Project Title:
Data Analysis for Diabetes Prediction

Contact Information:
Solae Kim (VC1B)
solae.kim82@bcmail.cuny.edu

Supervisor:
Joshua Abok Nyajuaya

The development of this initiative is currently ongoing, spanning from September 2024 to December 2024. This document represents an incomplete draft based on research conducted up to October 27, 2024.

Table of Contents

0. Introduction

0.1 Preface

In modern society, AI technology is rapidly spreading across nearly every industry, including healthcare, finance, education, and manufacturing, driving transformative changes. However, alongside these advancements, concerns over misuse and ethical issues have emerged. The rise of deepfake technology, which enables the dissemination of false information, privacy violations, and bias in AI-based hiring systems reveals the potential for severe negative consequences when AI is misapplied. As a result, there is a growing awareness that AI must transcend mere innovation to serve the public good and promote societal values.

With this sense of urgency and responsibility, I embarked on this project to explore and realize the positive impact AI can have. Through public healthcare data analysis powered by AI models, I aim to demonstrate that artificial intelligence is not merely a fleeting trend or subject of debate but a field with the potential to enhance public welfare and uphold social responsibility. Furthermore, I hope to highlight AI's potential as a fundamental tool in computer science, one that not only fuels technological progress but also plays a crucial role in creating a healthier environment for future generations.

Through this proejct, I aspire to share a vision where leading-edge technology can contribute to a better tomorrow by embracing social responsibility. It is my sincere hope that AI, when harnessed thoughtfully and ethically, will become a force for positive change, a tool that promotes public good and helps to create a more equitable and meaningful impact on society.

0.2 Abstract

This project focuses on independently mastering data analysis techniques through the development of a diabetes prediction model using public data. By leveraging Python and machine learning methodologies, I will preprocess and scale the data and train models to accurately identify risk patterns. The project will involve detailed data analysis, system optimization, and performance evaluation to deepen understanding of practical algorithm applications and explore their potential. Ultimately, this project aims to expand knowledge in computer science by reflecting current technological trends and exploring technologies that can contribute to the public good.

1. Resources

1.1 Key Tools

1. IDE

1) Google Colaboratory: A free web-based platform where you can write and run Python code directly in your browser, offering Jupyter Notebook features along with free access to GPUs and TPUs for high-performance computing.

2. Python Libraries

1) Pandas: For handling and analyzing data.
2) NumPy: For numerical operations.
3) Scikit-learn: For machine learning models.
4) Matplotlib/Seaborn: For creating charts and graphs.

3. Machine Learning Algorithms

1) Logistic Regression: In this project, linear regression was chosen for its simplicity and interpretability. Linear regression provides a clear understanding of the relationship between predictor variables and the target variable, making it easier to communicate results to non-technical stakeholders. Additionally, its computational efficiency minimizes resource consumption, and it offers a useful benchmark for evaluating model performance. For these reasons, linear regression is a practical and effective choice for this project.

4. Version Control Platform

1) GitHub: This platform enables the efficient storage and sharing of updated files, as well as the systematic tracking of project progress.

1.2 Dataset Description

In this project, I use the **Pima Indians Diabetes Database**, collected by the National Institute of Diabetes and Digestive and Kidney Diseases. This dataset includes health information on individuals of Pima Indian heritage aged 21 and older, designed to support diagnostic predictions of diabetes based on various health indicators. The primary features of the dataset are as follows:

- Pregnancies: Number of times pregnant. This factor is included under the hypothesis that pregnancy experience may impact diabetes risk.

- Glucose: Plasma glucose concentration after glucose tolerance test. Elevated glucose levels are considered one of the main indicators of diabetes.

- BloodPressure: Blood pressure (mm Hg). Blood pressure levels are closely associated with diabetes risk, with hypertension being a well-known related factor.

- SkinThickness: Triceps skinfold thickness (mm). This measurement reflects subcutaneous fat, which may relate to insulin resistance.

- Insulin: Serum insulin (μU/ml). Blood insulin concentration aids in assessing insulin resistance, a key factor in diabetes.

- BMI: Body Mass Index, calculated as weight (kg) divided by the square of height (m^2). Obesity is a known risk factor for diabetes, making BMI an important consideration.

- DiabetesPedigreeFunction: Diabetes pedigree function, indicating genetic influence. This index reflects the likelihood of diabetes based on family history, accounting for genetic factors in diabetes risk.

- Age: Age of the patient. Older age is associated with an increased risk of diabetes, making it a significant factor.

- Outcome: Diabetes status. This binary variable indicates diabetes presence, with 0 representing non-diabetic and 1 representing diabetic status. It serves as the target variable for binary classification.

This dataset is composed of clinically and physiologically significant factors for predicting diabetes onset, enabling healthcare providers to identify high-risk groups early and take preventive actions. Key variables, such as glucose levels, blood pressure, BMI, and family history, highlight important metabolic, genetic, and lifestyle factors associated with diabetes. Building predictive models based on these elements allows for the proactive identification of at-risk individuals, supporting appropriate lifestyle modifications, medical guidance, and monitoring.

