School of Engineering

PROJECT:
PREDICTING DIABETES OUTCOME FOR WOMEN

INSTRUCTOR:
CHRISTOPHE BÉCAVIN, PHD, UNIVERSITÉ CÔTE D'AZUR

MASTER'S PROGRAMS:

APPLIED MSC IN DATA ANALYTICS
APPLIED MSC IN DATA SCIENCE & ARTIFICIAL INTELLIGENCE
APPLIED MSC IN DATA ENGINEERING & ARTIFICIAL INTELLIGENCE

TEAM MEMBERS:
ALMENDRA PEREZ (DS)
RONALD LE PAPE (DE)
SHANCHUN YANG (DA)
SUJEENDRA KUMAR MARUBOINA (DS)
VANESSA GIRALDO VILLANUEVA (DA)

GITHUB LINK: HTTPS://GITHUB.COM/VANES-SA03/DIABETES-PREDICTION-ML

Paris, 30.03.2025

Document Overview

INTRODUCTION: *Understanding Diabetes and Project Motivation*

Diabetes is a chronic metabolic disorder that impairs the body's ability to regulate blood glucose levels. If left untreated, it can result in serious complications such as cardiovascular disease, kidney failure, neuropathy, and blindness. The global prevalence of diabetes has reached alarming levels: according to the International Diabetes Federation (2021), 537 million adults were living with diabetes in 2021, and this number is projected to increase to 783 million by 2045. In Taiwan specifically, more than 2.18 million people — nearly one in ten — are affected by the disease. One of the major challenges in managing diabetes is its silent progression; many individuals remain undiagnosed in the early stages due to the lack of noticeable symptoms, leading to delayed treatment and worsening health outcomes (IDF, 2021).

In recent years, machine learning has emerged as a powerful tool for early disease detection by uncovering hidden patterns in clinical data. In this context, the present project aimed to build a supervised learning pipeline capable of predicting diabetes status based on routine health measurements. The dataset used originates from a clinical study conducted at the Taipei Municipal Medical Centre between 2018 and 2022 (Chou et al., 2023). It contains anonymized medical records from 15,000 female patients aged 20 to 80 and includes eight numerical variables commonly collected in standard health check-ups: number of pregnancies, plasma glucose, diastolic blood pressure, triceps skinfold thickness, serum insulin level, body mass index (BMI), diabetes pedigree function, and age. The binary target variable indicates whether each patient was diagnosed with diabetes (1) or not (0).

To complete this project, the team adopted a structured and collaborative approach. The work was divided into key milestones, with responsibilities distributed according to each member's strengths and interests. The exploratory data analysis (EDA) was led by two members, while feature engineering, model selection, and evaluation were handled individually before being consolidated. Several classification models were trained and compared, including Logistic Regression, Decision Tree, and CatBoost. The final model was selected based on performance metrics such as accuracy, recall, and AUC. It was then integrated into a web application developed using the Flask framework.

The team made use of multiple collaborative tools to support project management. GitHub was used to version and document the codebase, OneDrive and SharePoint were used to share files and resources, and weekly meetings were held to monitor progress. Each step of the pipeline — from data cleaning to model deployment — was recorded through structured meeting minutes and shared checklists, which helped ensure that deadlines were met and tasks distributed fairly.

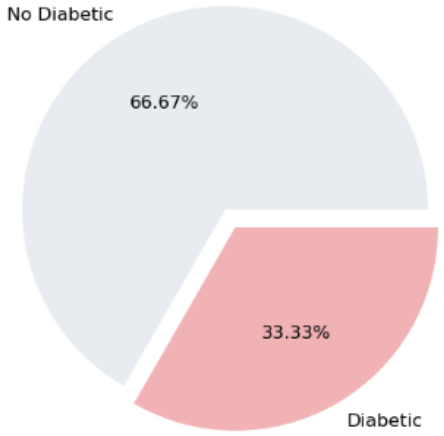
The final deliverables include a Jupyter Notebook summarizing the end-to-end technical pipeline, a written report, a web application that allows real-time predictions based on user input or CSV upload, a demo video, and a public GitHub repository containing the full code and documentation. This multidisciplinary and collaborative effort demonstrates the potential of applied data science in solving real-world healthcare problems.

