Capital University of Science and Technology

Department of SOFTWARE ENGINEERING

Course Title: AI ARTIFICAL INTELLEGENCE

PROJECT

Semester: Fall 2025

TEACHER: ADANA KARMAT

NAME: HAZIB RAHSEED

Reg No: BAI243037

Contents

1. Introduction

Predicting disease progression is a critical task in medical Artificial Intelligence, assisting healthcare professionals in identifying high-risk patients early. This project focuses on analyzing the **Diabetes Dataset** to predict a quantitative measure of disease progression one year after baseline.

Although **Linear Regression** is fundamentally a regression algorithm designed for continuous outputs, this project adapts it for a **binary classification** task (High Progression vs. Low Progression). By predicting the continuous progression score and applying a threshold, we classify patients into risk groups. This approach provides a valuable exercise in understanding the relationship between regression outputs and classification decision boundaries.

Main objectives of this project:

- To preprocess and analyze the Diabetes dataset.
- To train a Linear Regression model on physiological features.
- To evaluate performance using both Regression metrics (MSE, R^2) and Classification metrics (Accuracy, Precision, Recall).
- To visualize results using appropriate graphs.

2. Dataset Description

- **Dataset Name:** Diabetes Dataset (Scikit-learn)
- **Source:** Bradley Efron, Trevor Hastie, et al. (2004) "Least Angle Regression".

The dataset consists of ten baseline variables (physiological features) and one target variable for 442 diabetes patients. The features have been mean-centered and scaled by the standard deviation times the number of samples.

Key Attributes

- **age:** Age in years
- **sex:** Gender
- **bmi:** Body Mass Index
- **bp:** Average Blood Pressure
- **s1 (tc):** T-Cells (a type of white blood cell)

- **s2 (ldl):** Low-density lipoproteins
- **s3 (hdl):** High-density lipoproteins
- **s4 (tch):** Thyroid stimulating hormone
- **s5 (ltg):** Lamotrigine
- **s6 (glu):** Blo;od sugar level

Dataset Preview

The following table shows a sample of the processed feature data and the target variable:

Feature (age)	Feature (sex)	Feature (bmi)	Target (Y)
0.038	0.050	0.061	151.0
-0.001	-0.044	-0.051	75.0
0.085	0.050	0.044	141.0
0.041	0.050	-0.016	206.0

3. Libraries and Tools Used

The following Python libraries were used in this project:

- **Pandas:** For data loading and dataframe manipulation.
- **NumPy:** For numerical computations and array handling.
- **Scikit-learn:** For model training (LinearRegression), dataset loading, and evaluation metrics (mean_squared_error, r2_score).
- **Matplotlib:** For data visualization.

4. Data Preprocessing

4.1 Binary Target Creation

To evaluate the regression model as a classification model (as per the report requirements), the continuous target variable is converted into a binary class based on the median value:

- **1 (High Progression):** $Value > Median$
- **0 (Low Progression):** $Value \le Median$

4.2 Feature Selection

- The original quantitative target column is separated as Y.
- All 10 physiological columns (age through s6) are treated as input features X.

5. Train-Test Split

The dataset is divided into training and testing sets to ensure an unbiased evaluation:

- **Training Data:** 70% (309 samples)
- **Testing Data:** 30% (133 samples)
- **Random State:** 99

This split ensures reproducibility of the results.

6. Model Training

A **Linear Regression** model is trained using the training dataset ($train\_x$, $train\_y$).

The model learns to map the input features (BMI, Blood Pressure, etc.) to the disease progression score. The training process resulted in the following key coefficients:

- **Intercept:** 155.59
- **Highest Positive Coefficient:** s5 (661.96)
- **Second Highest Coefficient:** bmi (517.18)

7. Prediction Strategy

7.1 Continuous Predictions

These are the direct outputs of the Linear Regression model.

- *Example:* For a specific test case, the Actual value was **75.0** and the Predicted value was **77.99**.

7.2 Binary Predictions

Continuous predictions are converted into class labels using the median threshold strategy:

- Values greater than the median are classified as **High Progression**.
- Values less than or equal to the median are classified as **Low Progression**.

8. Evaluation Metrics

8.1 Regression Metrics

To evaluate how well the regression model fits the data, the following metrics are used:

- **Mean Squared Error (MSE): 3157.96**
    - *Analysis:* Represents the average squared difference between estimated values and the actual value.
- **R^2 Score: 0.4546**
    - *Analysis:* The model explains approximately 45.5% of the variance in the target variable.

