DIABETIC PROGRESSION PREDICTION USING LINEAR REGRESSION

A Comprehensive Report on predicting diabetic progression over 1-year using linear regression

Group Members:

1. Bisrat Asaye _____ UGR/8508/14
2. Naol Daba_____ UGR/4777/14
3. Yordanos Zegeye_____UGR/6316/14

Submitted to: Mr. Bisrat

11.04.2024

INTRODUCTION

What is Diabetes?

Diabetes is a chronic disease that features excessive levels of glucose in the blood, which, if not properly managed, might lead to a variety of health complications. The disease ensues when the body either does not produce enough insulin or cannot use the insulin it does produce appropriately. Insulin is a hormone that has a vital role in regulating blood glucose levels and converting them into energy for the cells in the body.

The two major forms of diabetes are Type 1 and Type 2. Type 1 diabetes is an autoimmune condition where the body's immune system turns against itself and completely destroys insulin-producing cells in the pancreas. There is complete loss of insulin production; hence, these individuals have to rely on insulin therapy for life.

Type 2 diabetes, on the other hand, is a progressive disease in which the body becomes resistant to insulin or unable to produce enough insulin to maintain normal glucose levels. Type 2 diabetes used to be associated with dietary factors and decreased physical activity, and it is more typical for adults. Type 2 diabetes can sometimes be controlled by adjusting the diet and exercise, along with medications, but it is different from Type 1 in that as the disease progresses, it may also require insulin.

This project will focus specifically on Type 2 diabetes, as it is the most prevalent form and its long-term progression can be complex to manage. The consequences of the development of diabetes, especially Type 2, are the onset of cardiovascular disease, damage to the kidneys, damage to the nerves, and damage to vision. One-year diabetes progression prediction allows understanding of a patient's trajectory better and taking proactive measures to manage potential risks. Such early prediction may help in tailoring treatment plans, improving quality of life, and decreasing the likelihood or severity of complications.

Aim of Using Linear Regression

Linear regression was chosen for its simplicity and because it's good at predicting continuous outcomes, which would be the case with diabetic indicator progression over a while. This model will give predictions on how features such as age, BMI, and BP

contribute to diabetic progression. Through this, the project will have a basic model that can provide quick predictions from available data to help in early treatment and intervention of Type 2 diabetes.

DATASET OVERVIEW

The dataset used in this project is publicly available on Kaggle. This dataset is designed to help one study the development of diabetes, in this case, the Type 2 diabetes over a one-year time period. It comprises about 400 entries, where each entry represents a different subject with various clinical and demographic attributes relevant for the prediction of diabetic progression. While each attribute in this dataset conveys different aspects of a patient's health profile, they offer a multi-dimensional view that can inform predictive models on disease progression.

The dataset includes the following attributes:

AGE: Refers to the age of the patient; an important feature in modeling diabetes progression since the risk of and damage from diabetes usually increases with age.

SEX: The sex of a patient; this is usually coded in binary form to manifest any sex-specific differences in progression and response to treatment of diabetes.

BMI (Body Mass Index): A commonly used indicator of health that gives an estimate of body fat percentage in relation to height and weight. Greater the BMI, greater is the risk of developing type 2 diabetes and the faster its progression.

BP: Systolic blood pressure measurements that may reflect the patient's cardiovascular health. Hypertension control is very important in the patient with diabetes, since it can worsen complications related to diabetes.

S1-S6: These represent six attributes, denoting measurements from serum and other biochemical data, including levels of cholesterol and blood lipid concentrations. All these metrics reflect the state of metabolism, liver function, and lipid profile of the patient.

S1: Level of total cholesterol, related to cardiovascular risk.

S2: Low-density lipoprotein (LDL) levels, a type of cholesterol associated with cardiovascular complications.

S3: High-density lipoprotein (HDL) level, also known as "good cholesterol."

S4: Total triglycerides—a type of fat in the blood, the high level of which can indicate metabolic syndrome.

S5: Measures blood glucose levels, directly relevant to diabetes monitoring.

S6: A marker of liver enzyme levels, possibly associated with metabolic and liver health.

Y (Outcome): This is the dependent variable and represents a quantitative measure of diabetic progress after one year. The unit of measurement of the outcome does not appear to be explicitly described in the dataset; however, it should be a composite score based on clinical and laboratory assessments.

Importance of the Dataset Features

The dataset has different features, each giving insight into different factors influencing diabetes progression. For example, well-known risk factors for diabetes and its progression include age, BMI, and blood pressure. However, biochemical measurements S1-S6 give much more detailed insight into the metabolic profile of each patient. Including such diversified metrics allows the model to capture the multifactorial nature underlying Type 2 diabetes progression.

Data Quality and Suitability for Predictive Modeling

The structure and variety in the dataset make it apt for building a predictive model. However, real-world datasets always bear the possibility of tedious handling of missing values, outliers, and normalization of data to ensure that no particular feature over-influences predictions of the model. Moreover, the size of this particular dataset (~400 samples) gives moderately sufficient data, which requires careful strategies for training, validation, and testing to ensure maximal accuracy without overfitting. This dataset allows for a comprehensive approach in the prediction of Type 2 diabetes progression and an assessment of how demographic and health metrics contribute to disease advancement. Since diabetes is multidimensional, the data is a good starting point for developing an interpretable model that practitioners can use to develop targeted treatment and management strategies.

