Behavioral Drivers of Diabetes

Classifying Diabetes Risk from Lifestyle Data

Rossana Antonella Sacco

Master's Degree in Artificial Intelligence and Data Engineering

Data Mining and Machine Learning project

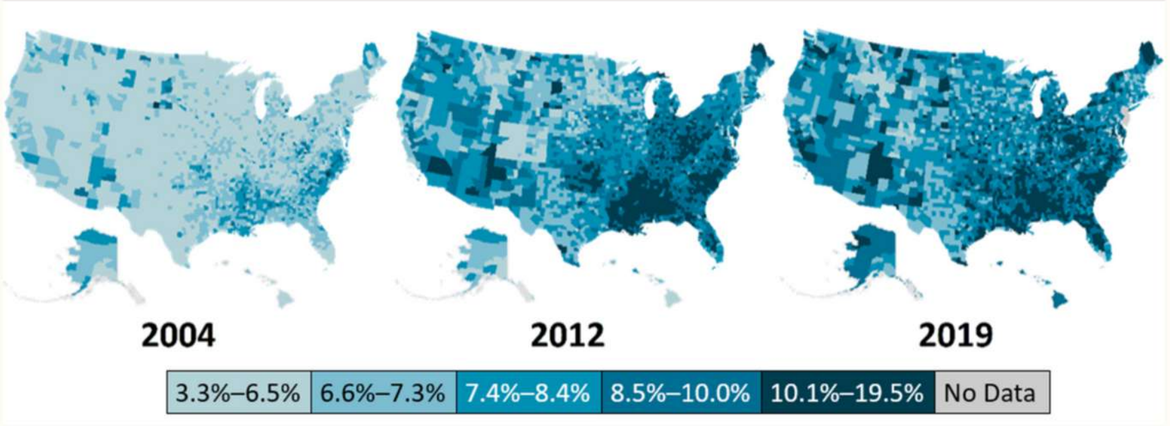
The problem

Diabetes is rapidly rising in the U.S:

- **+28%** between 2001–2020
- Today: **38 million Americans** (1 in 10 adults)
- Nearly 1 in 5 cases **undiagnosed,** leading to heart disease, kidney failure and amputations
- **High medical costs** limit access to standard treatment

Lifestyle is a key factor:

- Healthy diet & exercise → up to **40% lower risk**
- Smoking, alcohol, inactivity **aggravate outcomes**



Using Data Mining and Machine Learning techniques this research aims to build a classification model capable to identify **diabetes cases**, promote **early detection** and healthy behaviors.