1	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
2	6	148	72	35	0	33.6	0.627	50	1
3	1	85	66	29	0	26.6	0.351	31	0
4	8	183	64	0	0	23.3	0.672	32	1
5	1	89	66	23	94	28.1	0.167	21	0
6	0	137	40	35	168	43.1	2.288	33	1
7	5	116	74	0	0	25.6	0.201	30	0
8	3	78	50	32	88	31	0.248	26	1
9	10	115	0	0	0	35.3	0.134	29	0
10	2	197	70	45	543	30.5	0.158	53	1
11	8	125	96	0	0	0	0.232	54	1
12	4	110	92	0	0	37.6	0.191	30	0

Figure 1. Example of Original Pima Indians Diabetes Database Structure

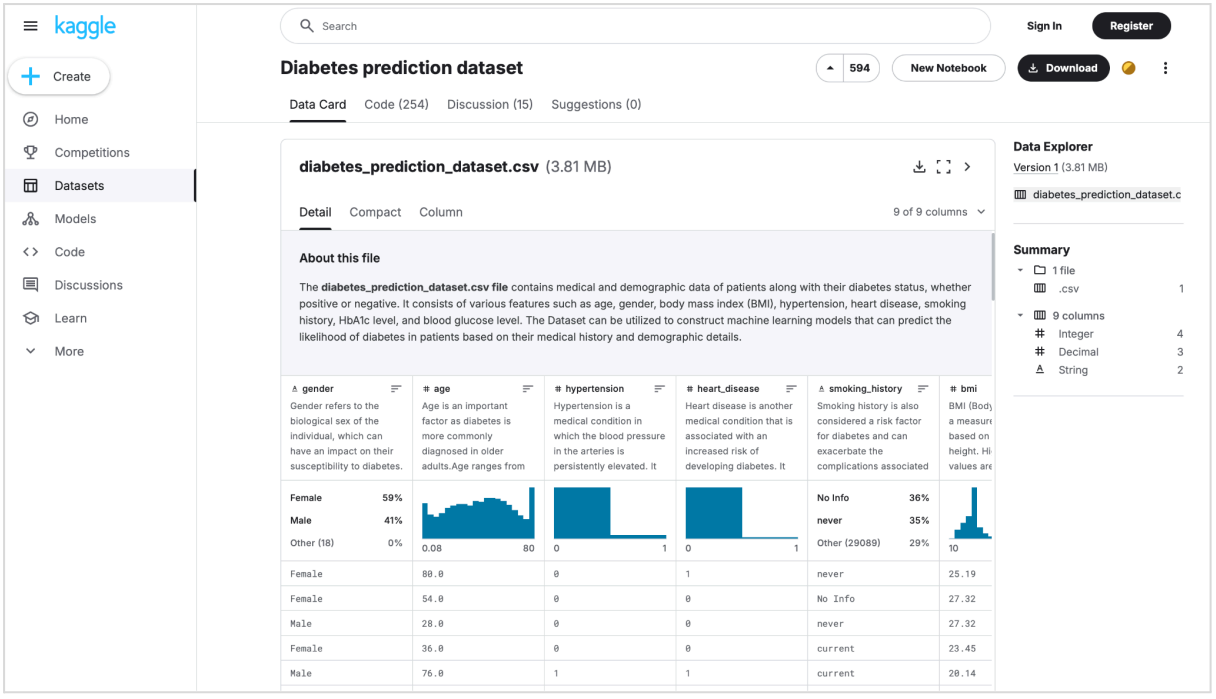


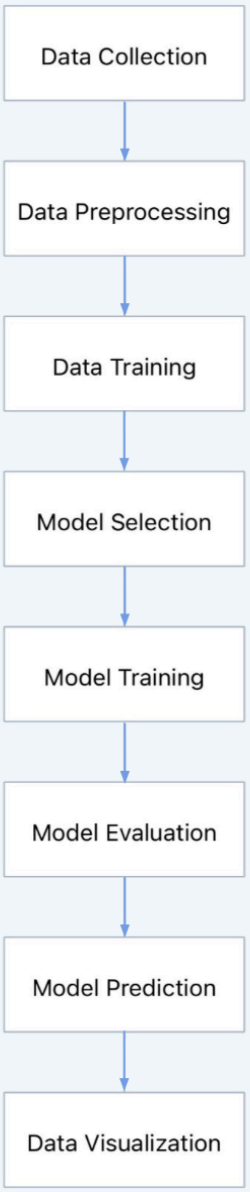
Figure 2. Dataset Source Website "Kaggle". As the world's largest data science community, Kaggle provides a platform where companies share large datasets, enabling numerous data scientists, teams, and individuals to work on solving real-world problems using big data.

Dataset Source: Mustafa, I. (n.d.). *Diabetes Prediction Dataset - A Comprehensive Dataset for Predicting Diabetes with Medical & Demographic Data*. Kaggle. Available at: https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset

2. Methodology

2.1 Data Analysis Design

As depicted in the flow diagram below, this project starts with data collection, advances through data preprocessing and model training, and culminates in the generation and visualization of predictive results.