MODEL TRAINING AND SELECTION

In this project, we trained and evaluated two models to predict diabetic progression: a **Linear Regression** model and a **Stochastic Gradient Descent (SGD) Regressor**. Both

approaches offer distinct advantages and were chosen to balance simplicity and interpretability with tuning flexibility for optimal predictive performance.

Data Splitting Strategy

To ensure robust training and evaluation, the dataset was divided into training, validation, and test sets:

- **Training Set (60%)**: Used to train the model and optimize for the lowest training error.
- **Validation Set (20%)**: Used to tune hyperparameters and evaluate performance on unseen data to prevent overfitting.
- **Test Set (20%)**: Used solely for final performance evaluation after training and tuning to provide an unbiased estimate of model performance.

This split ensures that the model is optimized while remaining generalizable to new data.

Linear Regression Model

The Linear Regression model serves as the baseline, providing a simple, interpretable relationship between features and the target outcome. This model was trained on the training set and evaluated on the validation and test sets. The performance metrics recorded in `evaluation_metrics.txt` for the Linear Regression model include:

- **Validation Mean Squared Error (MSE)**: 2415.69
- **Validation R^2 Score**: 0.5810
- **Test Mean Squared Error (MSE)**: 3245.61
- **Test R^2 Score**: 0.4389

The moderate R^2 values indicate that the model captures some but not all of the variability in diabetic progression, highlighting the potential for improved accuracy with further model tuning or feature engineering.

SGD Regressor Model

The SGD Regressor was chosen as a complementary approach to enable gradient-based optimization and potential performance improvement through regularization. Hyperparameter tuning was conducted by varying the `alpha` (regularization strength) parameter across values [0.0001, 0.001, 0.01, 0.1]. For each value of `alpha`, the model was trained on the training set and evaluated on the validation set, recording the MSE and R^2

scores.

Key results of the SGD tuning process:

- **Alpha = 0.0001**: Lowest MSE, indicating minimal regularization worked best in this instance, likely due to the limited dataset size and simpler relationships.
- The tuning results were stored and printed in `SGD Tuning Results`, revealing a balance between model complexity and predictive performance.

MODEL EVALUATION

The model evaluation focused on assessing the Linear Regression model's performance on both the validation and test datasets. Key metrics recorded include the Mean Squared Error (MSE) and the R^2 score, which indicate how well the model captures the variability of diabetic progression over time.

- **Validation Set**: The Linear Regression model achieved a **Validation MSE of 2415.69** and an **R^2 score of 0.5810**, indicating that approximately 58.1% of the variability in diabetic progression could be explained by the model. This result suggests a moderate level of accuracy and highlights the model's effectiveness at capturing the main trends in the validation data.
- **Test Set**: For the test set, the model yielded a **Test MSE of 3245.61** and an **R^2 score of 0.4389**. The decrease in the R^2 score from validation to test sets suggests slight overfitting, indicating the model may not generalize perfectly to new data. This finding reveals that although the model performs adequately on the validation data, further improvements, such as feature engineering or advanced regularization, may be needed to enhance generalization.

To better understand model predictions, we visualized the relationship between predicted and actual values on both the validation and test sets. The plots show a general alignment of predictions with actual values, though some divergence indicates areas where the model struggled to capture the precise diabetic progression rate. These findings underscore that while the model captures some variability, diabetic progression may have more complex dynamics that require additional features or advanced models for full accuracy.