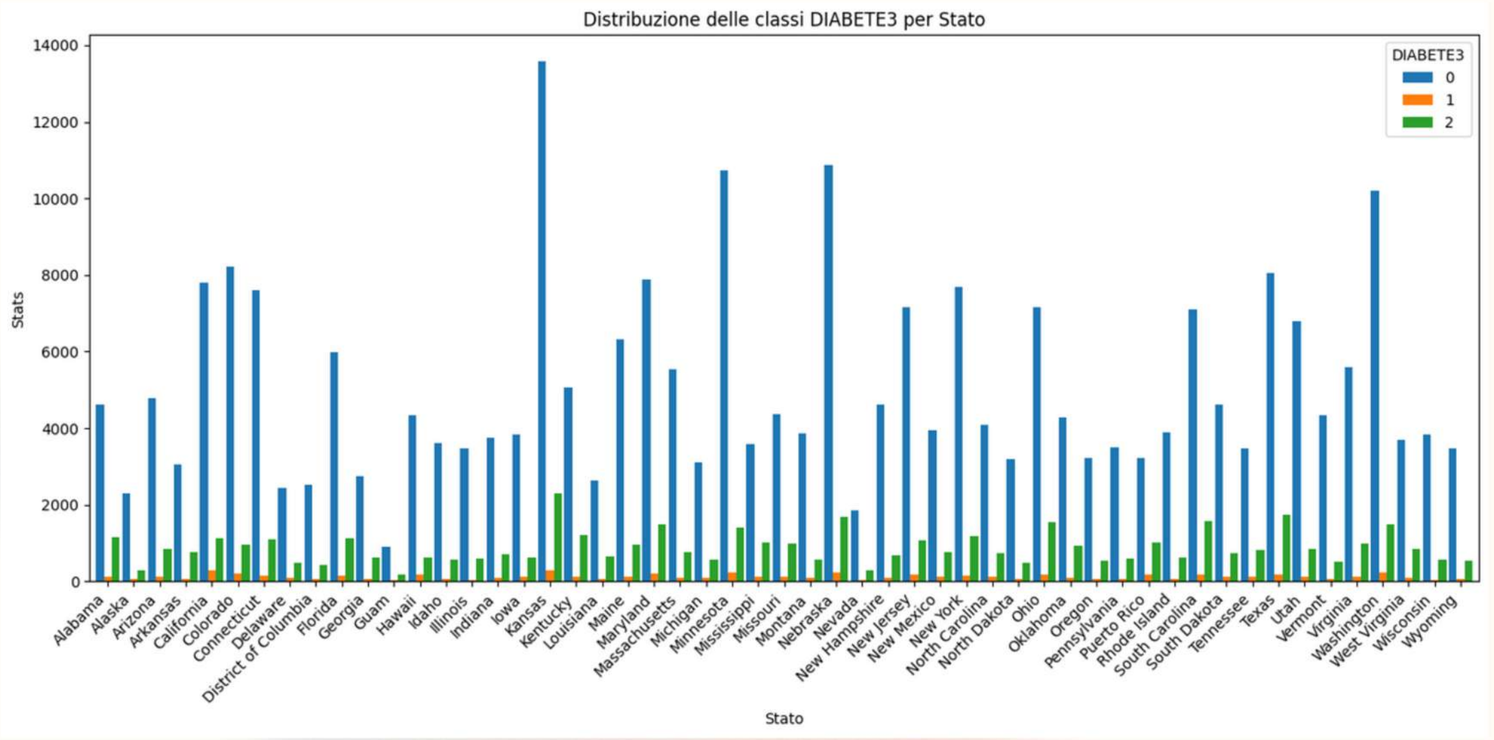
Dataset

Overview

Behavioral Risk Factor Surveillance System (BRFSS) – a dataset based on a health survey conducted by the U.S. Centers for Disease Control and Prevention (CDC)

- **430,000+** records
- **300+** features

The dataset includes a wide range of variables covering demographics, overall health status, chronic conditions, access to healthcare, and lifestyle habits.



Target variable: **DIABETE3**

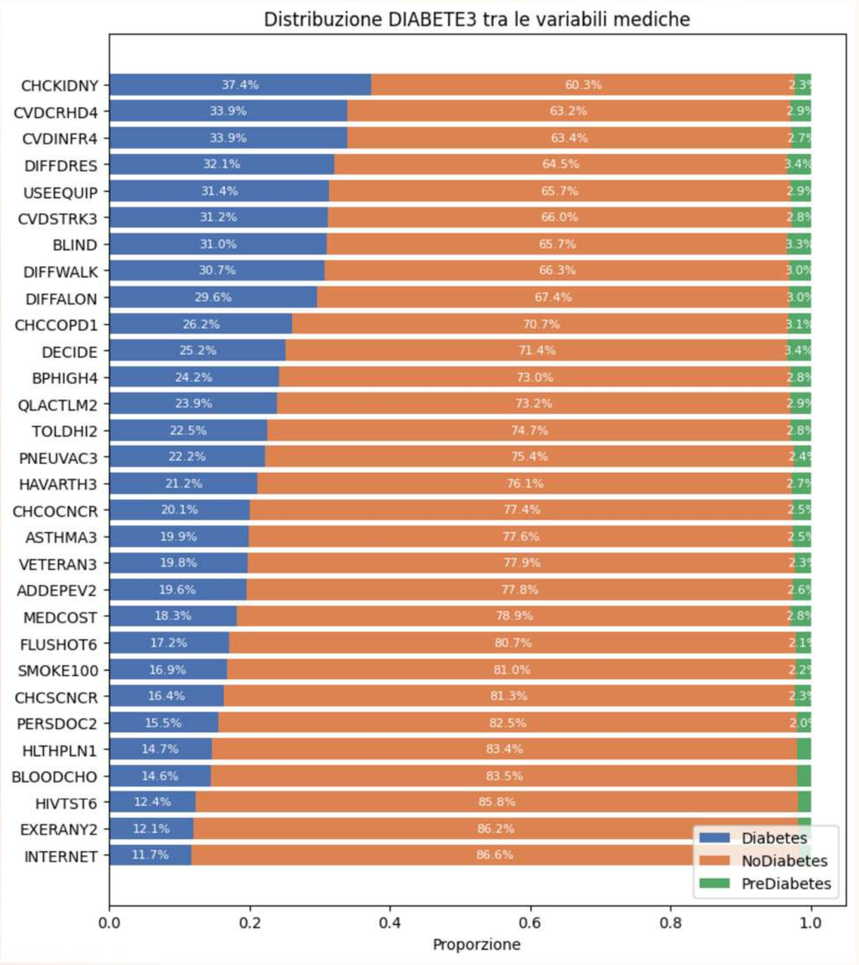
Dataset

Data cleaning

- Removed **admin/survey**-derived fields
- Dropped features with **>30%** missing values
- **Redundant** features were consolidated.
- **Recode** missing values (e.g. 9, 99 → NaN)
- Convert **Yes/No → 1/0**
- Align scale directions in **ascending orde**r for ordinal features
- Ambiguous or inconsistently scaled variables were transformed into **standardized and interpretable formats**

Target Variable

- Filtered out **unknown values** (codes 7, 9)
- Missing values within the majority class were removed to ensure **data integrity**.

