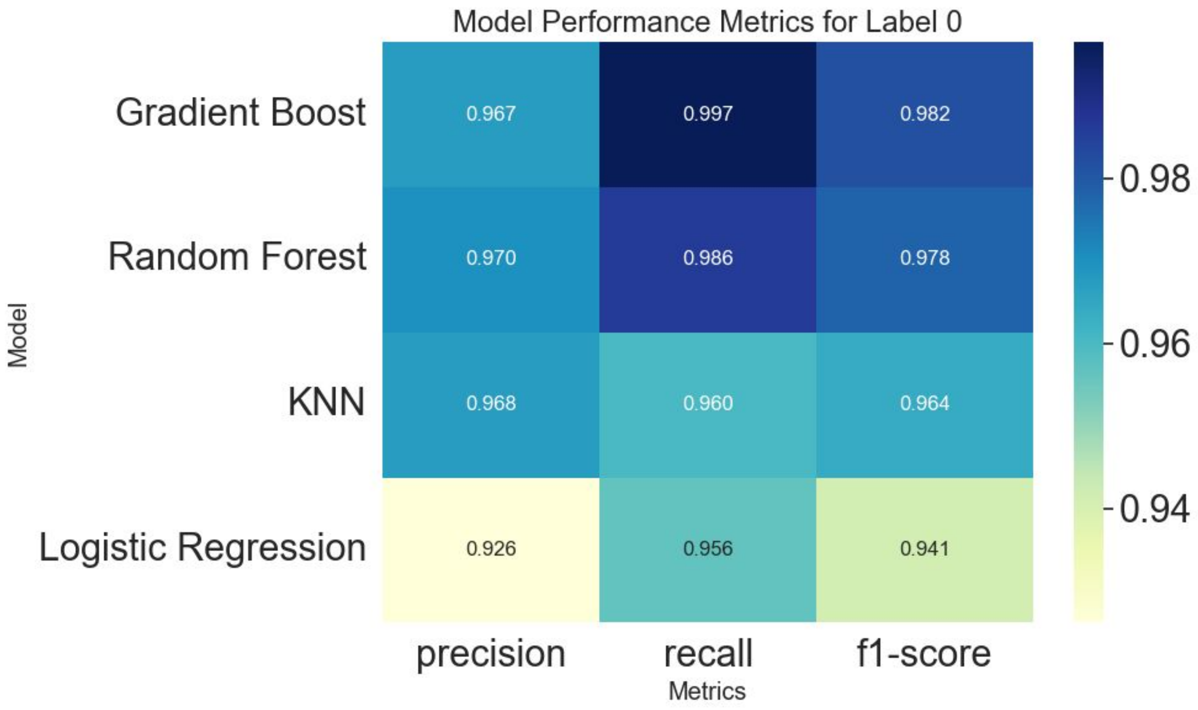
# Modeling - Overall Results

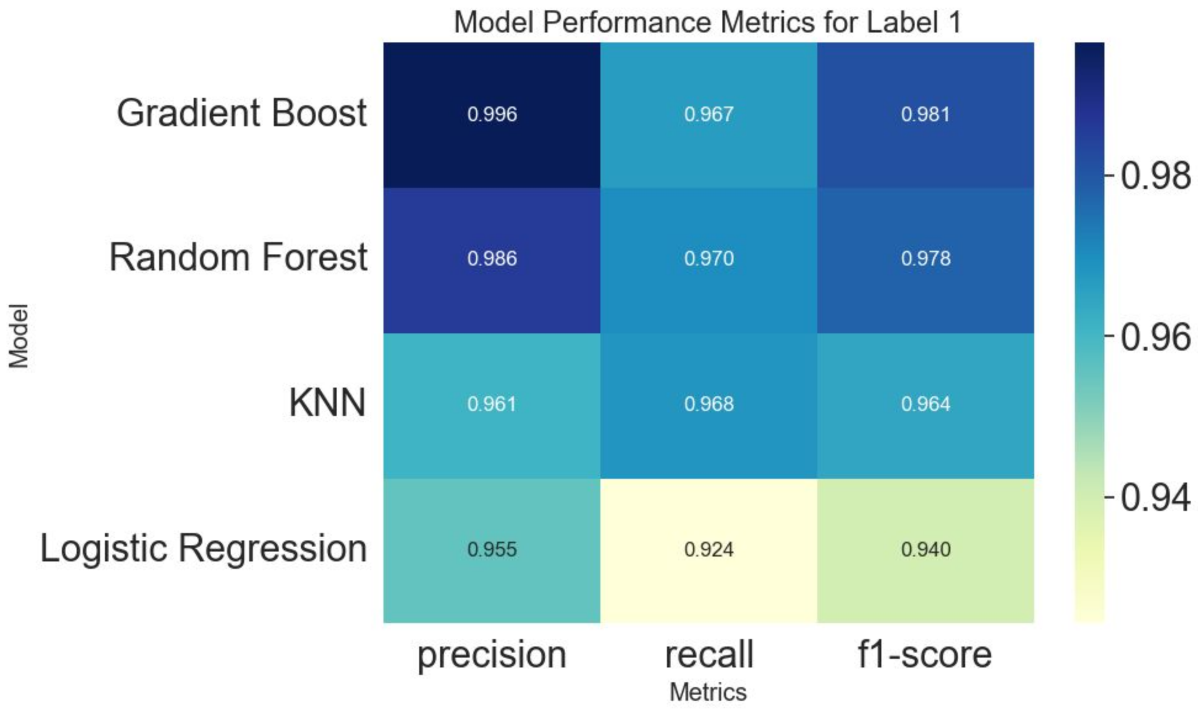



*Best Results:*

*Gradient Boost*

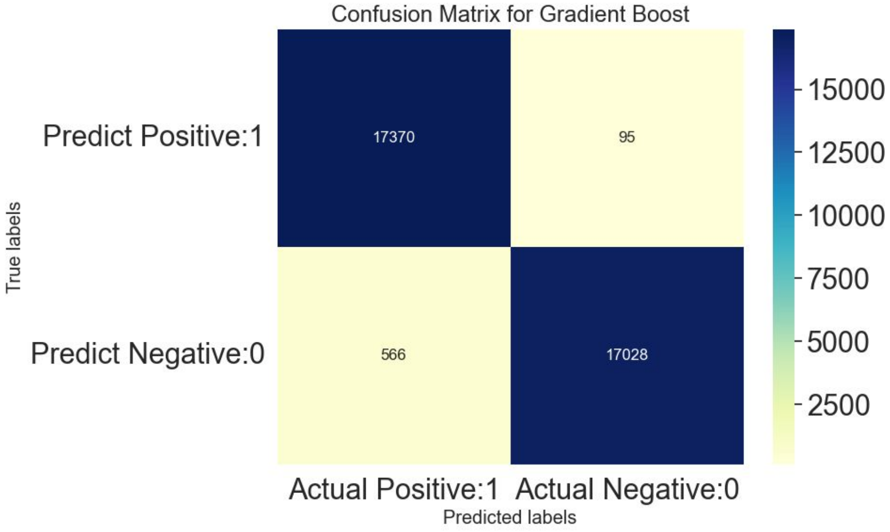
# Modeling - Performance per label

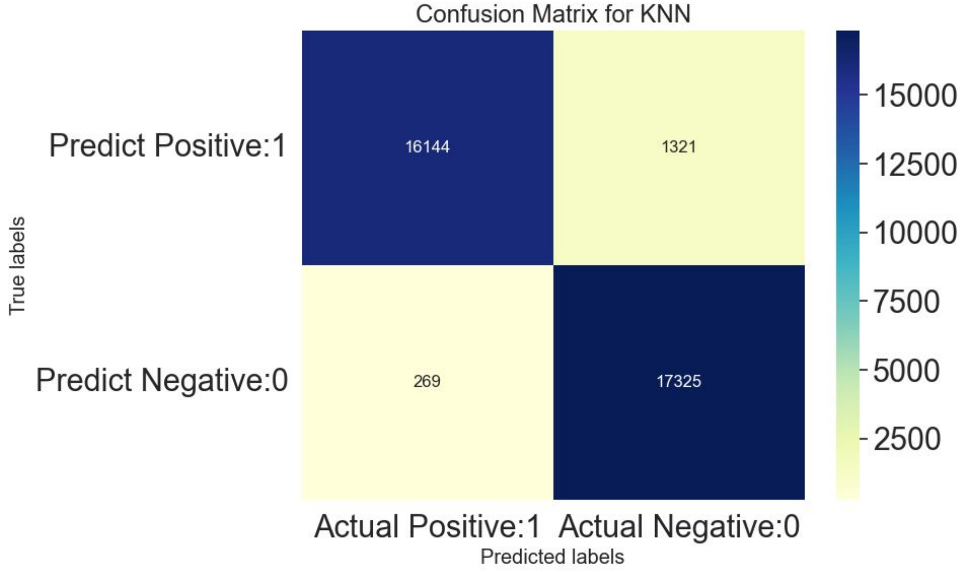



*Best Results: Gradient Boost and Random Forest*

# Modeling - Confusion Matrix



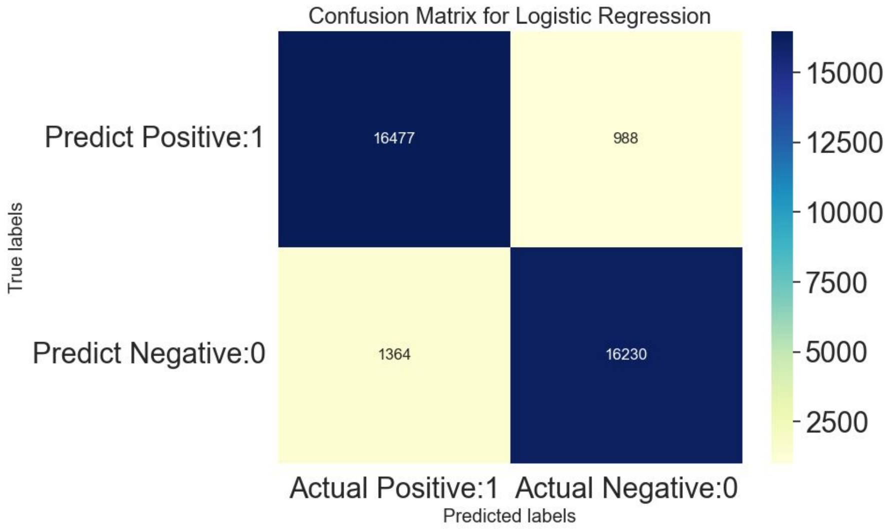



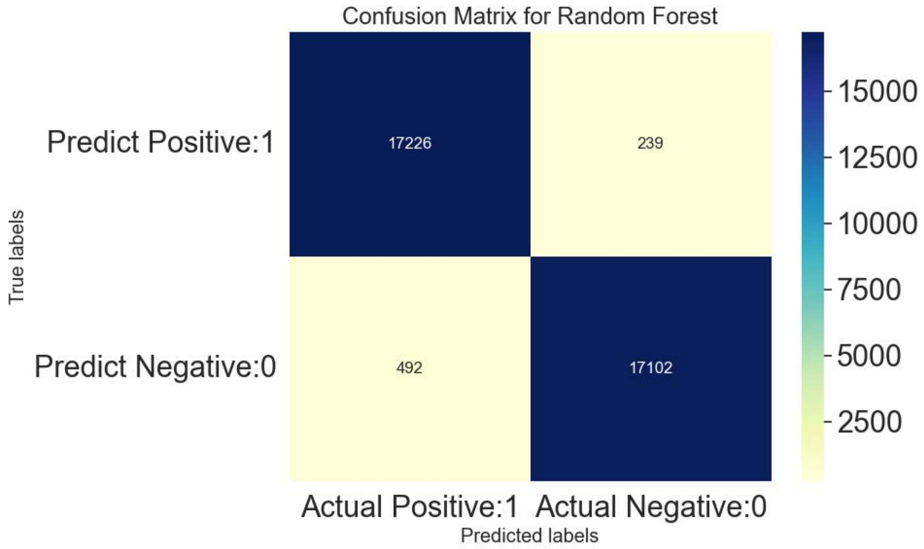

*Lowest False Negative:*

*KNN*

*Lowest False Positive:*

*Gradient Boost*



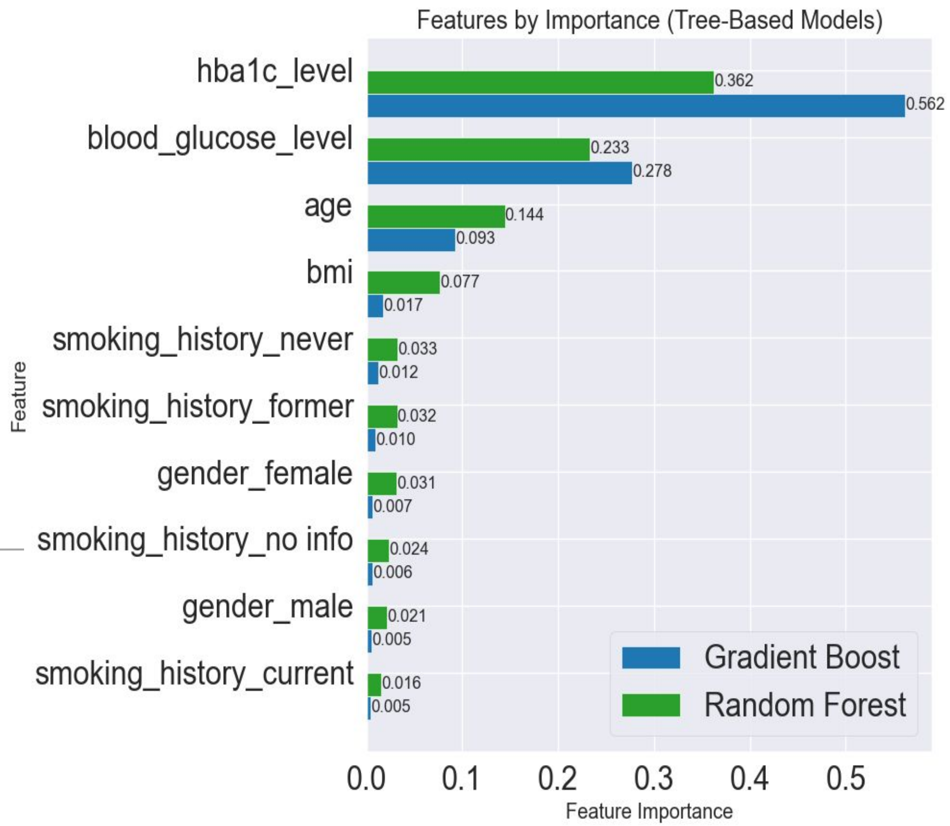
# Conclusion

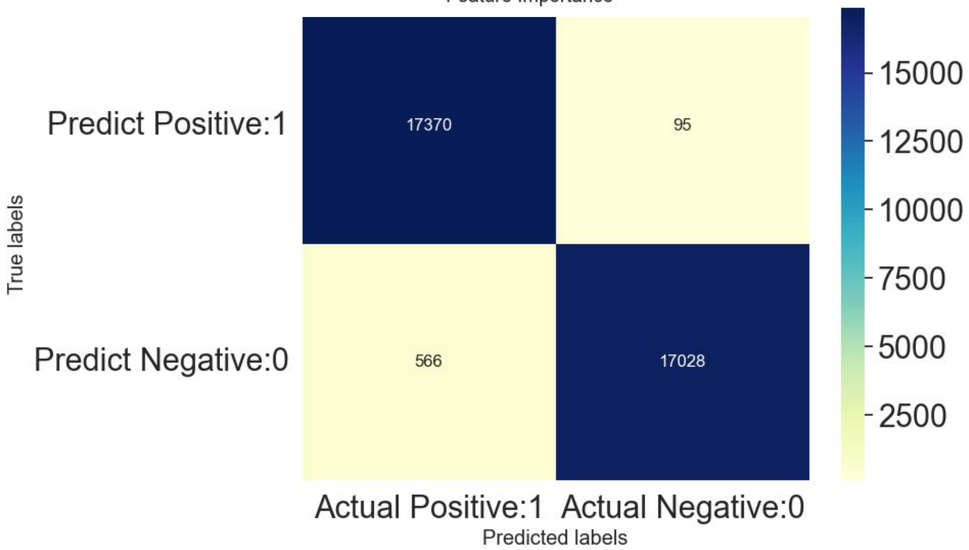
