UNIVERSITÀ DI PISA

Dipartimento di Ingegneria dell'Informazione

Master's Degree in Artificial Intelligence and Data Engineering

Behavioral Drivers of Diabetes: Classifying Diabetes Risk from Lifestyle Data

Candidate:

Rossana Antonella Sacco

Project GitHub Repository: github.com/RossSacco/DmmlPJ

Academic Year 2024–2025

Contents

Chapter 1

Introduction

Diabetes represents a rapidly escalating public health challenge in the United States. Between 2001–2004 and 2017–2020, the age-adjusted prevalence of diagnosed and total diabetes among U.S. adults rose from approximately 10.3% to 13.2% — an increase of nearly **28%** over two decades [1] [3].

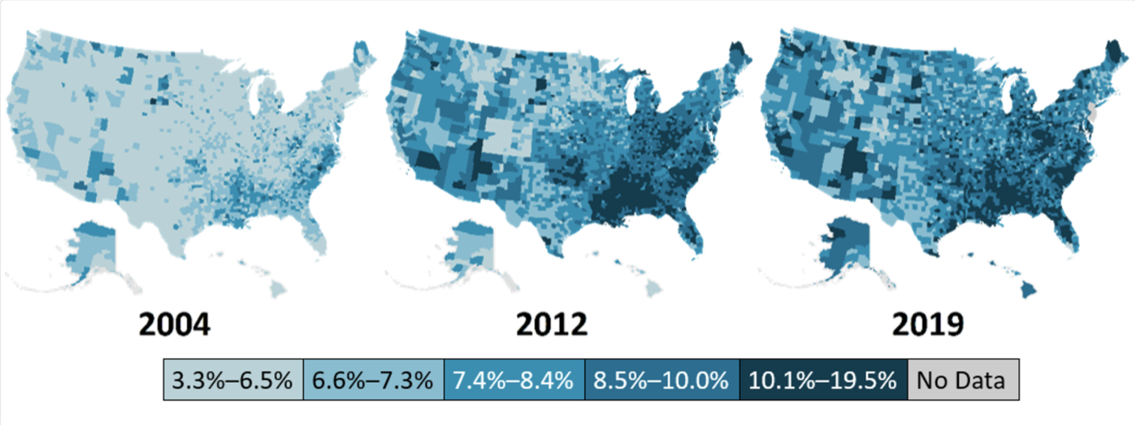


Figure 1.1: From CDC's National Diabetes Statistic Report

This upward trend translates into roughly **38 million Americans** living with diabetes, or approximately one in every ten adults, with nearly one fifth of them undiagnosed and therefore untreated. Such figures underscore not only the growing burden of disease but also the missed opportunities for **early detection and intervention**.

The human and economic consequences of this surge are profound. **Diabetes** is a leading driver of cardiovascular disease, end stage renal failure, neuropathy and lower limb amputations, and it imposes an annual cost to the U.S. health system that exceeds **hundreds of billions of dollars** [1]. However, despite the availability of effective therapies, many Americans face significant barriers to obtaining standard diabetes care, due to both direct medical costs and broader social-structural constraints [2].

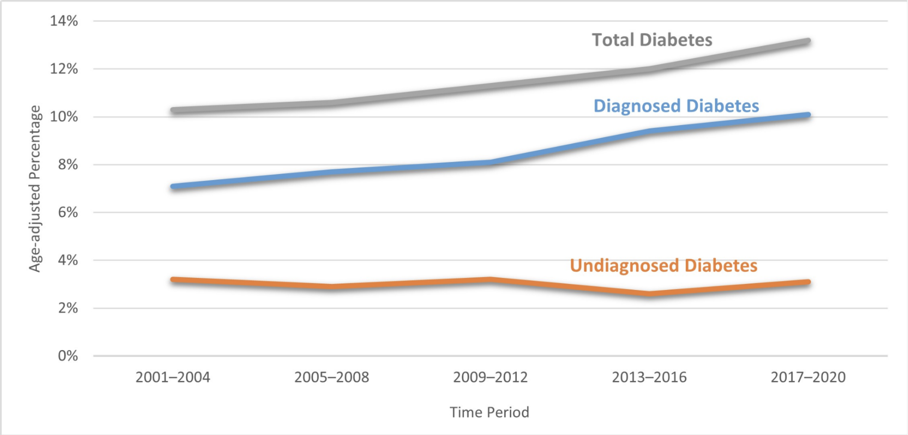


Figure 1.2: Trends in age-adjusted prevalence of diagnosed diabetes, undiagnosed diabetes, and total diabetes among adults aged 18 years or older, United States, 1999–2016.

Lifestyle behaviors are central to both the pathogenesis and prevention of diabetes. Prospective studies have consistently shown that diets high in whole grains and dietary fiber reduce diabetes risk by **18% − 40%**, while replacing even a single daily sugar sweetened beverage with water can reduce the risk by over 25%.

Physical inactivity, smoking, and excessive alcohol consumption further compound risk: adults adhering to a **healthy lifestyle** profile with regular exercise, abstinence from smoking, moderate or no alcohol use and balanced diet, exhibit a substantially lower incidence of diabetic outcomes [3]. Conversely, rising obesity rates, driven by caloric excess, sedentary occupations, and fast food availability, remain the single greatest **modifiable risk factor**, accounting for the majority of preventable diabetes cases. **The United States' environment** and socioeconomic disparities amplify these trends.