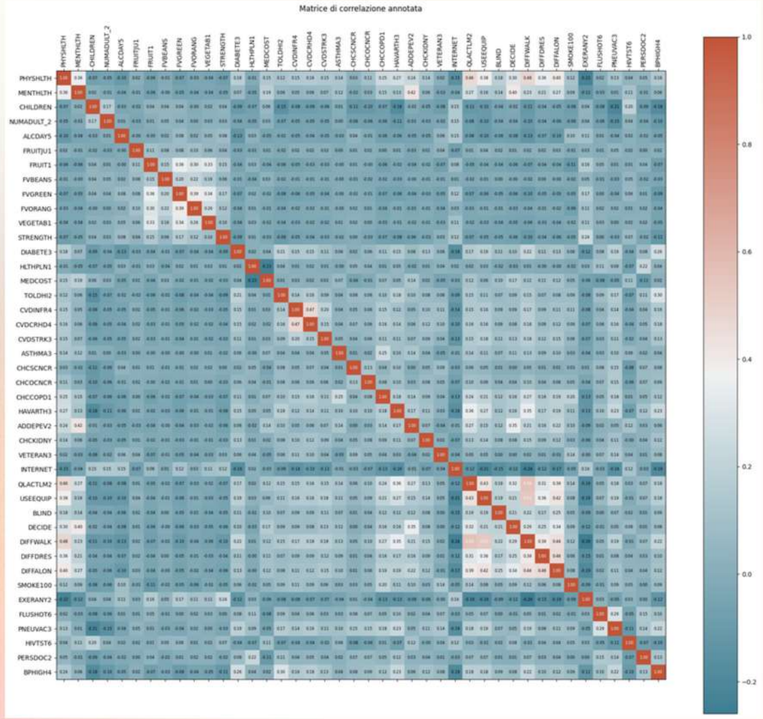


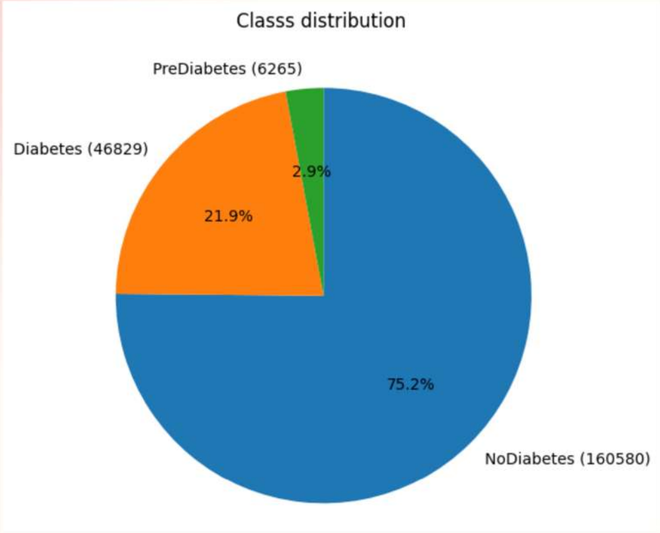
Dataset

Exploratory Data Analysis

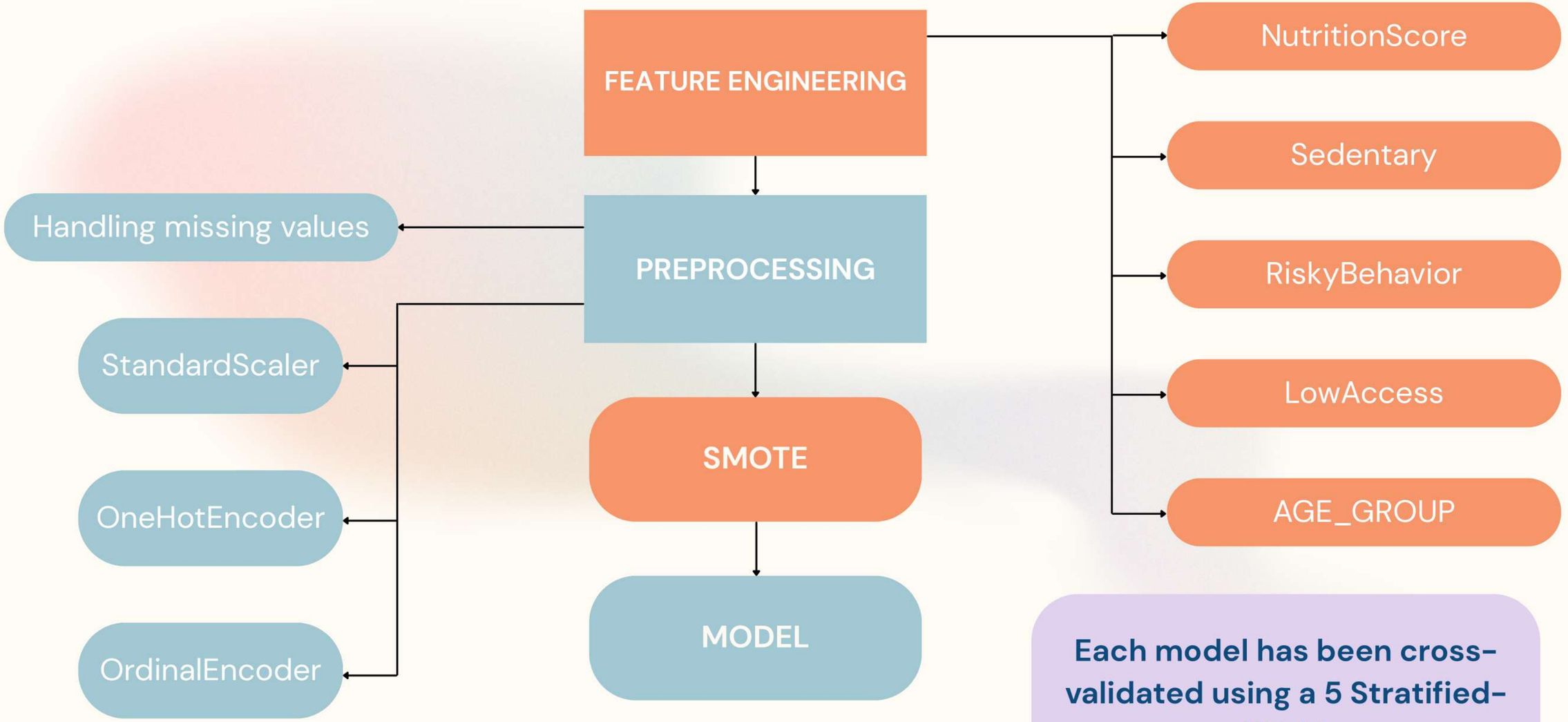
- Removed obvious **data entry errors** (e.g. “77 children”)
- Applied **winsorization** (1st–99th percentile)

- **Correlation Study**
  - Removed features with corr < **0.05** with the target variable
- **Categorical variables**