METHODOLOGY

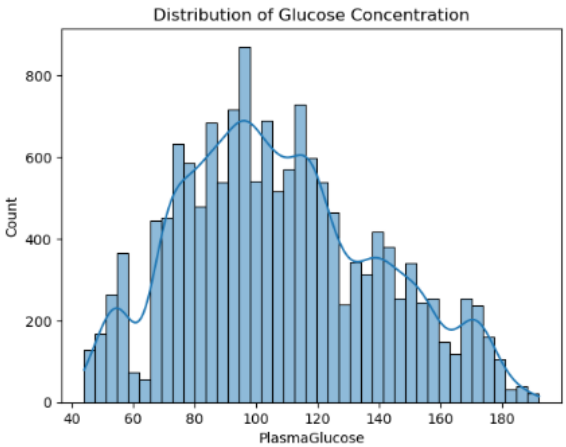
Exploratory Data Analysis

The exploratory data analysis (EDA) phase was a critical first step in understanding the structure, quality, and predictive potential of the dataset. It enabled us to formulate early hypotheses, identify patterns and inconsistencies, and guide downstream decisions regarding preprocessing, feature engineering, and model selection. The dataset consists of 15,000 rows and 9 columns: 8 numeric clinical features and 1 binary target variable (Diabetic), which indicates whether a patient was diagnosed with diabetes (1) or not (0). Each observation corresponds to a female patient aged between 20 and 80, collected from the Taipei Municipal Medical Center between 2018 and 2022.

An initial structural check confirmed that the dataset had no missing values or duplicated rows, allowing us to proceed without data imputation or deduplication. All features were numeric, which ensured compatibility with most machine learning models. We reviewed summary statistics such as mean, median, standard deviation, minimum, and maximum for each feature to evaluate data distribution and detect potential anomalies. During this inspection, we identified physiologically implausible values in the Diastolic Blood Pressure feature. Specifically, several entries contained values below 40 mmHg, which are extremely rare or critical in a clinical setting. While this variable did not contain missing values, the presence of these anomalies raised concerns about measurement accuracy or data entry errors. This issue was flagged for correction and later addressed through multiple strategies, including median replacement and K-Means-based imputation, as detailed in the feature engineering section.

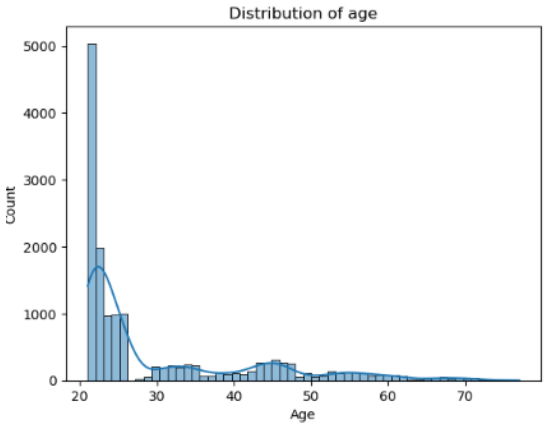


To understand the general structure of the dataset, we first examined the class distribution of the target variable. As shown in the figure below, the data exhibits a moderate class imbalance, with approximately 60% of records labeled as non-diabetic and 40% as diabetic.