1. Data Collection: Load the dataset to prepare the training and testing datasets.

2. Data Preprocessing: Normalize the data and save the transformers.

3. Data Training: Split the data into training and validation sets, and use the training data to train the model.

4. Model Selection: Define the appropriate algorithm or method for the model.

5. Model Training: Train the model on the training data, and monitor metrics such as loss and accuracy.

6. Model Evaluation: Evaluate the model's performance on the validation set to ensure it generalizes well.

7. Model Prediction: Use the model to make predictions on new data and output the results.

8. Data Visualization: Visualize training progress, performance metrics, and predictions to gain insights.

Figure 3. Basic Flow Diagram of the Project

2.2 Model Training Procedure

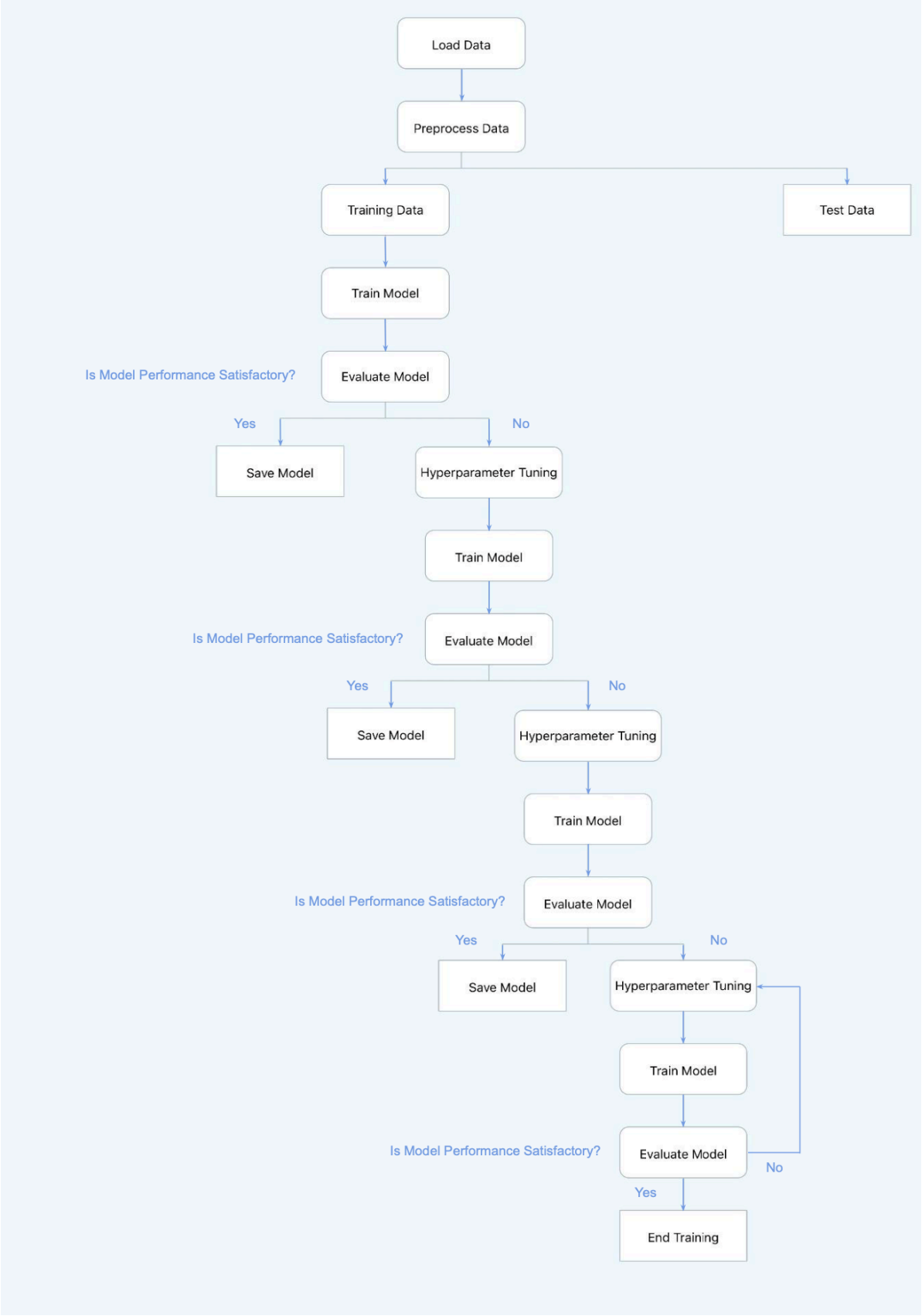


Figure 4. Flow Diagram of the Model Training Process

The diagram above illustrates the training process for obtaining predictive results using diabetes data. Initially, the data is loaded and preprocessed, which includes handling missing values and performing normalization and scaling. The data is then split into training and testing sets, and an appropriate machine learning algorithm is selected to train the model. Following the evaluation of the model's performance, if the results are satisfactory, the model is saved and the project is concluded. If the performance is insufficient, hyperparameter tuning is carried out to refine the model, with iterative adjustments made until satisfactory results are achieved. Once the tuned model meets the desired performance, it is saved, and the project is finalized.