Evidence supports that even modest improvements in diet quality, increases in physical activity, and enhanced access to preventive and self management services can significantly reduce diabetes incidence and mitigate its complications.

Early detection of diabetes and the promotion of **healthy lifestyle habits** are therefore of paramount importance. This project goal is to develop a **machine learning model** that, using behavioral data, risk factors and demographic information, can accurately distinguish between individuals with and without diabetes. Using **Data Mining and Machine Learning techniques** this research aims to identify the most significant determinants of diabetes risk, thereby informing targeted prevention and intervention strategies.

Chapter 2

Dataset

The dataset **'Behavioral Risk Factor Surveillance System (BRFSS)'** originates from a comprehensive **health-behavior survey** which captures self-reported demographic attributes, lifestyle habits, clinical measurements and derived indicators for a large number of respondents, conducted by the **U.S. Centers for Disease Control and Prevention**. The key variable, **DIABETE3**, encodes diabetes status , distinguishing individuals with diabetes, prediabetes or no diabetes. The official dataset **codebook** was consulted to interpret the meaning and encoding of each feature [4].

2.1 Data Overview

The dataset comprises over 430 000 individual records and more than 300 features, spanning multiple domains:

- **Record identifiers and sampling information**: state FIPS code, interview date, number of phone, etc.

- **Demographic characteristics**: age, sex, race/ethnicity, education level, and household income.

- **Health status indicators**: self-rated general and mental health, number of days of poor physical or mental health, and body mass index categories.

- **Chronic condition flags**: diagnoses and history of diabetes, hypertension, heart attack, stroke, asthma, cancer, chronic obstructive pulmonary disease (COPD), and arthritis.

- **Health care access and utilization**: health insurance coverage, cost related care postponement, routine check-ups, cholesterol screening, and immunizations.

- **Lifestyle and behavioral factors**: tobacco use, alcohol consumption, fruit and vegetable intake, sugar sweetened beverage consumption, physical activity frequency, and strength training.

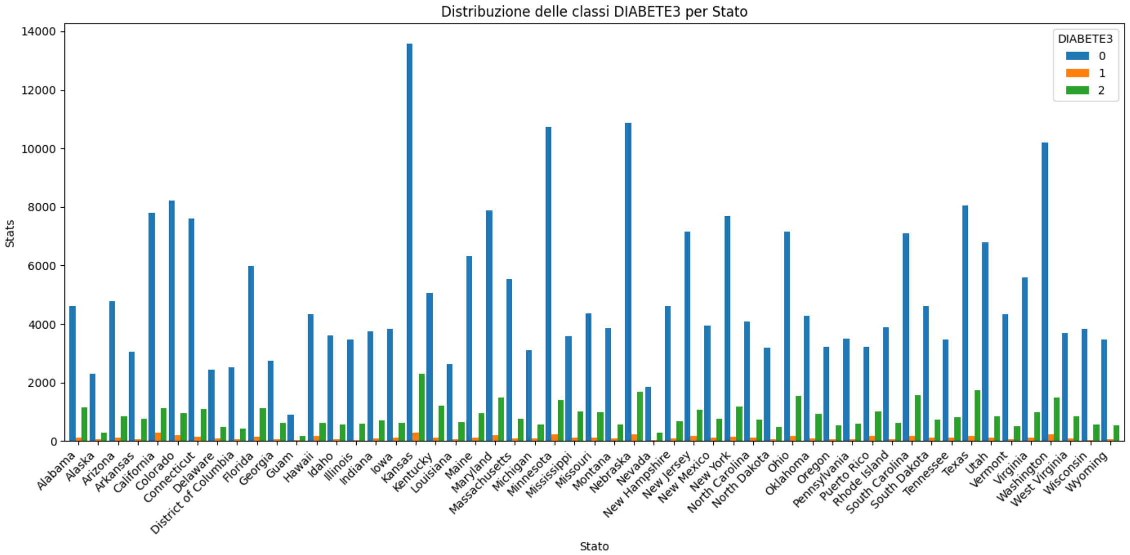


Figure 2.1: Distribution of the target variable by state

Variables are encoded using numeric codes. Data types include **continuous measures** (e.g. physical health), **ordinal categories** (e.g. education level), **nominal factors** (e.g. employment status), and **binary indicators** (e.g. asthma diagnosis). These resources enable in-depth analysis of demographic, clinical, behavioral and social determinants of **diabetes risk**.

2.2 Data Cleaning

Before any meaningful analysis or predictive modeling can take place, it is essential to transform the raw survey responses into a coherent, reliable dataset. In its original form, the **BRFSS** data contain mixed coding schemes, placeholder values for nonresponse, and a multitude of administrative or derived fields that do not directly inform diabetes risk. Left unaddressed, these inconsistencies can introduce **noise** and **bias**.

Administrative and **survey-derived** fields (e.g. `_RFHLTH`, `_RFBMI5`, `_CHLDCNT`) were excluded to preserve only the original variables directly relevant to this study and to eliminate superfluous or redundant information. Columns exhibiting over **30% missingness** were completely removed to minimize noise and improve model stability. Furthermore, the two separate adult-count variables (`NUMADULT_1` and `NUMADULT_2`), which recorded the same household information for two different kinds of telephone lines, were consolidated into a single `NUMADULT` feature, since the distinction between lines was considered irrelevant.