**Main research question:**

"What methodologies and techniques should be used for developing a machine learning model to assist general practitioners in accurately diagnosing diabetes, while simultaneously alleviating their workload, considering key objectives, available data, data preprocessing, choice of algorithms, and hyperparameter tuning?"

- Gradient Boosting technique
- Based on: HBA1C-level and Blood Glucose level
- 566 false negatives of the 22484 patients

- Reducing workload GP's
- Increasing accuracy in identifying diabetes



# Challenges & Future Work

## Limitations:

Number of (relevant) features to further:

- Decrease the workload of a GP
- Create a machine learning model based on demographic data

## Future work:

- Cluster different types of patients to further increase accuracy
- Create a risk predictor algorithm which predicts the risk of getting/having diabetes expressed in percentages



JADS Jheronimus Academy of Data Science

Slides courtesy of Prof. Dario Di Nucci

# References


- Mahabub, Atik. “A Robust Voting Approach for Diabetes Prediction Using Traditional Machine Learning Techniques.” *SN Applied Sciences* 1, no. 12 (November 25, 2019): 1667. https://doi.org/10.1007/s42452-019-1759-7.
- Ganie, Shahid Mohammad, Pijush Kanti Dutta Pramanik, Majid Bashir Malik, Saurav Mallik, and Hong Qin. “An Ensemble Learning Approach for Diabetes