IDS homework 10 - Project Report

Predicting and Analysing Diabetes Risk Using Patient Health Data

Sofia Apri, Helena Rääsk

Repository: https://github.com/sofiaapri/IDS2025-Diabetes-Prediction

Task 2. Business understanding

Our business goals

Background

Diabetes is a chronic and potentially life-threatening disease where late detection can lead to serious complications. In everyday healthcare practice many possible risk factors must be assessed at the same time, such as age, body weight, cholesterol, family history, physical inactivity, blood pressure and glucose-related measures. Because these indicators interact and vary between people, it is not always clear which factors contribute most to risk or how to combine them into a consistent assessment. Comprehensive patient data can help both to understand key drivers and support building models that can assist risk assessment and monitoring. However, this is a student project using synthetic data and is not intended for clinical deployment.

Business goals

This project is intended to benefit individuals and healthcare professionals by providing a data-driven approach to estimating diabetes risk and predicting outcomes. The direct goals are to analyse how demographic, lifestyle and clinical indicators relate to diabetes diagnosis and stage, to identify groups of patients with distinct risk profiles, and to build predictive models for diabetes diagnosis and stage. In addition, the project aims to identify the most influential predictors and derive a simple, interpretable low/medium/high risk grouping that can also be understood by non-experts.

Business success criteria

The project is successful if it provides actionable and understandable outcomes rather than only technical results. This means that the analysis should clearly explain which factors are most associated with diabetes outcomes, the identified risk profiles should be meaningful and easy to interpret, and the produced risk grouping should align with higher observed diabetes prevalence and/or more severe stages. The evaluation should explicitly consider clinically relevant errors, especially missed cases, and the final workflow should be reproducible and well documented.

Assessing our situation

Inventory of resources

The team consists of two students with an expected workload of about 60 total hours. We use a Kaggle dataset of 100,000 synthetic patient records (≈14 MB). Development will be done in Python using Jupyter notebooks and common data-science libraries (pandas, scikit-learn, matplotlib).

Requirements, assumptions, and constraints

The key requirement is to state clear objectives, apply relevant methods, and produce sufficient results considering the project hours. We assume synthetic data is suitable for demonstrating data-mining methods, but it may not reflect real clinical distributions. We must avoid data leakage by excluding the provided precomputed diabetes risk score as a model input when predicting diagnosis or stage. A central constraint is class imbalance in diabetes stages (gestational diabetes is rare), which affects modelling.

Risks and contingencies

The main risk is obtaining overly optimistic results. We will reduce this risk by defining the model input features in advance, excluding leakage-prone variables such as the provided diabetes_risk_score, and validating performance against simple baseline models. Another risk is unstable multi-class stage prediction because some stage classes are very rare (e.g., Gestational). To address this, we will report class-wise metrics in addition to overall scores and also evaluate a variant where rare classes are handled separately. If time becomes limited, we will prioritize a solid baseline, clear evaluation, and interpretability rather than extensive hyperparameter tuning.

Terminology

Diabetes diagnosis refers to the binary target (0/1). Diabetes stage refers to the multi-class label (e.g., No Diabetes, Pre-Diabetes, Type 1, Type 2, Gestational). Recall measures how many true diabetes cases are correctly detected, while precision measures how many predicted cases are correct. ROC AUC summarizes ranking performance for binary prediction across thresholds.

Costs and benefits

Since this is an academic project, costs are limited to development time and standard compute on a personal computer. The benefit is a reproducible analysis and models that demonstrate how diabetes-related risk can be analysed and predicted.

Our data-mining goals

Data-mining goals

We will perform exploratory data analysis to quantify relationships between patient features and diabetes outcomes, and apply clustering to identify distinct risk profiles. We will train supervised models to predict diabetes diagnosis and diabetes stage from patient features, while excluding the precomputed risk score from model inputs to avoid leakage. We will identify influential predictors using interpretable model outputs (e.g., coefficients or feature importance) and derive a simple low/medium/high risk score.

Data-mining success criteria

For diagnosis prediction, we target ROC AUC ≥ 0.80 on a held-out test set and aim for recall ≥ 0.85 for the positive class, while monitoring false positives. For stage prediction, we will evaluate using a confusion matrix and class-wise precision and recall due to imbalance. Clustering is successful if clusters show clearly different outcome distributions and can be explained as meaningful risk profiles.

Task 3. Data understanding

Gathering data

Data requirements

To achieve the business and data-mining goals, the project needs patient-level information covering demographics, lifestyle factors, clinical indicators, glucose-related lab values and medical history. Two target variables are needed: diagnosed_diabetes (binary) and diabetes_stage (multi-class). These variables are required for assessing risk factors, identifying patient profiles, and building predictive models.