2.2.1 Feature Standardization

Target Variable

Rows with `DIABETE3` codes of 7 or 9 (unknown/refused) were **excluded** from analysis. The remaining integer codes were then mapped to three distinct categories: **"Diabetes," "NoDiabetes," and "PreDiabetes"** in order to provide a **multi-class** target for modeling.

Missing Values Codes

Standard survey **codes for nonresponse** (7, 9, 77, 99, etc.) were uniformly mapped as **NaN** (after this transformation, it was verified that no variable had more than 30% of missing values). For some variables, such as `PHYSHLTH, MENTHLTH` and `CHILDREN`, specific recoding rules (e.g. 88 → 0 to denote "none") were applied before mapping nonresponses as missing. This approach ensures a **consistent** representation of missing data across all formats.

Metrics Transformation

Frequency metrics with ambiguous scales (for instance, `ALCDAY5` which originally mixed daily/weekly counts) were recast into **interpretable measures** (e.g. average days of alcohol consumption per week). Binary variables originally coded as 1 = yes and 2 = no (e.g. `MEDCOST`) were transformed respectively into **1/0** integers. Data types were defined: **nominal attributes** (e.g. `EMPLOY1`) were converted to string categories, **ordinal variables** (e.g. `EDUCA`) were converted to numeric values with explicit ordering (any scales originally defined in descending order were inverted to ascending order through value mapping), and **continuous measures** (e.g. `STRENGTH`) were trasformed into float.

By **standardizing** missing value codes, recasting ambiguous frequency measures into interpretable metrics and removing features with excessive sparsity, it ensure that each variable genuinely reflects the behavior, demographic trait, or health status it was **intended to capture**. Moreover, reconciling ordinal scales and converting nominal fields to categorical types preserves the natural ordering of responses and prevents inadvertent numerical assumptions.

2.3 Exploratory Data Analysis

EDA provides a snapshot of a dataset's key characteristics while revealing interesting relationships between features and the target.

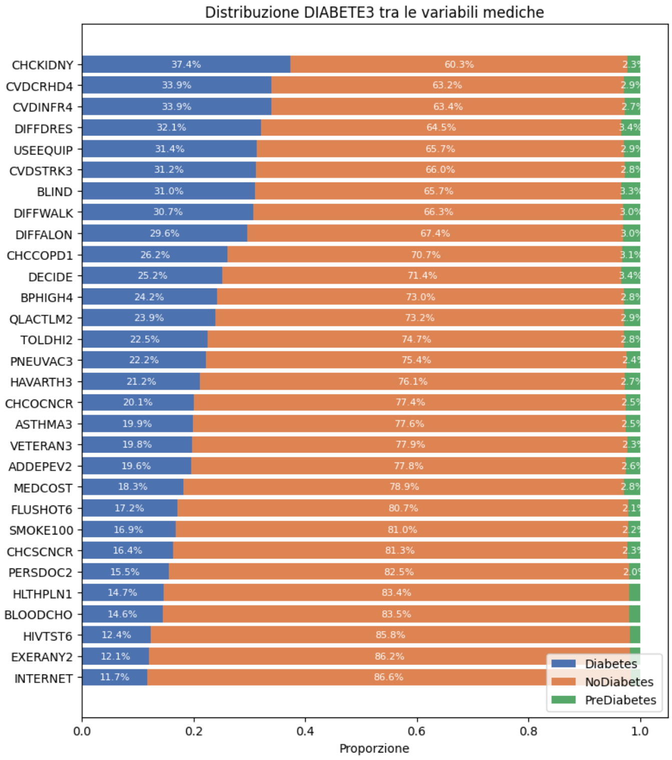


Figure 2.2: Distribution of classes among respondents reporting "yes" for additional medical conditions.

2.3.1 Data Integrity

To ensure a reliable data for modeling, **continuous features** were inspected for anomalous values and extreme observations using boxplots, which revealed **pronounced right skewness** and isolated data-entry errors (e.g., "77 children"). Winsorization at the **1st and 99th** percentiles constrained outliers without discarding them outright, thereby preserving variability while reducing their influence on summary statistics. Clearly erroneous rows were subsequently removed to ensure that all downstream analyses reflect plausible clinical scenarios.

2.3.2 Correlation study

Correlations were calculated among all continuous and binary numeric features and the target. Predictors with an absolute correlation below **0.05** with the target were excluded, with the exception of those designed for later grouping in preprocessing, while there were none of the remaining features exceeded a correlation of 0.40.

2.3.3 Cathegorical variables

Frequencies for each categorical feature were tabulated to detect sparse categories, which were merged into a single **"Other"** level when below a frequency threshold. **Chi-square tests** of independence were then performed between each feature and the target; only variables achieving statistical significance were maintained.

2.3.4 Data Distribution

The target variable is **heavily imbalanced**, with the majority class ("NoDiabetes") comprising over 75% of all cases. To ensure a fully observed dataset, any records in this dominant class that contained missing values in any attribute were removed. This filtering slightly reduced the overall sample size but maintained data integrity without worsening the existing class imbalance.