8.2 Classification Metrics

To assess classification performance (High vs. Low progression), the following metrics are calculated:

- **Confusion Matrix**
- **Accuracy**
- **Precision**
- **Recall**
- **F1 Score**

9. Visualizations

9.1 Confusion Matrix Heatmap

A heatmap visualization of the confusion matrix helps identify correct predictions versus misclassifications (False Positives and False Negatives) between the High Progression and Low Progression classes.

9.2 Classification Metrics Bar Chart

A bar chart comparing Accuracy, Precision, Recall, and F1 Score provides a quick visual summary of the model's reliability in distinguishing between patients with mild vs. severe disease progression.

10. Results Summary

- The model achieves an **R^2 score of ~0.45**, suggesting a moderate correlation between the physiological features and disease progression.
- The features **s5 (ltg)** and **bmi** have the strongest influence on the predictions, as indicated by their high coefficient values.
- The MSE value indicates that while the model captures the general trend, individual predictions can vary significantly from the actual values.

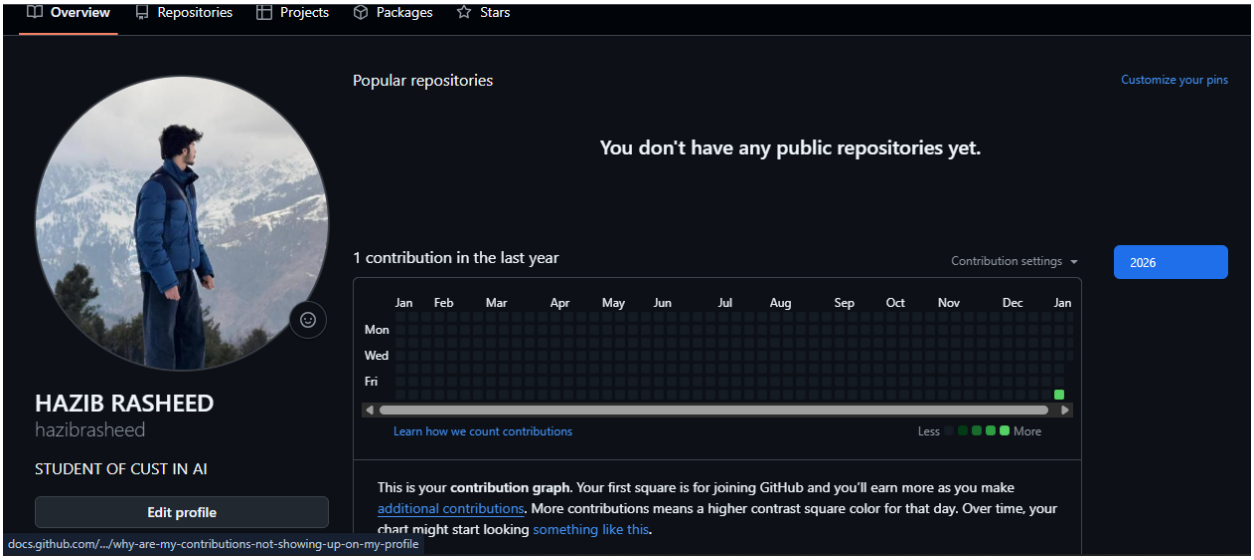
11. Limitations

- **Linearity:** Linear Regression assumes a straight-line relationship between physiological markers and disease progression, which may oversimplify complex biological processes.
- **Algorithm Suitability:** Linear Regression is not the optimal algorithm for classification problems; it is sensitive to outliers which can skew the decision threshold.
- **Feature Complexity:** The dataset relies on baseline variables and may miss other external factors (diet, lifestyle) that influence diabetes progression.

12. Future Improvements

- **Logistic Regression:** Replace Linear Regression with Logistic Regression to natively handle the binary classification of High/Low risk.
- **Advanced Models:** Experiment with non-linear models such as **Random Forest Regressors** or **Support Vector Machines (SVM)** to potentially improve the R^2 score.
- **Hyperparameter Tuning:** Perform Grid Search or Cross-Validation to optimize model parameters.
- **Feature Engineering:** Explore interaction terms (e.g., combining BMI and Age) to capture more complex relationships in the data.

REFERENCE:


Overview
Repositories
Projects
Packages
Stars
Popular repositories
Customize your pins
You don't have any public repositories yet.
1 contribution in the last year
Contribution settings
2026
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec Jan
Mon
Wed
Fri
Learn how we count contributions
Less More
This is your contribution graph. Your first square is for joining GitHub and you'll earn more as you make additional contributions. More contributions means a higher contrast square color for that day. Over time, your chart might start looking something like this.
HAZIB RASHEED
hazibrasheed
STUDENT OF CUST IN AI
Edit profile
docs.github.com/.../why-are-my-contributions-not-showing-up-on-my-profile