  - Sparse categories merged to “Other”
  - Chi-square test
- **Class Balance study**





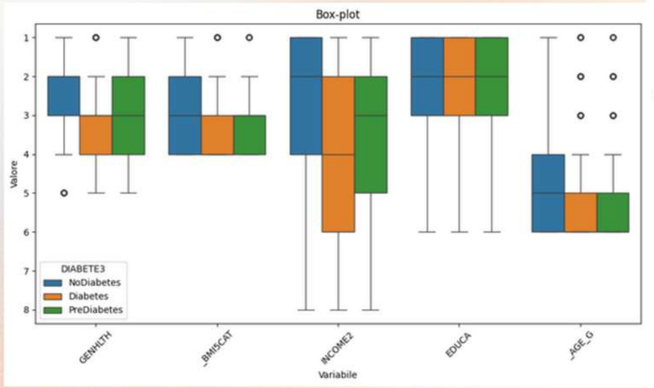
Preprocessing and pipeline building

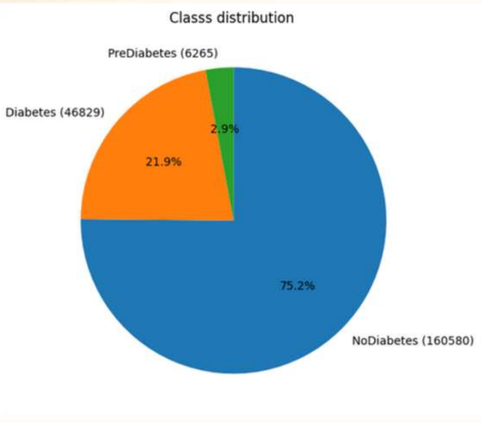


Each model has been cross-validated using a 5 Stratified-KFold

Multiclass Classification

Identifying **PreDiabetes** would support early intervention and prevention.