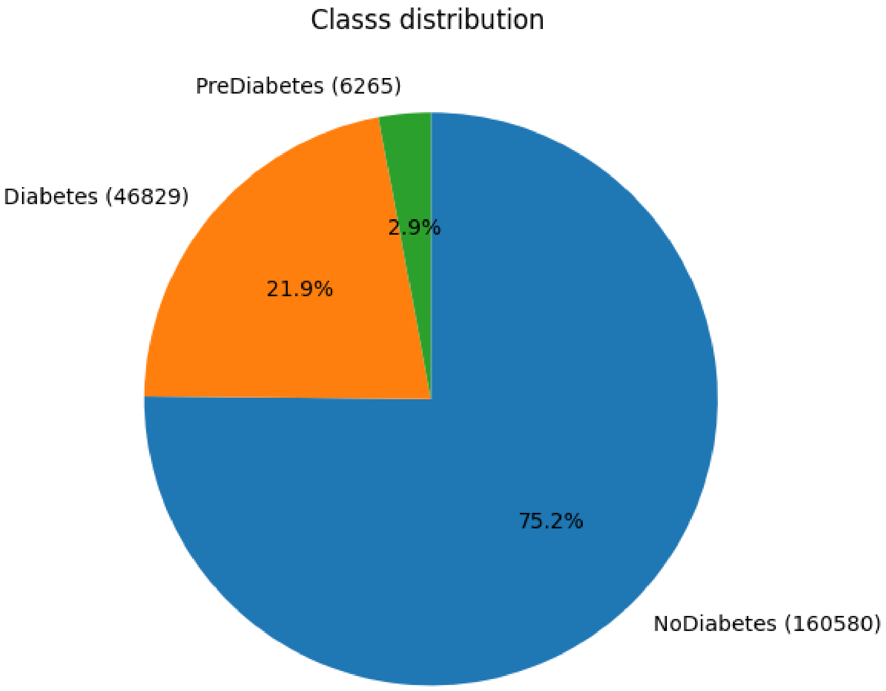


Figure 2.3: Class Distribution

Chapter 3

Diabetes Classification

3.1 Preprocessing phase

To prepare the dataset for machine learning, a series of preprocessing steps were applied.

Handling missing values

For numerical features, missing values are imputed using the median, whereas for all other feature types, they are imputed using the most frequent value.

Normalization

Numerical features were standardized using StandardScaler() to ensure that differences in scale do not bias the model.

Encoding

Since the Nominal categorical features did not exhibit high cardinality, each was transformed using OneHotEncoder. For the ordinal categorical features, OrdinalEncoder was applied.

Feature engineering

New features were derived from selected original ones to better capture key behavioral, nutritional, and demographic risk factors for diabetes, thereby improving the predictive performance and interpretability of models.

- **NutritionScore**: Computes a nutritional score as a weighted sum of fruit and vegetable eating frequency indicators (`FRUIT1`, `FVBEANS`, `FVGREEN`, `FVORANG`, `VEGETAB1`), using predefined weights for each item.
- **Sedentary**: Defines a sedentariness indicator.

$$\texttt{Sedentary} \;=\; \big(1 - \texttt{EXERANY2}\big) \;+\; \mathbf{1}_{\{\texttt{\_PACAT1}=4\}}.$$

Here,

- `EXERANY2` = 1 indicates the person did any physical activity in the past month, and `EXERANY2` = 0 indicates they did not.
- `_PACAT1` = 4 identifies respondents classified as "inactive" in their overall activity category.
- $\mathbf{1}_{\{\texttt{\_PACAT1}=4\}}$ is an indicator that equals 1 exactly when `_PACAT1` is 4 (inactive), and 0 otherwise.

Thus, `Sedentary` can take values 0, 1, or 2 according to this table:

EXERANY2	_PACAT1	Sedentary
1 (did exercise)	2 (somewhat active)	0
0 (no exercise)	2 (somewhat active)	1
0 (no exercise)	4 (inactive)	2
1 (did exercise)	4 (inactive)	1

- **RiskyBehavior**: Generates a risk behavior score by summing:
  - `SMOKE100` (smoking status), is equal to 1 if the respondent has smoked at least 100 sigarets during his life.
  - A binary indicator equal to 1 if `ALCDAY5` (alcohol consumption frequency) is greater or equal than 3 (times per week).
- **LowAccess**: Creates a "low access to care" variable as the sum of:

$$\texttt{LowAccess} \;=\; \big(1 - \texttt{PERSDOC2}\big) \;+\; \texttt{MEDCOST}.$$

Here,

- `PERSDOC2` = 1 indicates the respondent has a personal doctor; `PERSDOC2` = 0 means he do not.
- `MEDCOST` = 1 indicates the respondent reported difficulty paying for medical care; `MEDCOST` = 0 means no such difficulty.