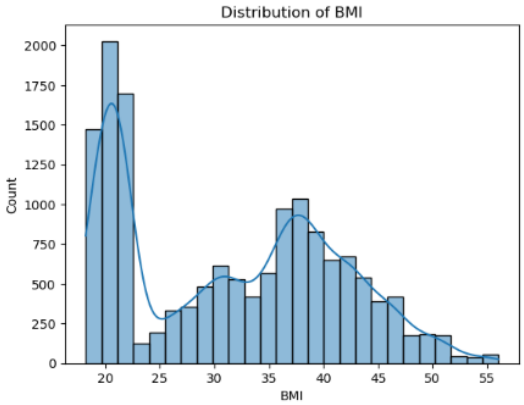


Next, we explored the distribution of individual features and their relation to diabetes status. One of the most critical variables was Plasma Glucose. The kernel density estimation plot below clearly shows that diabetic patients tend to have significantly higher glucose levels than non-diabetic ones.

Age also showed a strong correlation with diabetes. Most diabetic individuals were over 50 years old, while few cases appeared in patients younger than 30. This trend is reflected in the following distribution plot:



The Body Mass Index (BMI) followed a similar trend. Diabetic patients generally had higher BMI values, indicating a relationship between excess weight and diabetes, which is consistent with medical literature.



These three visualizations allowed us to highlight early patterns in the data and confirmed that Plasma Glucose, Age, and BMI are key features distinguishing diabetic from non-diabetic individuals. Although further analysis such as boxplots and correlation heatmaps could have been included to quantify inter-feature relationships, the current insights were sufficient to guide the next phase of preprocessing and model development.

Overall, the EDA phase confirmed that the dataset was clean, well-structured, and suitable for machine learning. Plasma Glucose, BMI, and Age emerged as the most informative features. Anomalies were identified and flagged for correction. These insights shaped the downstream pipeline and provided a solid foundation for modeling.

Feature Engineering and Selection

Based on the EDA, the following steps were taken for feature engineering and selection:

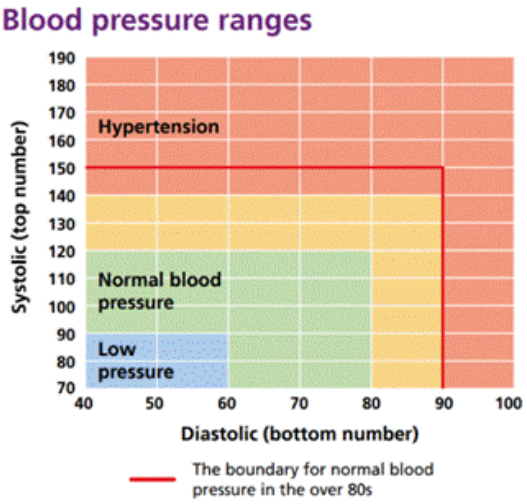
1. Handling Missing Values:

The dataset had no missing value, so no imputation was necessary.

2. Outlier Treatment:

We found that we had some very lows values for diastolic blood pressure. We considered that values below 40 were outliers as values in the range (40-60) are medically classified as low (See picture "Blood pressure range" below). They represent 221 records (1.5% of the dataset), with a percentage of 12% of diabetic person. Leaving them would have led to underestimate the prediction of becoming diabetic.

*Image source: And Precision. (n.d.). *Causes of high blood pressure*. https://andprecision.com/causes-of-high-blood-pressure/*

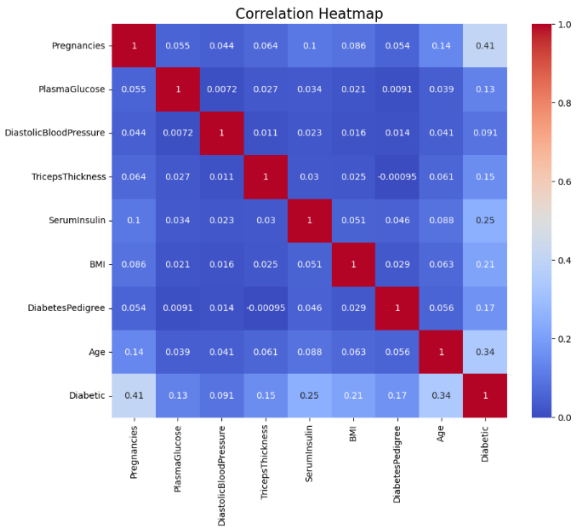


3. Feature Scaling:

All features were normalized to ensure that each feature contributes equally to the model. The library StandardScaler from ScikitLearn has been used for this purpose.