Heavy Class Imbalance

- Prediabetes is only the **3%** of data
  - Too few samples to learn class's caratheristics
- **Hybrid behavior**: prediabetes overlaps both other classes
  - Difficult to separate

To **balance** the dataset, several techniques were applied:

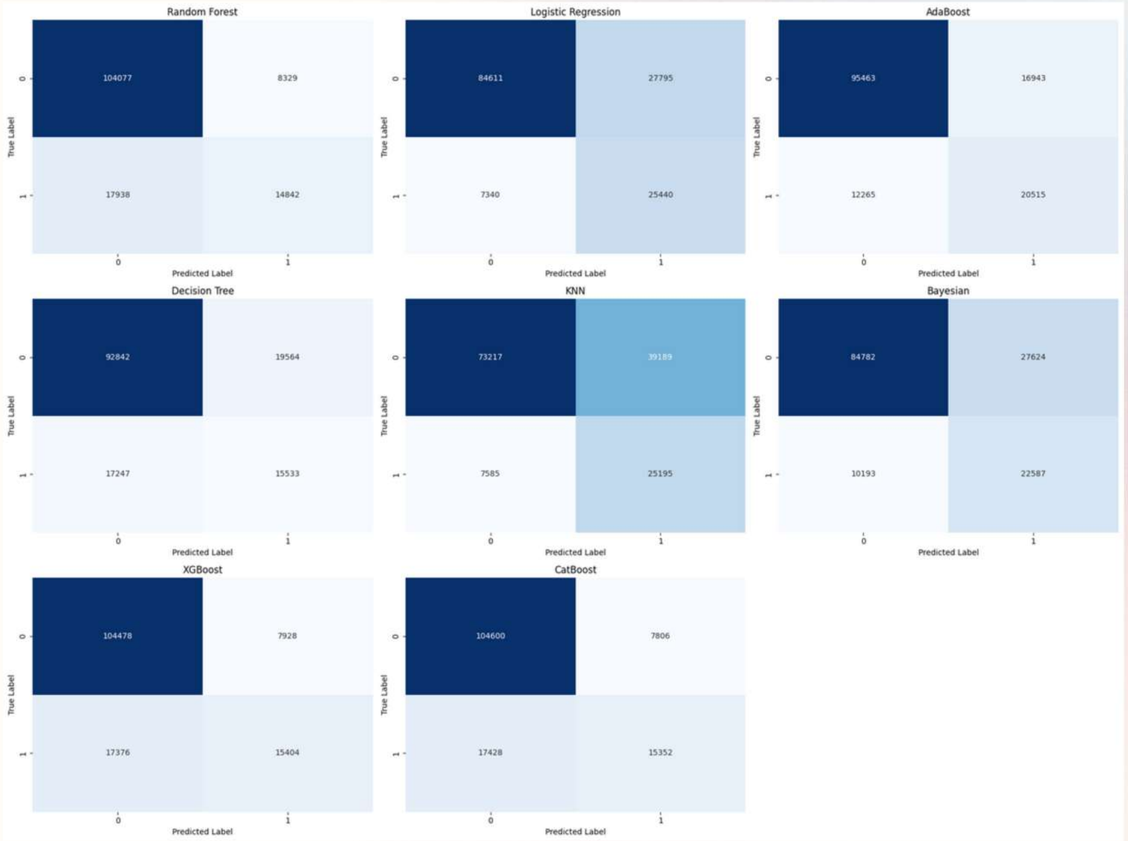
- **Oversampling** (SMOTE):
  - Not effective due to limited variability in real PreDiabetes cases
- **Undersampling**:
  - Required removing too much valid data; insufficient separation
- **Hybrid Sampling**:
  - Balanced all classes to ~40K records → still unclear boundaries