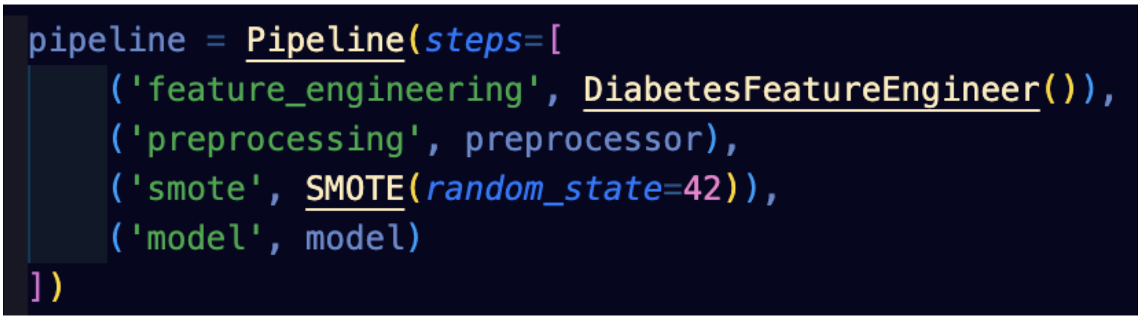
Therefore, `LowAccess` can take values 0, 1, or 2 according to this table:

PERSDOC2	MEDCOST	LowAccess
1 (has doctor)	0 (no cost barrier)	0
0 (no doctor)	0 (no cost barrier)	1
1 (has doctor)	1 (cost barrier)	1
0 (no doctor)	1 (cost barrier)	2

- **AGE_GROUP**: Bins the raw age `_AGE_G` into three ordered categories (*Young*, *Middle*, *Older*), producing a categorical age variable.

3.2 Pipeline

The classification model was built using a pipeline to ensure consistent application of feature engineering, preprocessing and model training.



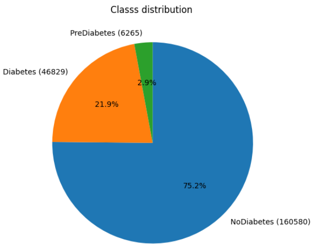
```
pipeline = Pipeline(steps=[
    ('feature_engineering', DiabetesFeatureEngineer()),
    ('preprocessing', preprocessor),
    ('smote', SMOTE(random_state=42)),
    ('model', model)
])
```

Figure 3.1: Pipeline Code

3.3 Multiclass Classification

The initial goal of this research was to build a model that could distinguish among three categories: **Diabetes, NoDiabetes and PreDiabetes**. Identifying the "prediabetes" group would have been especially valuable, because these are individuals who have not yet developed full-blown diabetes but whose behaviors and clinical measures place them on a trajectory toward it. From a preventive-medicine standpoint, correctly flagging prediabetic cases allows for early lifestyle interventions that can halt or delay disease onset.

However, only about 3% of all samples fell into the prediabetes category. This extremely low prevalence poses two main problems:



- **Insufficient examples to learn a reliable pattern:** With so few prediabetic cases, the model cannot see enough variability to learn what truly differentiates "prediabetes" from "no diabetes" or "diabetes."
- **Hybrid behavior of prediabetic individuals:** in many features prediabetic subjects often exhibit values that lie somewhere between the "no diabetes" and "diabetes" groups. This makes it even harder for the classifier to carve out a clear, compact region in feature space that consistently corresponds to prediabetes.

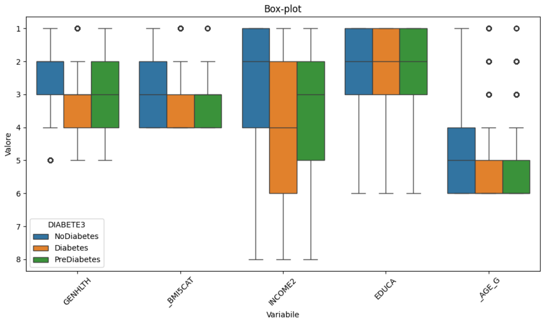


Figure 3.2: Box plot to compare class values

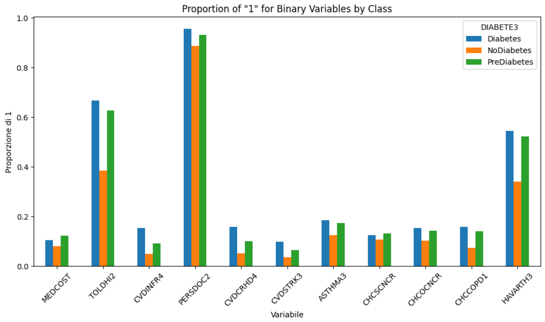


Figure 3.3: Distribution of classes among respondents reporting "yes" in medical question.

3.3.1 Sampling Strategies and Results

To mitigate the extreme imbalance in the "prediabetes" class, several **sampling techniques** were evaluated. In each case, the goal was to increase the representation of the minority class without compromising the integrity of the data distribution. The following approaches were attempted:

- **Oversampling:** Synthetic oversampling (SMOTE) was applied to augment the minority class. However, due to the very limited number of true "prediabetes" examples, the generated samples failed to capture the genuine variability of this subgroup.
- **Undersampling:** Randomly reducing the majority classes ("no diabetes" and "diabetes") to balance with the minority class would have required discarding a large portion of valid data. Even after aggressive undersampling, the remaining majority-class samples did not sufficiently separate from the hybrid behavior of prediabetic individuals.
- **Hybrid Sampling:** First undersampling the majority classes, then oversampling the minority. It was tested by adjusting all three classes to approximately 40,000 observations each. Despite equalizing class counts, this approach still failed to produce a coherent decision boundary for "prediabetes.".