2.3 Data Training Procedure

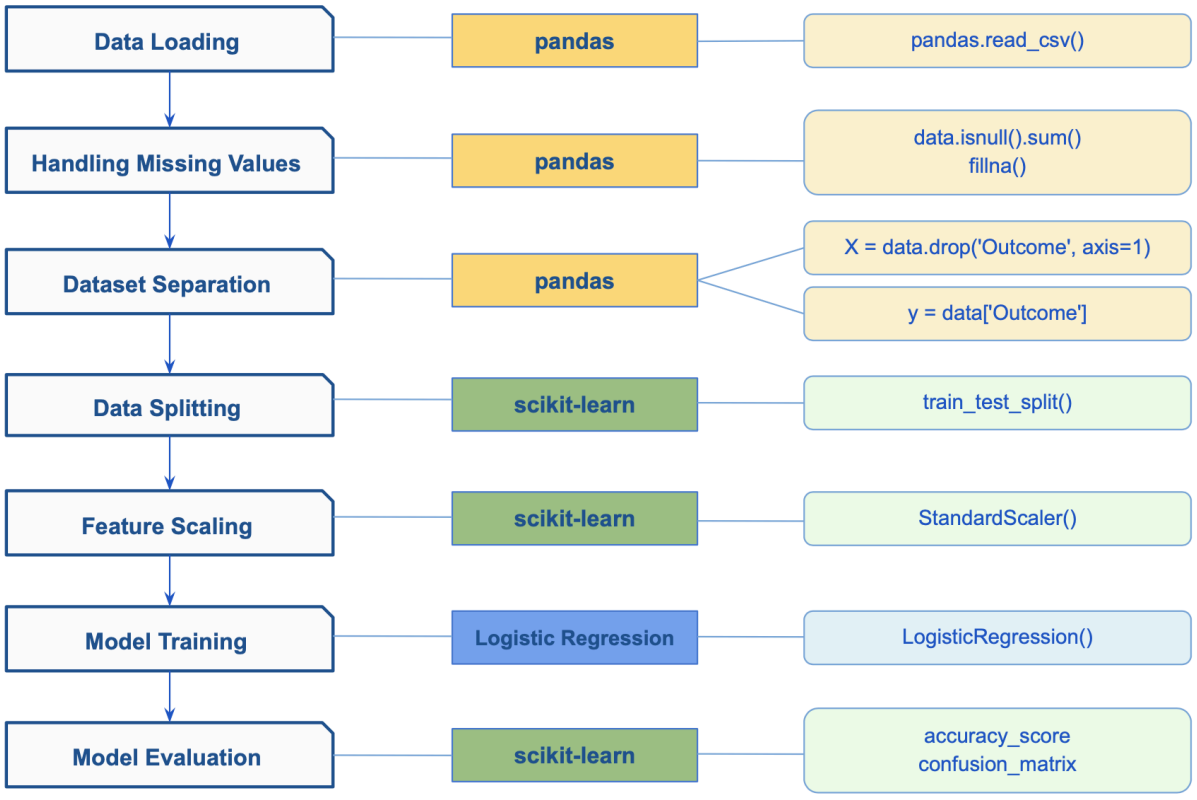


Figure 5. Flow Diagram for the key steps in data training. It illustrates the stages of data preprocessing, training, and evaluation, with each step representing a fundamental component of the machine learning pipeline.

This project utilizes the Python programming language, with key libraries including pandas and scikit-learn. Pandas is employed for data processing and analysis, supporting tasks such as data loading and preprocessing. Scikit-learn is used for constructing and evaluating machine learning models, encompassing tasks such as data splitting, feature scaling, model training, and performance evaluation. Notably, Logistic Regression is applied in this project to address binary classification problems by predicting the probability of specific events

(such as diabetes) based on various input features. The following details effectively illustrate how these tools and libraries are utilized in the data training stages.

1. Data Loading (pandas): Utilize pandas.read_csv() to import the dataset from a CSV file, which will be used for training the machine learning model.

2. Handling Missing Values (pandas): Identify missing values using data.isnull().sum(), and address them with fillna(). Proper handling of missing data is crucial as it can impact model performance and lead to errors, thus ensuring the dataset remains clean and reliable.

3. Feature and Label Separation (pandas): Separate the dataset into features (X) and labels (y) using X = data.drop('Outcome', axis=1) and y = data['Outcome']. This allows the machine learning model to predict the label based on the input features.

4. Data Splitting (scikit-learn): Use train_test_split() to divide the dataset into training and testing sets, with 80% allocated for training and 20% for testing. This separation is essential to evaluate the model's generalization performance, ensuring that the evaluation is not biased by the training data.

5. Feature Scaling (scikit-learn): Standardize the feature values using StandardScaler() to achieve a mean of 0 and a standard deviation of 1. Scaling features improves model efficiency and prevents any particular feature from disproportionately influencing the model due to differences in scale.

6. Model Training (Logistic Regression): Train the Logistic Regression model using LogisticRegression(). Logistic Regression is a binary classification algorithm that predicts the probability of a specific event, such as diabetes, based on various input features (e.g., age, glucose levels, blood pressure, body mass index). This model is particularly well-suited for binary classification problems, where the goal is to distinguish between two outcomes—such as the presence (1) or absence (0) of diabetes. During the training process, the model learns patterns from the provided training data that relate features to outcomes. These learned patterns are then applied to predict the likelihood of diabetes in unseen test data, thus equipping the model with the capability to make predictions on new data.

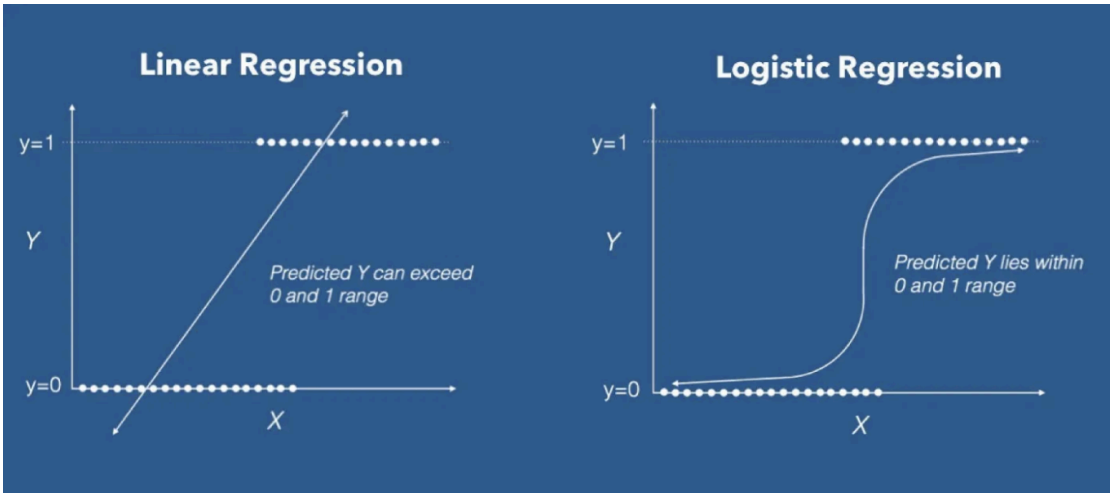


Figure 6. A comparison between Linear and Logistic Regression (Image:https://medium.com/@maithilijoshi6/a-comparison-between-linear-and-logistic-regression-8aea40867e2d)