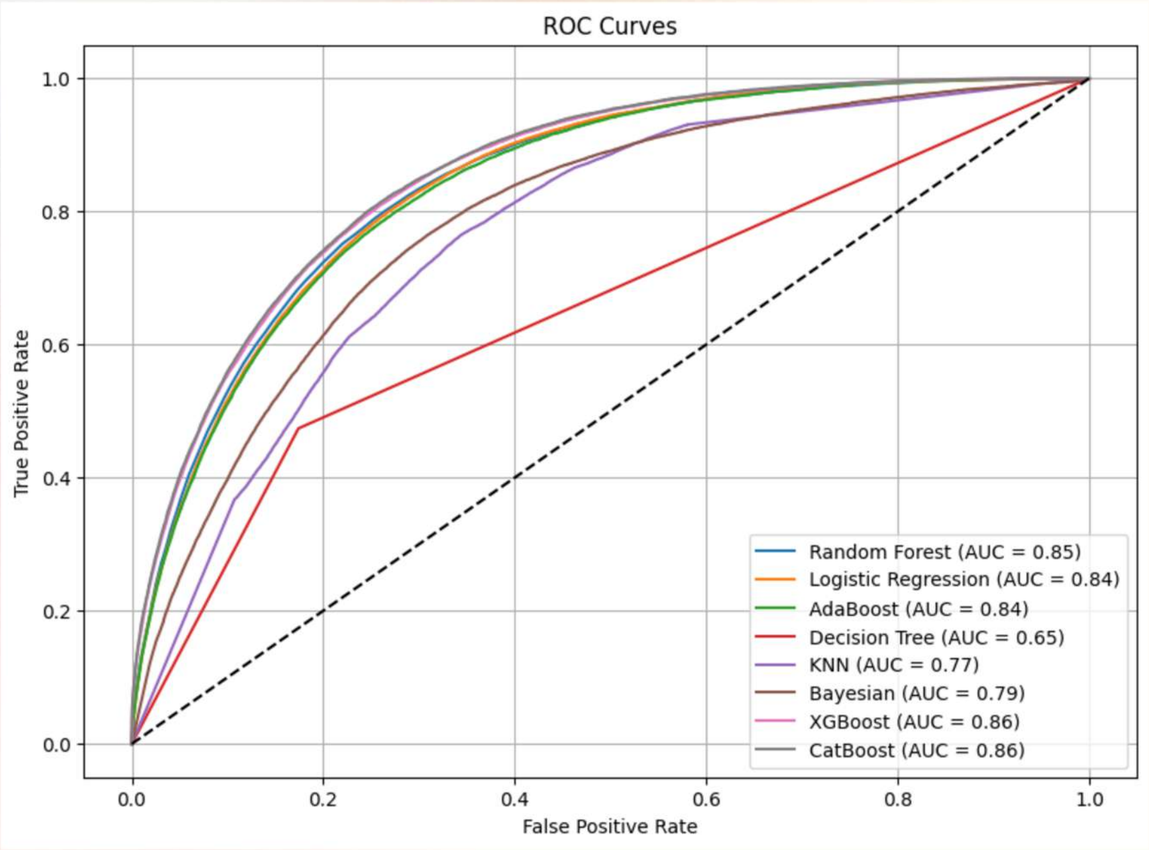
	Precision	Recall	F1Score
NoDiabetes	~ 0.9	< 0.55	< 0.48
PreDiabetes	< 0.05	~0.2	< 0.1
Diabetes	< 0.5	~0.5	~0.5

Despite different approaches, performance remained consistently **low**:

Balanced Accuracy	Accuracy	Precision	Recall	F1Score
< 0.53	~ 0.68	< 0.55	< 0.53	< 0.48

Binary Classification





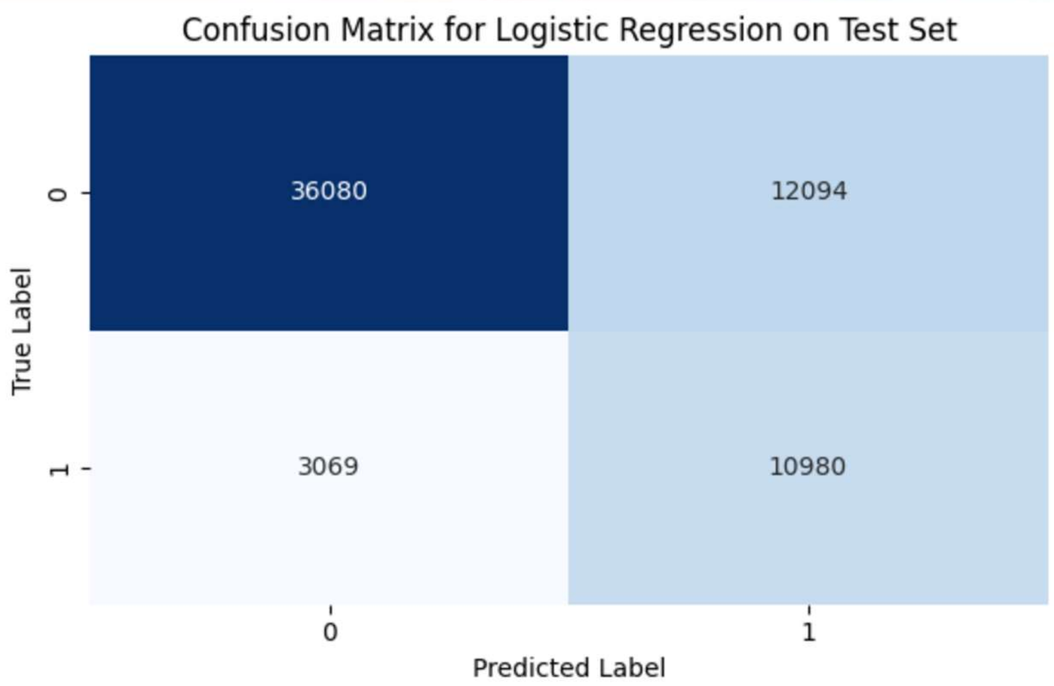
Given the nature of the problem, **recall** is the most critical metric.