Table 3.1: Model Performance Metrics using SMOTE, obtained with 5-fold cross validation

Model	Balanced Accuracy	Accuracy	Precision	Recall	F1 Score
Random Forest	0.463593	0.793308	0.529370	0.463593	0.467522
Logistic Regression	0.530903	0.635620	0.492329	0.530903	0.465319
AdaBoost	0.495586	0.740870	0.468451	0.495586	0.474805
Decision Tree	0.433242	0.702782	0.429197	0.433242	0.430839
KNN	0.462141	0.568912	0.436176	0.462141	0.407837
Bayesian	0.490094	0.597625	0.465661	0.490094	0.435631
XGBoost	0.476640	0.806556	0.557720	0.476640	0.481548

Because none of these sampling methods produced a reliable classifier for the multiclass problem, and given the inherent hybrid nature of prediabetic feature profiles, was concluded that multiclass modeling was **not feasible** with the available sample sizes. Consequently, the analysis focus was shifted to a **binary classification** paradigm (presence vs absence of diabetes), which offers a more stable and generalizable solution under the current data constraints.

3.4 Binary Classification

Given the large number of samples in the diabetes and Nodiabetes groups, all prediabetes samples were excluded for the binary classification, ensuring that only **pure examples** of the two remaining classes were used.

3.4.1 Model Comparison

To evaluate the performances of various machine learning models for predicting the presence of diabetes, several model were **cross-validated** using a 5-fold **Stratified-KFold**.

Table 3.2: Binary Classification Performance Metrics

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**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

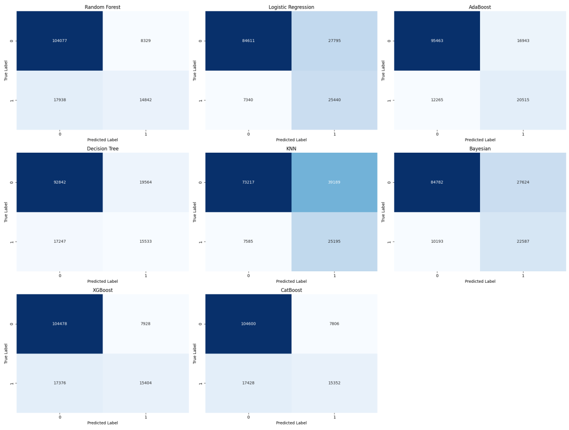


Figure 3.4: Box plot to compare class values

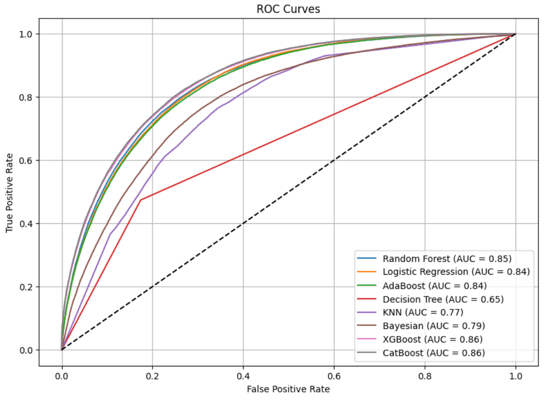


Figure 3.5: Distribution of classes among respondents reporting "yes" in medical question.

Overall, **Logistic Regression emerges** as the strongest model for this application despite not having the highest overall accuracy (0.76 compared to CatBoost's 0.83). In a preventive-focused diabetes study, correctly identifying positive cases (i.e., maximizing recall) is key: Logistic Regression achieves a **recall** of approximately 78%, whereas CatBoost's recall (47%) is far lower. Although KNN also attains a high recall (77%), its performance on accuracy, precision, and ROC AUC is substantially weaker, making it less reliable for balanced decision-making. Thus, the combination of strong recall with competitive performance on other metrics makes Logistic Regression the **preferred model** in this context.

3.4.2 Model Evaluation

Fine Tuning using GridSearchCV

To improve the performance of Logistic Regression model, a hyperparameter **tuning process** was conducted using GridsearchCV with a 5-fold Stratified-KFold. The following hyperparameters were optimized:

- **C:** Inverse regularization strength.
- **Solver:** Specifies which optimization algorithm the model uses to fit the coefficients.
- **Class_weight:** Controls how the algorithm weights each class during training.

The grid search was performed with two score metrics: Recall and ROC AUC. The best hyperparameter were selected based on the highest recall score, as predicting the presence of Diabetes was the priority. The best parameters found were:

- **C:** 0.01.
- **Solver:** liblinear.
- **Class_weight:** balanced.

Using best parameters, the best model results were:

- **Best_Recall:** 0.78
- **Best_AUC:** 0.84

Then, the fine-tuned Logistic Regression Model was then retrained on the entire training set and evaluated using the test set, with the following results:

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Linear Regression	0.756312	0.47586	0.78155	0.591547	0.765251