4. Feature Selection:

The features with the highest correlation to the target variable are as follows: pregnancies (0.41), age (0.34), serum insulin (0.25), BMI (0.21).



- Serum Insulin: Higher median insulin level is observed in diabetic individuals.
- BMI: Individuals with a higher BMI are more likely to have diabetes.
- Age: Individuals with diabetics are generally older.
- Pregnancies: A higher number of pregnancies is linked to an increased risk of diabetes.

We decided to keep all the features in our model because, based on our exploratory data analysis (EDA), each feature still seems to provide useful information. We chose to keep them since machine learning models can sometimes make incorrect assumptions, and having more features can help improve performance. Additionally, our dataset is not very large, so including all the features doesn't pose any computational issues. We also want to avoid introducing bias by removing features unnecessarily.

Model Selection, Comparison and Evaluation

Several machine learning models were trained and evaluated; some of these models were chosen as they were mentioned in the research paper in the original project:

1. Logistic Regression: A statistical model used to predict the probability that an instance belongs to a particular class, typically for binary classification tasks, which is suitable for the diabetes prediction (0 or 1).

2. Extra Trees: An ensemble method that combines multiple different models' predictions to improve accuracy and reduce overfitting. In our case, we combine three gradient boosting models — XGBoost, LightGBM, and CatBoost — by building a stacking classifier. Each of these models learn independently then their predictions are combined optimally.

3. Decision trees: A model that splits data by features and creates an upside-down tree to make predictions, starting at the top with a question about an important feature in the data, then branches out based on the answers.

4. Gradient boosting: An algorithm that builds an ensemble of decision trees, each one trying to correct the errors of the previous ones. It is specially designed to handle categorical features efficiently and requires minimal preprocessing.

The models were evaluated using the following metrics:

- Accuracy: The percentage of correct predictions out of all predictions made.
- Precision: The proportion of true positive predictions among all positive predictions made by the model.
- Recall: The proportion of actual positives that were correctly identified.
- F1 Score: The harmonic means of precision and recall.
- AUC-ROC: The area under the receiver operating characteristic curve, which measures the model's ability to distinguish between classes.

After the evaluation, we chose the model that achieved the best performance across the metrics mentioned above. In this case, it was CatBoost. To further improve its performance, we used hyperparameter optimization with Optuna to find the best combination of parameters.

RESULTS: *Model Performance, Interpretation and Justification*

The evaluation results for the models are summarized below:

Model	Accuracy
Logistic Regression	78.28%
CatBoost	96.24%
LGBM	95.87%

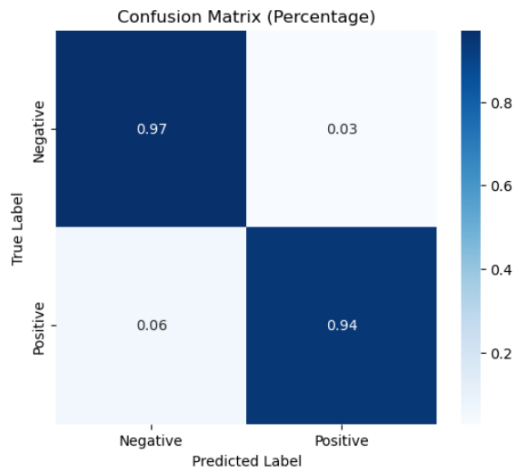
0 (non-diabetic)

Model	Precision	Recall	F1 Score
Logistic Regression	0.81	0.88	0.84
CatBoost	0.97	0.97	0.97
LGBM	0.97	0.97	0.97

1 (Diabetic)

Model	Precision	Recall	F1 Score
Logistic Regression	0.72	0.59	0.65
CatBoost	0.94	0.94	0.94
LGBM	0.94	0.93	0.94