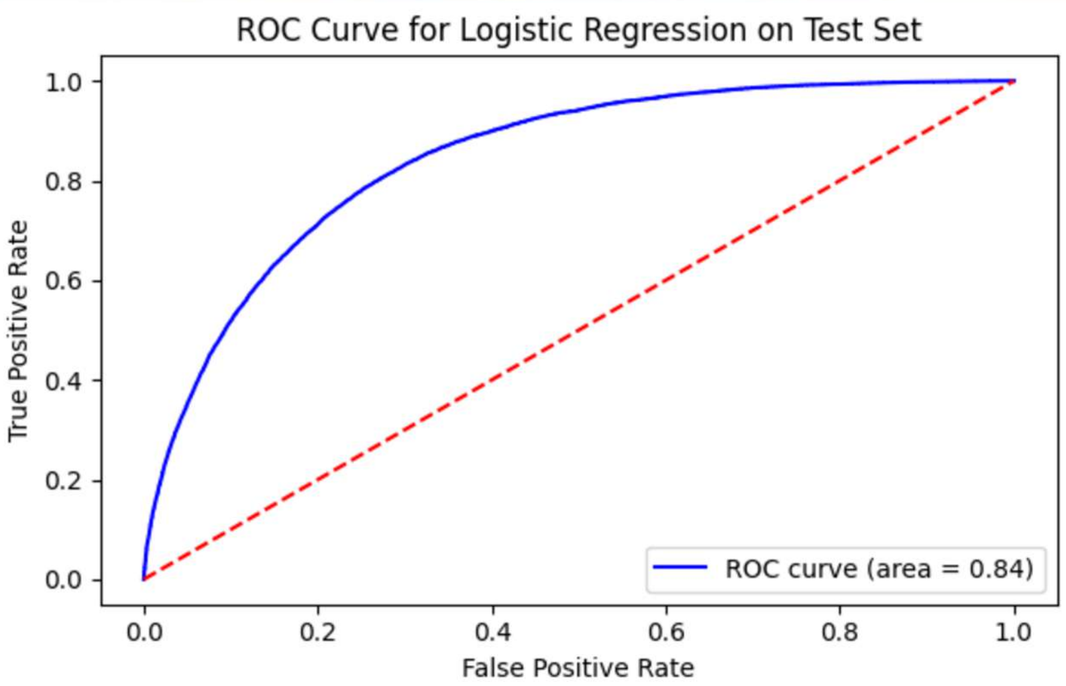
Binary Classification

	Accuracy	Precision	Recall	F1Score	ROC AUC
Random Forest	0.819080	0.640542	0.452776	0.530536	0.846516
Logistic Regression	0.758000	0.477881	0.776083	0.591525	0.844316
AdaBoost	0.798824	0.547680	0.625839	0.584157	0.841173
Decision Tree	0.746456	0.442573	0.473856	0.457681	0.649933
KNN	0.677834	0.391324	0.768609	0.518608	0.766200
Bayesian	0.739527	0.449842	0.689048	0.544324	0.789219
XGBoost	0.825713	0.660209	0.469921	0.549045	0.857348
CatBoost	0.826195	0.662924	0.468334	0.548893	0.859398

Binary Classification

Model Evaluation





	Accuracy	Precision	Recall	F1Score	ROC AUC
Logistic Regression	0.76	0.48	0.78	0.60	0.77