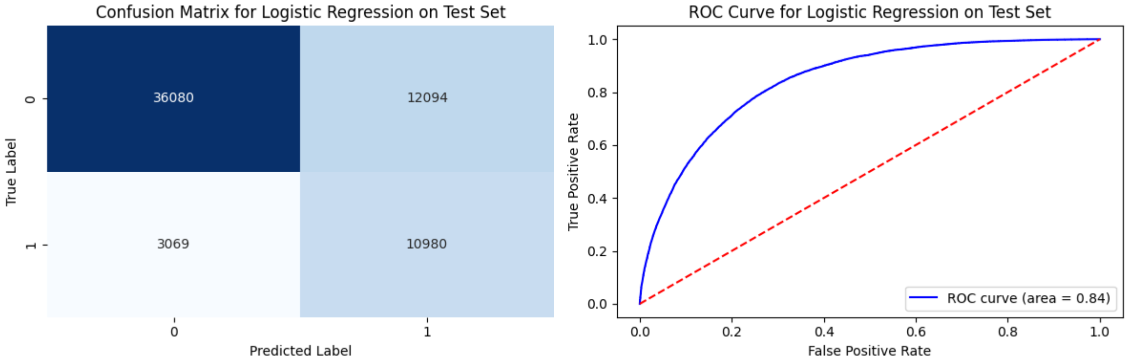
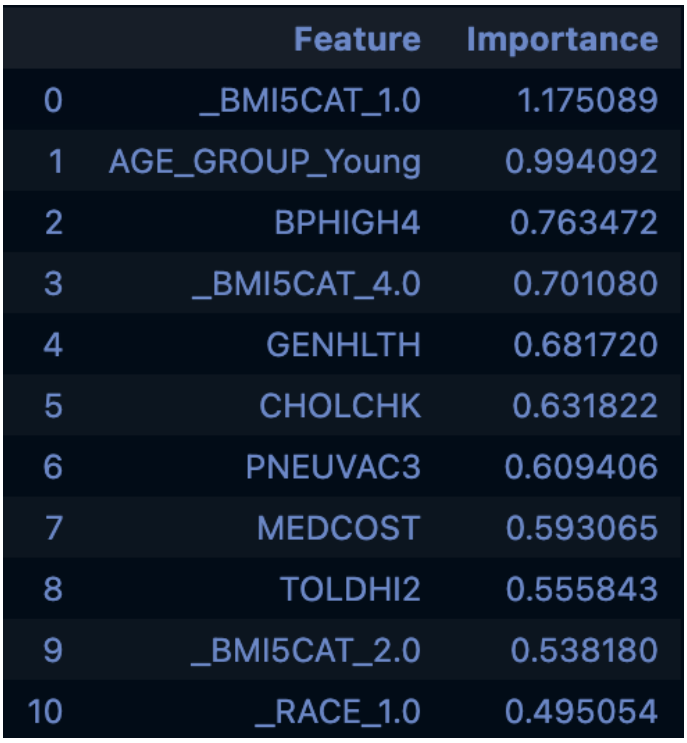


Figure 3.6: Fine-Tuned model performances

3.4.3 Feature Importance

To gain insights into the most influential features for predicting diabetes, the feature importance of the fine-tuned Logistic Regression model were extracted. The following lists the top 10 features contribuiting to the model's decisions in descending order:



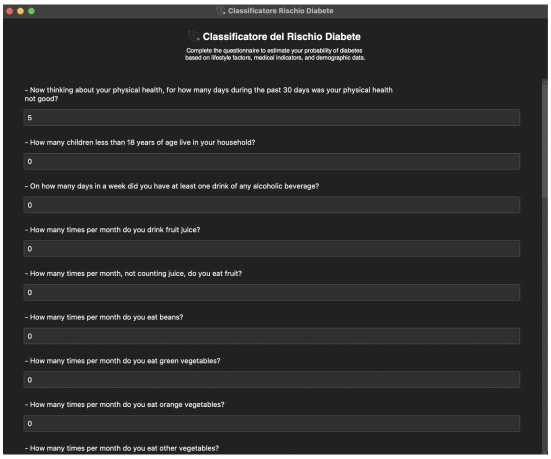
	Feature	Importance
0	_BMI5CAT_1.0	1.175089
1	AGE_GROUP_Young	0.994092
2	BPHIGH4	0.763472
3	_BMI5CAT_4.0	0.701080
4	GENHLTH	0.681720
5	CHOLCHK	0.631822
6	PNEUVAC3	0.609406
7	MEDCOST	0.593065
8	TOLDHI2	0.555843
9	_BMI5CAT_2.0	0.538180
10	_RACE_1.0	0.495054

Figure 3.7: Top 10 Important features

Chapter 4

User Interface

A **user interface** was developed to let individuals enter personal, clinical, and behavioral data and instantly receive a diabetes risk classification. The form guides users through fields such as age, diet, and activity, validating inputs in real time. Upon submission, a clear binary outcome (diabetes vs. no diabetes) is displayed along. This interface translates the underlying machine learning model into an accessible tool for both lay users and healthcare professionals.


Classificatore Rischio Diabete

Classificatore del Rischio Diabete

Complete the questionnaire to estimate your probability of diabetes
based on lifestyle factors, medical indicators, and demographic data.

- Now thinking about your physical health, for how many days during the past 30 days was your physical health not good?

5

- How many children less than 18 years of age live in your household?

0

- On how many days in a week did you have at least one drink of any alcoholic beverage?

0

- How many times per month do you drink fruit juice?

0

- How many times per month, not counting juice, do you eat fruit?

0

- How many times per month do you eat beans?

0

- How many times per month do you eat green vegetables?