Prediction Using Boosting Techniques.” *Frontiers in Genetics* 14 (October 26, 2023). https://doi.org/10.3389/fgene.2023.1252159.
- Li, Mingqi, Xiaoyang Fu, and Dongdong Li. “Diabetes Prediction Based on XGBoost Algorithm.” *IOP Conference Series: Materials Science and Engineering* 768, no. 7 (March 2020): 072093. https://doi.org/10.1088/1757-899X/768/7/072093.
- Lai, Hang, Huaxiong Huang, Karim Keshavjee, Aziz Guergachi, and Xin Gao. “Predictive Models for Diabetes Mellitus Using Machine Learning Techniques.” *BMC Endocrine Disorders* 19, no. 1 (October 15, 2019): 101. https://doi.org/10.1186/s12902-019-0436-6.
- Mushtaq, Zaigham, Muhammad Farhan Ramzan, Sikandar Ali, Samad Baseer, Ali Samad, and Mujtaba Husnain. “Voting Classification-Based Diabetes Mellitus Prediction Using Hypertuned Machine-Learning Techniques.” Mobile Information Systems 2022 (March 19, 2022): e6521532. https://doi.org/10.1155/2022/6521532.
- Tasin, Isfafuzzaman, Tansin Ullah Nabil, Sanjida Islam, and Riasat Khan. “Diabetes Prediction Using Machine Learning and Explainable AI Techniques.” Healthcare Technology Letters 10, no. 1–2 (December 14, 2022): 1–10. https://doi.org/10.1049/htl2.12039.
- “Patient Registration - RCM Glossary | MD Clarity.” Accessed March 01, 2024. https://www.mdclarity.com/glossary/patient-registration.

# Thank you!