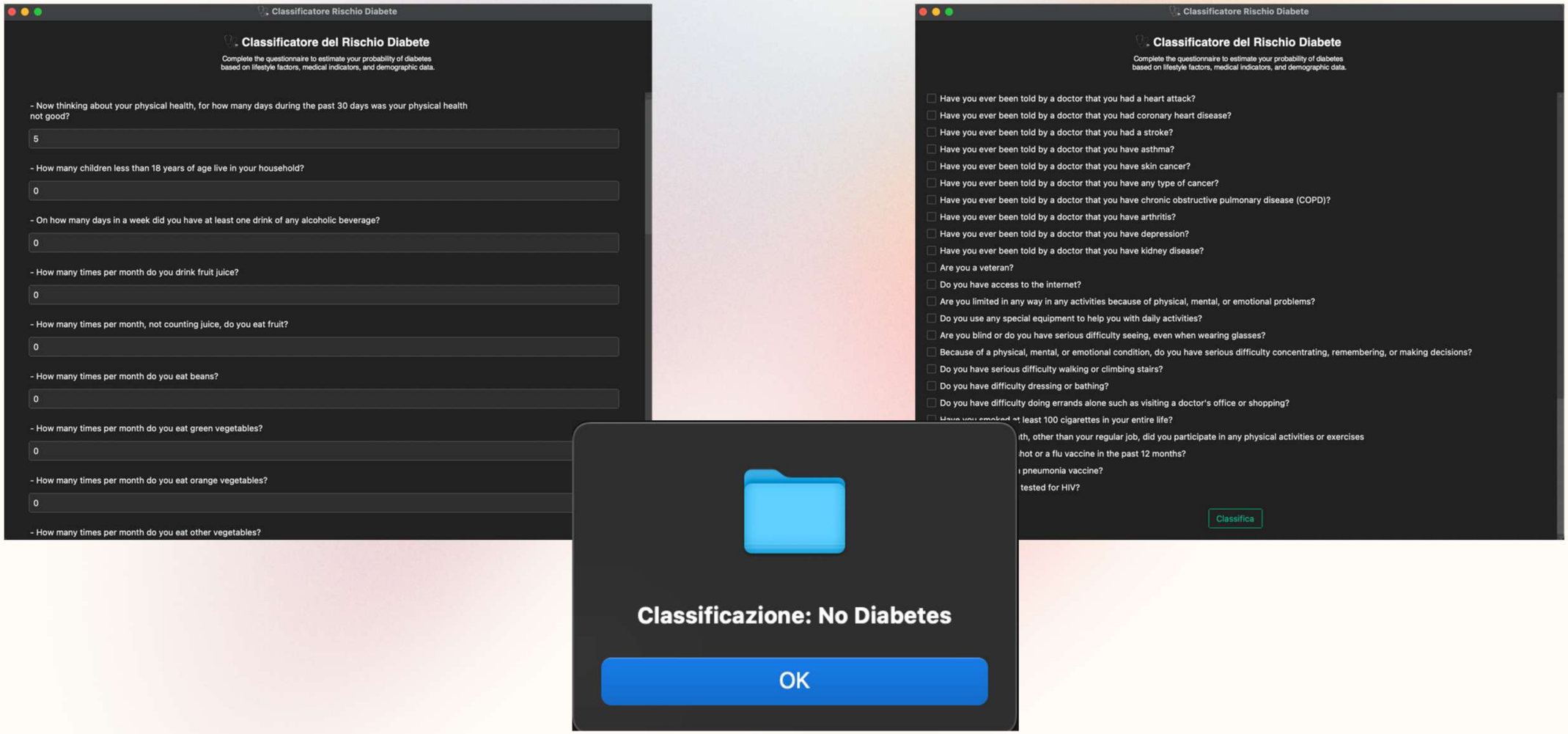
Binary Classification

Feature Importance

	Feature	Importance
0	_BMI5CAT_1.0	1.175089
1	AGE_GROUP_Young	0.994092
2	BPHIGH4	0.763472
3	_BMI5CAT_4.0	0.701080
4	GENHLTH	0.681720
5	CHOLCHK	0.631822
6	PNEUVAC3	0.609406
7	MEDCOST	0.593065
8	TOLDHI2	0.555843
9	_BMI5CAT_2.0	0.538180
10	_RACE_1.0	0.495054

Binary Classification

User Interface



Bibliography

- **Dataset**: https://www.kaggle.com/datasets/cdc/behavioral-risk-factor-surveillance-system?resource=download&select=2015.csv
- **Codebook**: https://www.cdc.gov/brfss/annual_data/2015/pdf/codebook15_llcp.pdf
- **CDC's Study:** https://www.cdc.gov/pcd/issues/2019/19_0109.htm
- **Lancaster University's study**: https://www.researchgate.net/publication/377829779_Predictions_of_diabetes_through_machine_learning_models_based_on_the_health_indicators_dataset

THANK YOU
for your attention

Rossana Antonella Sacco