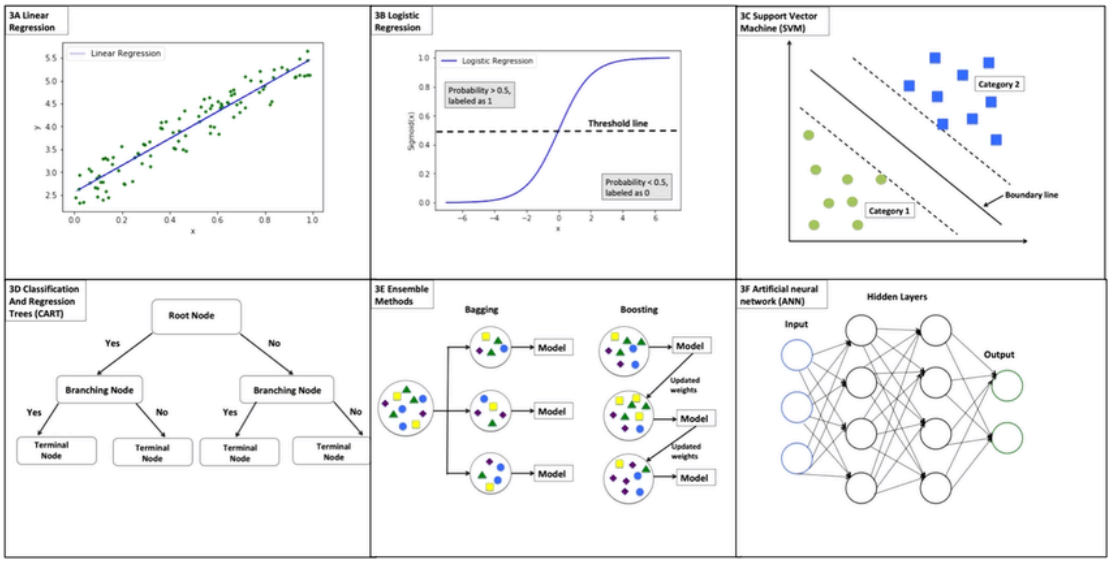


Figure 7. Illustrations of machine learning models. 3A. Linear regression; 3B. Logistic regression; 3C. Support vector machine; 3D. Classification and regression trees (CART); 3E. Ensemble methods; 3F. Artificial neural network (ANN) (Image:https://www.researchgate.net/figure/Illustrations-of-machine-learning-models-3A-Linear-regression-3B-Logistic-regression_fig3_339541927)

7. Model Evaluation (scikit-learn): Assess the model's performance using accuracy_score and confusion_matrix. These metrics provide insights into the model's accuracy and the rate of misclassifications, enabling an evaluation of the model's effectiveness and potential areas for improvement.

2.4 Prototype Code

Initial prototype code and data visualizations related to this project are available for review on GitHub: https://github.com/solaecloud/DiabAnalysis/blob/main/DiabAnalysis.ipynb

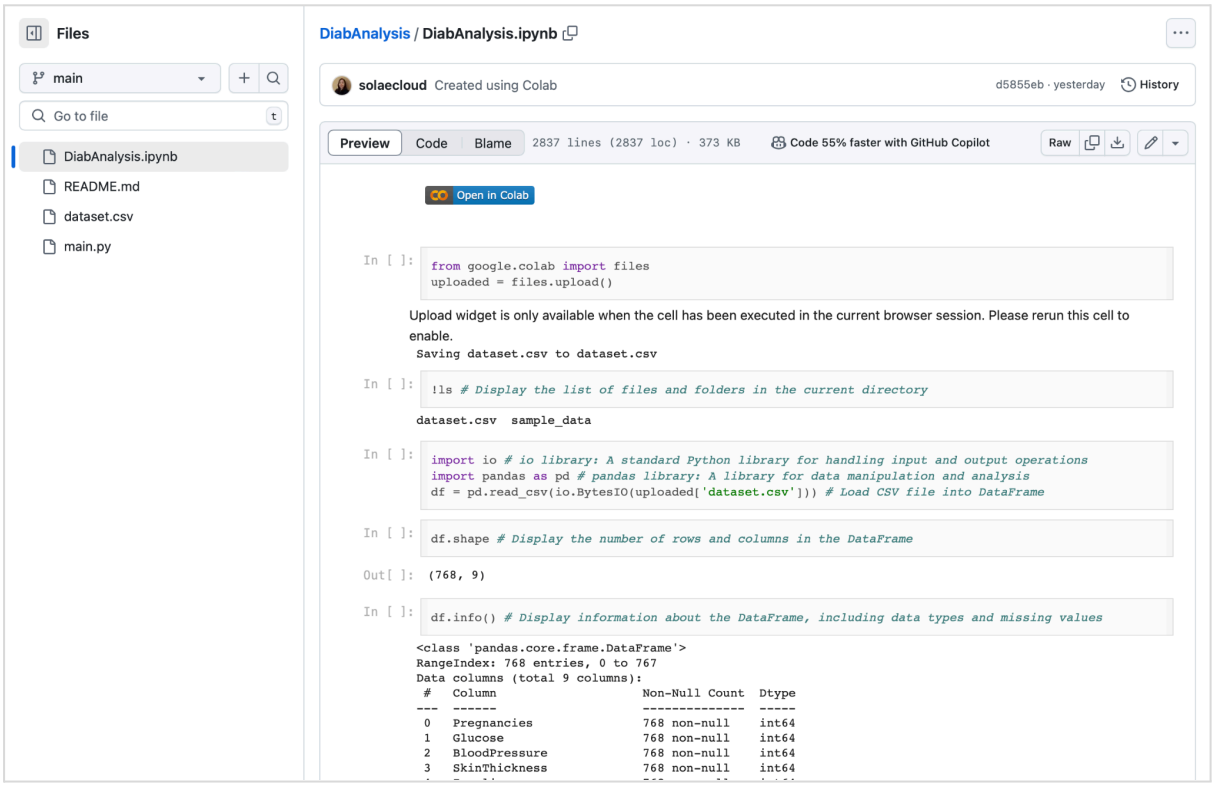


Figure 8. Initial Prototype Code and Visualization Examples

3. Project Timeline

3.1 Tentative Schedule

Dates	Activities	Objectives
10/28-11/03	Initial Model Selection and Training	Select appropriate algorithms, train initial models
11/04-11/10	Model Evaluation and Tuning	Evaluate model performance, tune hyperparameters
11/11-11/21	Advanced Model Improvement	Implement advanced techniques and models for better performance
11/22-11/30	Result Visualization and Interpretation	Create visualizations, interpret results
12/01-12/12	Final Review and Presentation Preparation	Conduct final review and prepare presentation materials