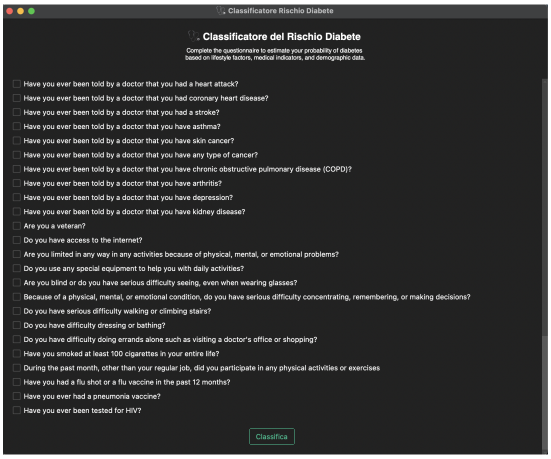
0

- How many times per month do you eat orange vegetables?

0

- How many times per month do you eat other vegetables?

Figure 4.1


Classificatore Rischio Diabete

Classificatore del Rischio Diabete

Complete the questionnaire to estimate your probability of diabetes
based on lifestyle factors, medical indicators, and demographic data.

☐ Have you ever been told by a doctor that you had a heart attack?
☐ Have you ever been told by a doctor that you had coronary heart disease?
☐ Have you ever been told by a doctor that you had a stroke?
☐ Have you ever been told by a doctor that you have asthma?
☐ Have you ever been told by a doctor that you have skin cancer?
☐ Have you ever been told by a doctor that you have any type of cancer?
☐ Have you ever been told by a doctor that you have chronic obstructive pulmonary disease (COPD)?
☐ Have you ever been told by a doctor that you have arthritis?
☐ Have you ever been told by a doctor that you have depression?
☐ Have you ever been told by a doctor that you have kidney disease?
☐ Are you a veteran?
☐ Do you have access to the internet?
☐ Are you limited in any way in any activities because of physical, mental, or emotional problems?
☐ Do you use any special equipment to help you with daily activities?
☐ Are you blind or do you have serious difficulty seeing, even when wearing glasses?
☐ Because of a physical, mental, or emotional condition, do you have serious difficulty concentrating, remembering, or making decisions?
☐ Do you have serious difficulty walking or climbing stairs?
☐ Do you have difficulty dressing or bathing?
☐ Do you have difficulty doing errands alone such as visiting a doctor's office or shopping?
☐ Have you smoked at least 100 cigarettes in your entire life?
☐ During the past month, other than your regular job, did you participate in any physical activities or exercises
☐ Have you had a flu shot or a flu vaccine in the past 12 months?
☐ Have you ever had a pneumonia vaccine?
☐ Have you ever been tested for HIV?

Figure 4.2

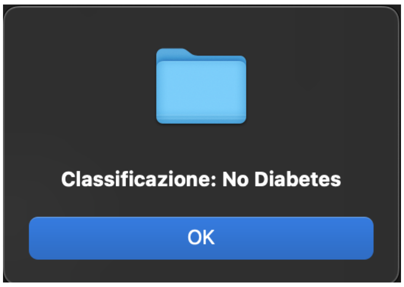

Classificazione: No Diabetes

Figure 4.3

Chapter 5

Conclusion

The **binary logistic regression model** built in this study on a large set of behavioral, demographic, and clinical variables, achieved a **recall** of approximately 78% and an **AUC** of 0.8424 when distinguishing diabetes from non-diabetes (accuracy: 75.6%). Other studies on BRFSS data have often begun by selecting a small number of **'ad hoc'** risk indicators through univarite testing or domain knowledge [5] [6]. Although their model achieved higher overall accuracy, they did not prioritize sensitivity (recall) and used a predefined subset of variables. In contrast, this study started from a much **larger pool** of features. This comprehensive approach allowed us to assess the influence of a wider range of factors on diabetes risk and optimize recall, a critical metric in preventive screening.

Similarly, the **CDC's PCD study** on 2014 BRFSS data [6] built several machine learning models, then compared performance across algorithms. Their best model achieved 82.4% accuracy and an AUC of 0.7949, while the decision tree attained the highest **recall** at 51.6%. This work diverges by incorporating many additional behavioral indicators (dietary and physical activity scores, healthcare access metrics) directly into the modeling process. As a result, we achieve substantially **higher recall** (78%) without sacrificing AUC, demonstrating that a larger feature set, together with SMOTE can produce a effective screening tool when the goal is early detection of true positive cases.

In summary, by prioritizing recall over raw accuracy and embracing a wide range of predictive features rather than selecting a small, predefined subset, this model provides a practical and interpretable tool for early **diabetes risk detection**.

Bibliography

[1] NDSR, Available at: https://diabetesresearch.org/wp-content/uploads/2022/05/national-diabetes-statistics-report-2020.pdf, diabetes Statistic Report.

[2] DiabetesJournal, Available at: https://diabetesjournals.org/care/article/24/2/268/24025/Barriers-to-Providing-Diabetes-Care-in-Community.

[3] PMC, Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC6125024/.

[4] codebook, Available at: https://www.cdc.gov/brfss/annual_data/2015/pdf/codebook15_llcp.pdf.

[5] L. U. study, Available at: https://www.researchgate.net/publication/377829779_Predictions_of_diabetes_through_machine_learning_models_based_on_the_health_indicators_dataset.

[6] C. study, Available at: https://www.cdc.gov/pcd/issues/2019/19_0109.htm.

[7] Dataset, Available at: https://www.kaggle.com/datasets/cdc/behavioral-risk-factor-surveillance-system?resource=download&select=2015.csv.