My First Machine Learning Model

Diabetes Progression Prediction – Regression Project

Project Overview

This project focuses on predicting the progression of diabetes in patients based on clinical measurements. The objective was to build a regression model that estimates disease progression one year after the baseline using patient data. The dataset was obtained from the sklearn.datasets module, which is a well-known benchmark dataset for regression tasks.

Dataset Description

The dataset consists of 442 observations and 10 standardized input features representing physiological measurements, along with one target variable indicating disease progression.

Features:

- **age** – Patient's age
- **sex** – Patient's sex (numerically encoded)
- **bmi** – Body Mass Index
- **bp** – Average blood pressure
- **s1 to s6** – Blood serum measurements

Target:

- A quantitative score measuring disease progression after one year

Objective

To predict the target value (diabetes progression score) using regression techniques and evaluate how well the model performs in making these predictions.

Tools & Technologies

- **Language**: Python
- **Libraries**: Scikit-learn, Pandas, Matplotlib, NumPy
- **Model Used**: Linear Regression
- **Evaluation Metrics**: Mean Squared Error (MSE), R^2 Score

Workflow Summary

1. Data Loading & Preparation

The dataset was loaded and converted into a pandas DataFrame for easier analysis and manipulation. Features and target variables were separated for training.

2. Train-Test Split

The dataset was divided into training and testing sets using an 80-20 split to ensure fair evaluation of the model's performance on unseen data.

3. Model Training

A Linear Regression model was used as the baseline algorithm to learn from the training data and capture relationships between features and the target variable.

4. Evaluation

The model was evaluated using:

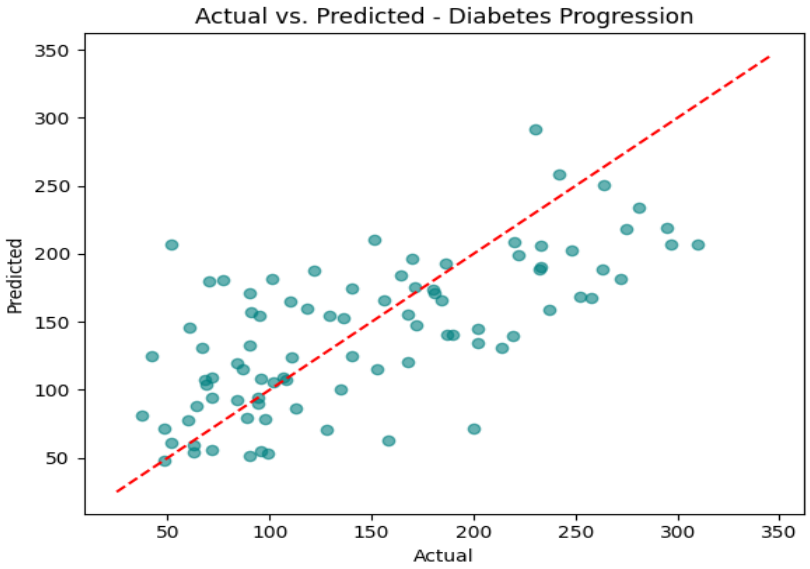
- **Mean Squared Error (MSE)**: Measures the average squared difference between actual and predicted values.
- **R^2 Score**: Indicates how much variance in the target variable is explained by the model (a higher value is better).

5. Visualization

A scatter plot of predicted vs. actual target values was created to visualize the model's accuracy. A perfect model would have all points lying on the diagonal line.

Results

- The Linear Regression model achieved an **R^2 score of approximately 0.45**, meaning it could explain about 45% of the variation in diabetes progression.
- The **MSE was around 2900**, indicating the average error between predicted and actual values.
- Features like **BMI**, **BP**, and **serum s5** showed strong influence on the target prediction.



Insights & Future Scope

- The model provides a basic yet effective regression approach to predicting disease progression.
- Although Linear Regression gave a good starting point, other models such as **Random Forest**, **Ridge Regression**, or **XGBoost** could further improve prediction performance.
- Feature engineering and hyper parameter tuning are recommended as next steps for optimization.

Conclusion

- This project demonstrates how a supervised learning approach, particularly regression, can be used in the healthcare domain to predict disease progression. By leveraging patient clinical data, we can build models that provide valuable support in medical decision-making.