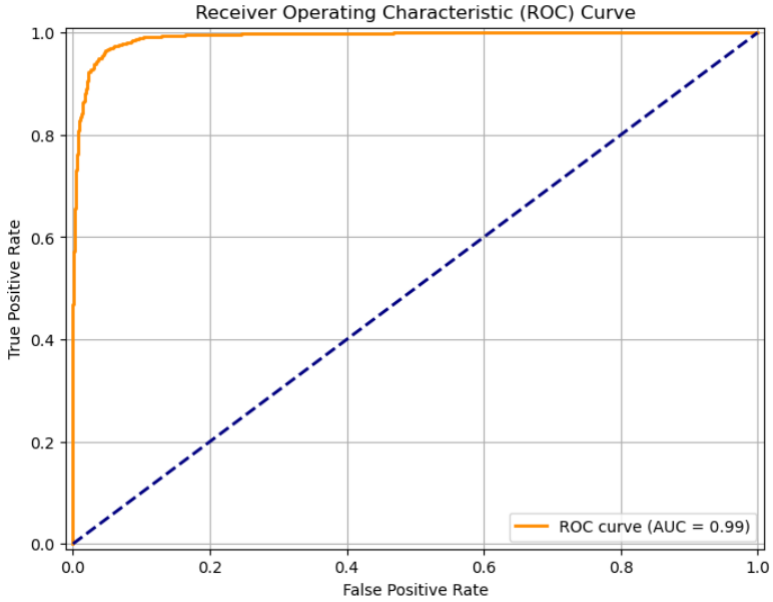
The CatBoost model performed the best, with an accuracy of 96.24%. This model was chosen for deployment due to its superior performance in predicting diabetic outcomes.



The confusion matrix allows us to evaluate the model's performance in predicting both possible outcomes: 1 (Diabetic) and 0 (non-diabetic). It is important not only to consider overall accuracy, but also to pay close attention to the percentages of false positives and false negatives, as minimizing these errors is crucial in medical predictions. In our model, we are successfully achieving low rates of both, indicating reliable and accurate classification.

Interpretation of Prediction Results:

The CatBoost model's high AUC-ROC score indicates that it is excellent at distinguishing between diabetic and non-diabetic individuals. The model's predictions are most influenced by the number of pregnancies, plasma glucose level, and age, which aligns with the findings from the EDA.



To evaluate the classification performance beyond standard accuracy metrics, we computed the ROC curve and calculated the Area Under the Curve (AUC). The ROC curve plots the true positive rate (sensitivity) against the false positive rate for different classification thresholds. AUC represents the model's ability to distinguish between diabetic and non-diabetic patients.

A value of AUC close to 1 indicates excellent classification performance, while a value around 0.5 suggests a random model. In our case, the final model (CatBoostClassifier within a stacked ensemble) achieved a high AUC score, confirming its robustness even with slight class imbalance.

The ROC curve shows a strong performance of the CatBoost model with an AUC score of 0.99, confirming its reliability on imbalanced binary classification.

DEPLOYMENT: *Web Application for Real-Time Diabetes Prediction*

A web application was built using Python and the Flask framework to serve our machine learning model. The trained model, originally developed in a Jupyter Notebook, was saved using the Pickle library, which allows for easy serialization of Python objects.