Data availability

We use a synthetic Kaggle dataset "diabetes_dataset.csv" containing 100,000 patient records and over 35 features. The dataset is complete, clean, and free of missing values or duplicates. It is available as a CSV file and can be easily accessed and processed using Python, Jupyter notebooks, and common data-science libraries (pandas, scikit-learn, matplotlib). All necessary variables are available and in usable format.
(https://www.kaggle.com/datasets/mohankrishnathalla/diabetes-health-indicators-dataset)

Selection criteria

All rows in the dataset are retained for analysis, as the data is synthetic, complete, and free of missing or duplicated entries. The only variable excluded from modelling is diabetes_risk_score, a pre-computed indicator that could cause data leakage by implicitly encoding information about the targets. In supervised learning, the two target variables (diagnosed_diabetes and diabetes_stage) are kept strictly separate from the predictor set to

ensure unbiased model training. For unsupervised methods such as clustering and PCA, both target labels and the pre-computed risk score are removed so that any discovered structure is driven solely by the underlying predictors. After these exclusions, all remaining variables are used throughout the exploratory analysis and modelling stages.

Describing data

The dataset contains 100 000 synthetic patient records and more than 35 features that describe demographic characteristics, lifestyle behaviours, medical history, and a wide set of clinical and laboratory measurements. The variables include numerical fields such as age, BMI, blood pressure, glucose values, cholesterol levels, insulin, and HbA1c, as well as categorical fields such as gender, ethnicity, smoking status, education level, income level, and employment status. All variables fall within medically plausible ranges. For example, age spans from 18 to 90 years, BMI ranges from 15 to 45 kg/m², fasting glucose from 70 to 250 mg/dL, and HbA1c from 4 to 14%. The dataset also contains two target variables: diagnosed_diabetes, which is binary, and diabetes_stage, which contains five clinically meaningful categories. Because the dataset is synthetic and preprocessed, there are no missing values or duplicated entries, and all features are clearly defined and easy to interpret.

Exploring data

Exploratory analysis shows that the dataset displays realistic variation across demographic, lifestyle, and clinical features. Age is broadly distributed across adulthood, and categorical variables such as gender, ethnicity, smoking status, and education level are well represented across their categories. Lifestyle indicators such as physical activity, alcohol consumption, sleep duration, and diet score show natural patterns. For example, physical activity is right-skewed because many individuals report low weekly activity. Clinical measurements also behave as expected. Individuals with diabetes tend to have higher fasting glucose, postprandial glucose, HbA1c, and BMI values compared to those without diabetes. The relationships between predictors and outcomes generally match medical knowledge. Higher glucose and HbA1c levels are associated with more severe diabetes stages, and healthier lifestyle indicators tend to be linked with lower diabetes risk. The distribution of the outcome variables is plausible, with roughly 20 to 25 percent diagnosed cases and a multi-class stage distribution in which Gestational diabetes is the rarest category.

Verifying data quality

The quality assessment confirms that the dataset is complete, consistent, and suitable for modelling. There are no missing values, duplicated rows, or values that fall outside realistic medical ranges. All numerical features remain within expected limits, and categorical features contain only valid category labels. As the dataset is synthetic, some relationships may be somewhat simplified compared to real clinical data, although this does not limit its use for analysis and model development. The main quality concern is the class imbalance in

the diabetes_stage variable, especially the rarity of the Gestational category, which may affect the stability of multi-class models and requires the use of class-wise evaluation metrics. Data leakage is avoided by removing the precomputed diabetes_risk_score from all modelling stages. Other than basic preprocessing steps such as encoding categorical variables and scaling numerical features, the dataset is ready for exploratory analysis and modelling.

Task 4. Project plan

Task 1

Sofia 2h. We have set up the project repository on GitHub, imported the dataset and run quick checks (missing values, duplicates, target distributions) using Python in Jupyter with pandas, tracked in Git.

Task 2

Sofia 6h, Helena 6h. We have performed exploratory analysis (descriptive statistics, distributions, group comparisons, correlations) and produced the main plots using pandas and matplotlib in Jupyter.

Task 3

Sofia 5h, Helena 7h. We will build clustering pipelines (one-hot encoding + scaling, K-means; hierarchical clustering/PCA for visualisation) using scikit-learn, then interpret clusters afterwards with diabetes outcomes, implementation will be in Jupyter with pandas/matplotlib.

Task 4

Sofia 9h, Helena 9h. We will train and evaluate models for diagnosed_diabetes and diabetes_stage (baseline logistic regression, random forest, SVM), using stratified splits and metrics such as ROC AUC and recall for diagnosis, plus confusion matrices and class-wise precision/recall for stage, implementation will use scikit-learn pipelines and pandas.

Task 5

Sofia 4h, Helena 4h. We will identify key predictors via coefficients/feature importance and derive an interpretable low/medium/high risk score, comparing it against the provided diabetes_risk_score baseline, using scikit-learn, pandas and matplotlib.

Task 6

Sofia 3h, Helena 3h. We will summarise the project and use Canva to design the poster.

Task 7

Sofia 1h, Helena 1h. We will prepare for the presentation of our project, prepare the text, assign speaking roles.