3.2 Repository

The project timeline and updated deliverables will be stored on GitHub:

- Github
  https://github.com/solaecloud/DiabAnalysis

- Github Project Management Board
  https://github.com/users/solaecloud/projects/2

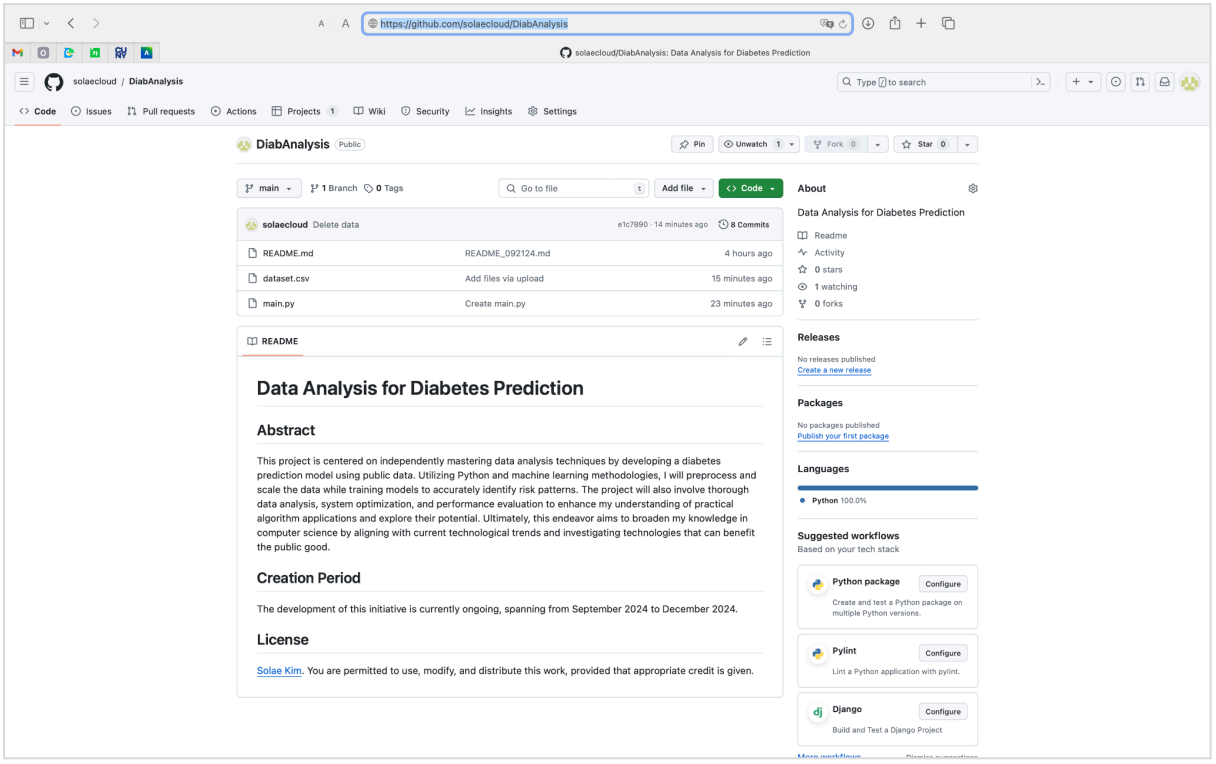
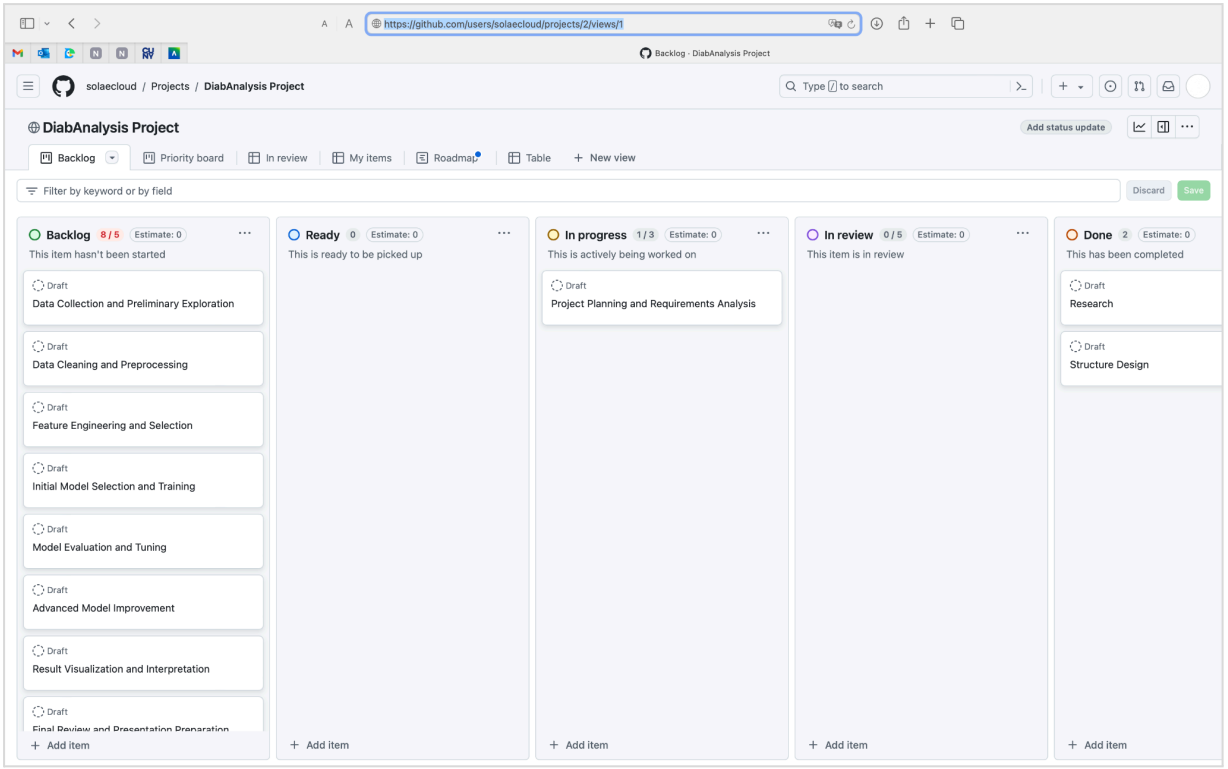
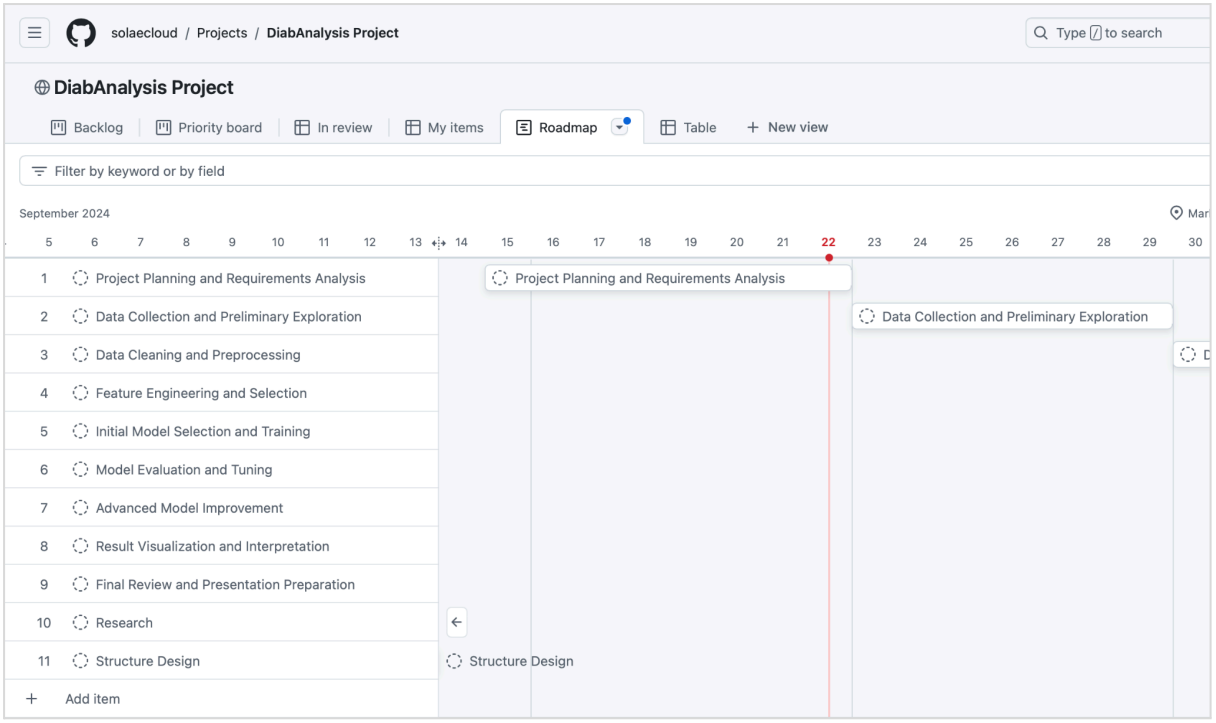