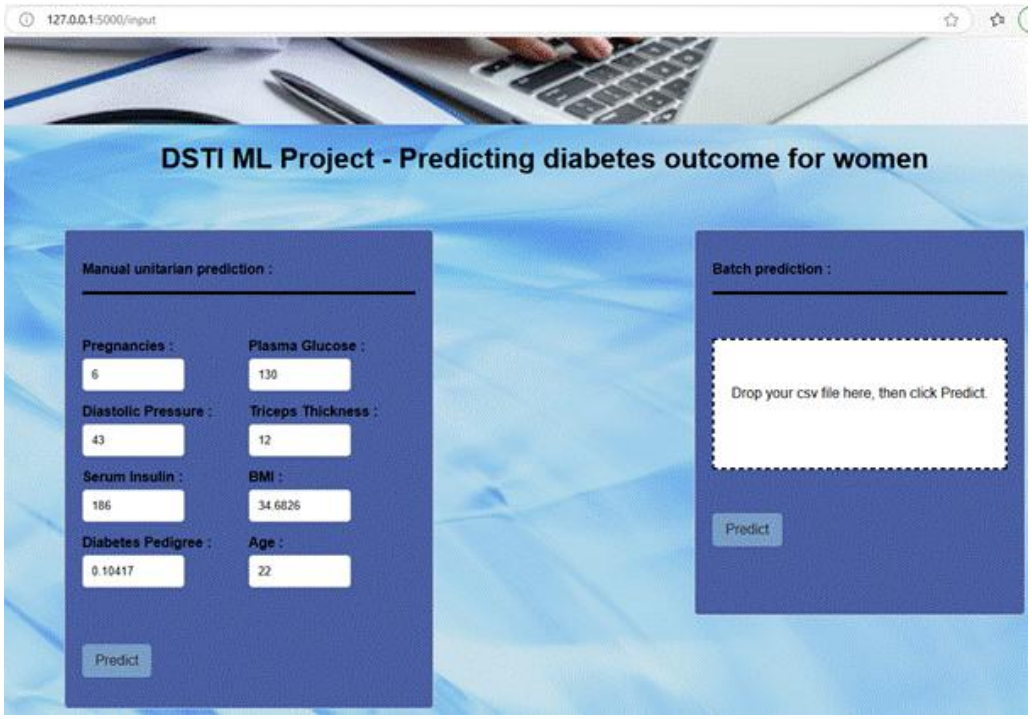
At runtime, the Flask app loads the model from a .pkl file and uses it to generate predictions based on user input. This setup enables a lightweight, interactive tool for real-time diabetes risk prediction, showcasing how machine learning can be integrated into practical applications.

The web application provides two ways to make predictions: a unitarian mode (values set by hand) and a batch mode, using a file drop zone created with the Flask_DropZone library, which can process a text file containing many rows.

The unitarian mode's result page displays the model's prediction: "DIABETIC" (orange background) or "NOT DIABETIC" (green background), along with the computed probability of the result (thanks to the predict_proba() method).

The batch mode result's page displays a preview of the first 10 results and provides a link to upload the full results set.

The application uses Fickle to load a fitted model. On the server side, color logging has also been implemented to monitor prediction activity, using Colorama library.



CONCLUSION: *Key Learnings, Challenges and Future Perspectives*

This project successfully built and deployed a machine learning model to predict diabetic outcomes for women based on eight key features. The Catboost gradient boosting model (a decision tree algorithm) outperformed other models, achieving high accuracy and AUC-ROC scores.

Our dataset had no real flaws or difficulty: no missing data, no categorical feature, no normalization problem (applying such method did not show noticeable difference).

In the EDA phase we only got rid of 221 records (1,5% of the data frame, of which 12% were diabetic). The work essentially consisted in testing different models and fine-tune them to produce the better result. Apart from logistic regression algorithms, several algorithms lead to an accuracy above 90%. When the accuracy is above 90% this task becomes harder to improve.

To make our solution accessible and practical, we also developed a web-based application. This user-friendly interface allows healthcare professionals or users to input patient data and receive real-time predictions regarding diabetes risk. Such a tool can play a crucial role in early diagnosis and patient management.

For future iterations of the project, we propose several avenues for improvement:

- Incorporating additional features such as lifestyle habits (e.g., physical activity, diet) and genetic predispositions, which could enhance the predictive accuracy.
- Exploring deep learning techniques, which may capture more complex patterns and interactions within the data.
- Expanding the dataset to include a more diverse and representative population, ensuring broader generalizability and fairness across different demographic groups.

You can access the full project and source code on our GitHub repository:

HTTPS://GITHUB.COM/VANES-SA03/DIABETES-PREDICTION-ML