Figure 9. Github Repository





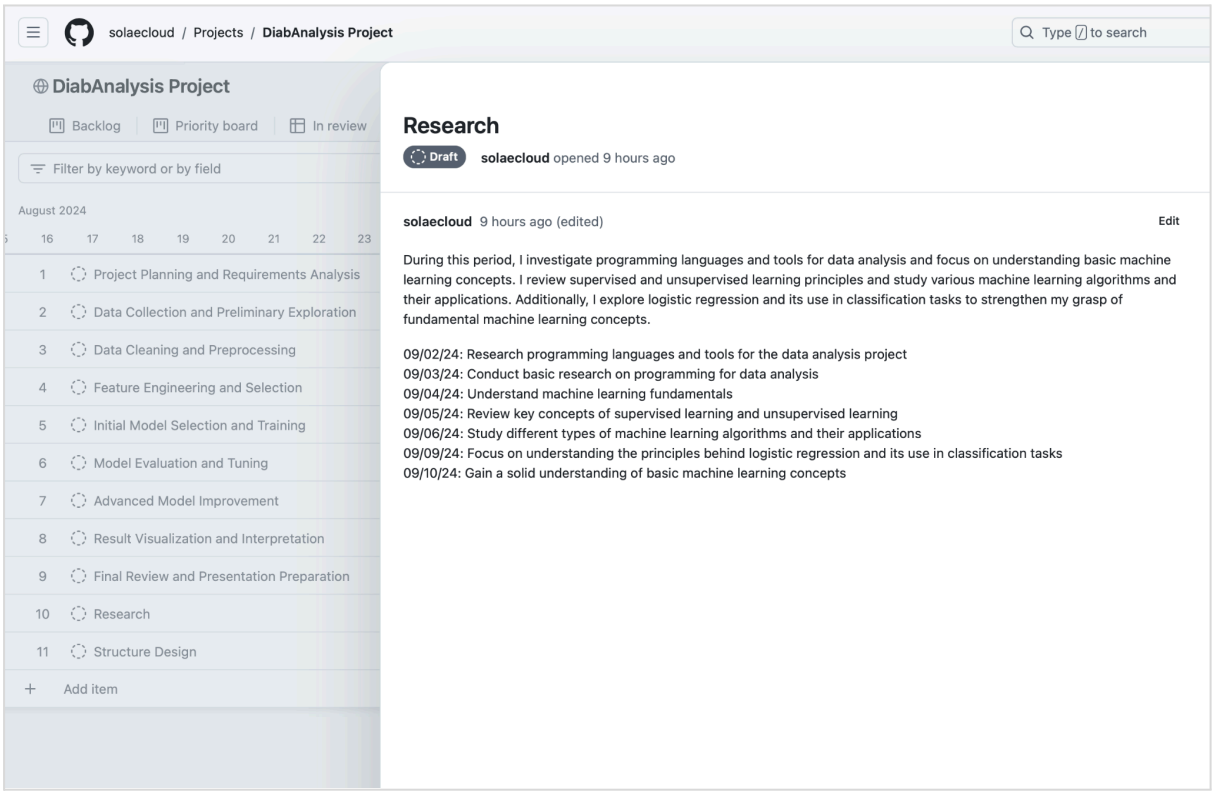


Figure 10. Sample GitHub Project Management Boards

4. Practical Applications

4.1 Use Cases

- **Case 1:**
    - **Input:** Individual health data including various features such as age, Body Mass Index (BMI), blood glucose levels, blood pressure, and family history.
    - **Output:** Probability of diabetes (a value between 0 and 1) and classification of risk level (e.g., low, medium, high).

- **Case 2:**
    - **Input:** Patient health records based on the Kaggle dataset, including data such as blood glucose levels, weight, and family history.
    - **Output:** Diabetes risk prediction and personalized management recommendations (e.g., 'regular check-ups recommended,' 'lifestyle changes advised').

- **Case 3:**
    - **Input:** Aggregated health information from the Kaggle dataset, such as average blood glucose levels, BMI, and diabetes prevalence by region.
    - **Output:** Analysis results for public health policy and strategy development (e.g., 'identification of high-risk areas,' 'assessment of prevention program needs').

4.2 Examples of Potential Applications

The insights and methodologies derived from this project can be applied across various domains to address pressing health challenges and improve public health outcomes. The following examples illustrate how the techniques and findings from this project can be utilized in real-world scenarios:

1. **Medical Research Data Analysis:** According to the 2023 data from the International Diabetes Federation (IDF), approximately 530 million people worldwide suffer from diabetes, which accounts for about 10.5% of the global adult population. To address this issue, researchers can analyze global health datasets to study diabetes onset patterns, understand risk factors, and develop new treatment methods and preventive strategies. This analysis provides essential foundational data for advancing research and improving public health interventions.

2. **Personal Health Management Apps:** The prevalence of diabetes continues to rise,

particularly in low-income countries. Given the substantial costs associated with diabetes management and treatment, providing a personal health management app that assesses diabetes risk can enable individuals in underserved areas to monitor their health status in real time. This proactive approach allows users to identify risk factors early and take preventive measures, thereby improving health outcomes even in regions with limited medical resources.

3. **Addressing the Healthcare Workforce Shortage:** To alleviate the shortage of healthcare professionals, hospitals and clinics can implement automated diabetes prediction systems. These systems evaluate patients' health statuses in advance, enabling medical staff to prioritize and focus on those requiring urgent care. This approach can reduce the burden on healthcare professionals and enhance the overall quality of care, contributing to more efficient and effective healthcare delivery.

---

5. Conclusion

This section will be added upon project completion.

(These fictional results below illustrate potential factors that might be associated with diabetes risk:

1. Blood Glucose Levels and Diabetes Risk: The analysis showed that individuals with a blood glucose level of 130 or higher had about a 70% higher likelihood of developing diabetes. This suggests that blood glucose management could play a key role in diabetes prevention.

2. Pregnancy Count and Diabetes Correlation: A trend was observed where an increase in the number of pregnancies correlated with a higher risk of diabetes. Women with five or more pregnancies had a 45% higher incidence of diabetes compared to the average.

3. BMI and Diabetes Association: Higher BMI (Body Mass Index) was associated with increased diabetes occurrence. Women with a BMI of 28 or higher were about 1.8 times more likely to develop diabetes compared to the average.

4. Age and Diabetes Incidence: The likelihood of diabetes increased with age. Among those aged 45 or older, the risk was approximately 1.4 times higher than among individuals under 35.

5. Skin Thickness and Diabetes Correlation: Skin thickness measurements of 25mm or more showed a slight increase in diabetes risk, although the association was less pronounced

compared to other factors.

6. Insulin Levels and Diabetes: Individuals with fasting insulin levels above 120 were about 60% more likely to develop diabetes than the average.)

---

6. References

1. Ng, A. (2017, August 15). *Machine Learning Specialization* [Video series]. YouTube. DeepLearning.AI. https://www.youtube.com/watch?v=vStJoetOxJg&list=PLkDaE6sCZn6FNC6YRfRQc_FbeQrF8BwGI
2. GeeksforGeeks. (2023, October 28). *Libraries in Python*. GeeksforGeeks. https://www.geeksforgeeks.org/libraries-in-python/
3. Geekster. (n.d.). *Data Analyst Project for Beginner: Analysis of Diabetes Health*. Geekster. https://www.geekster.in/articles/data-analyst-project-for-beginner-analysis-of-diabetes-health/