CONCLUSION

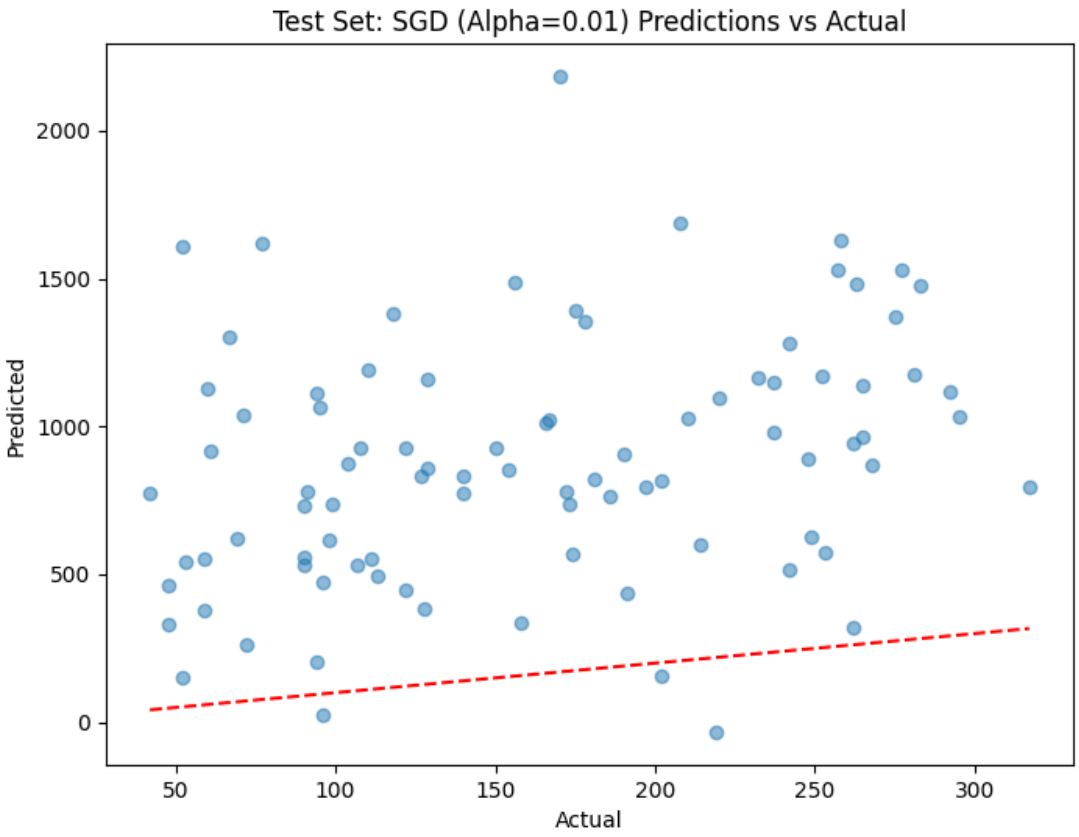
This project aimed to develop a predictive model for diabetic progression over one year using a simple and interpretable approach. By employing a Linear Regression model, we achieved moderate success in predicting diabetic progression, as evidenced by validation and test R^2 scores of 0.5810 and 0.4389, respectively. The model demonstrated a reasonable fit on the training and validation datasets, though a slightly reduced performance on the test data suggests opportunities for further refinement.

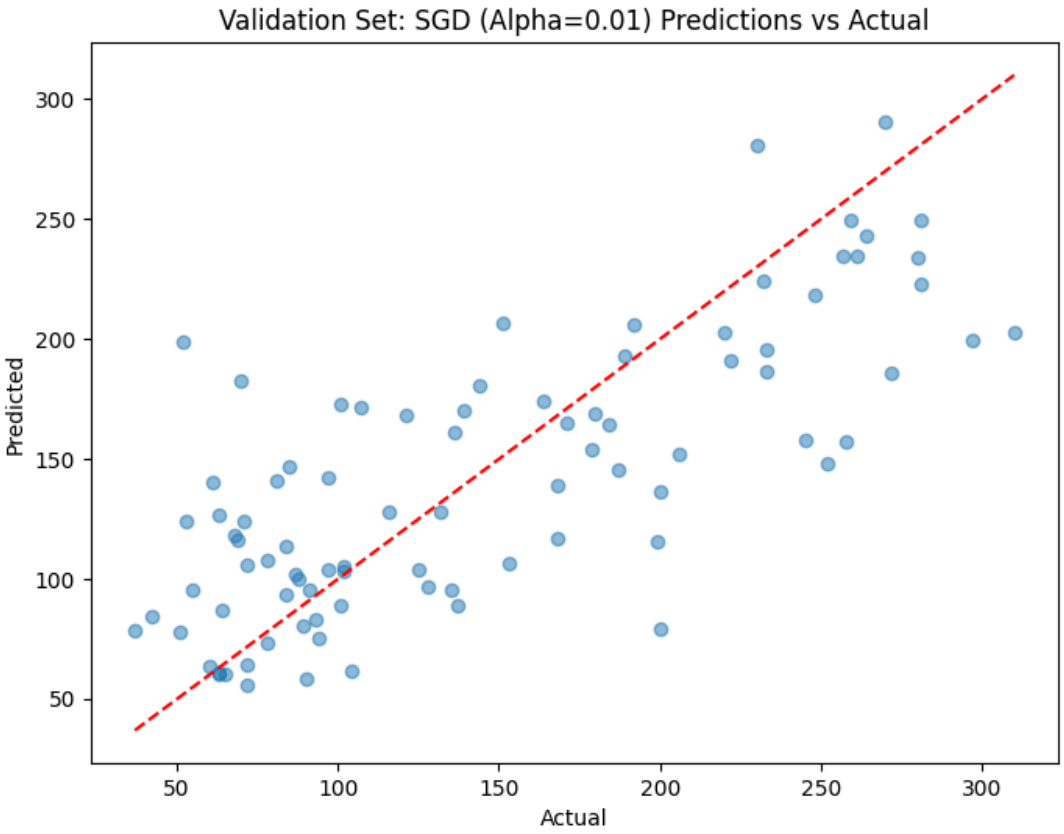
The results confirm that while linear regression can capture broad trends in diabetic progression, the complexity of diabetes and its multifactorial nature may require more advanced models to capture nuanced interactions between features. Future work could include exploring nonlinear models, incorporating domain-specific knowledge for feature engineering, or evaluating ensemble approaches. These steps could yield greater predictive accuracy and robustness, advancing efforts to improve diabetic care and management through predictive analytics.

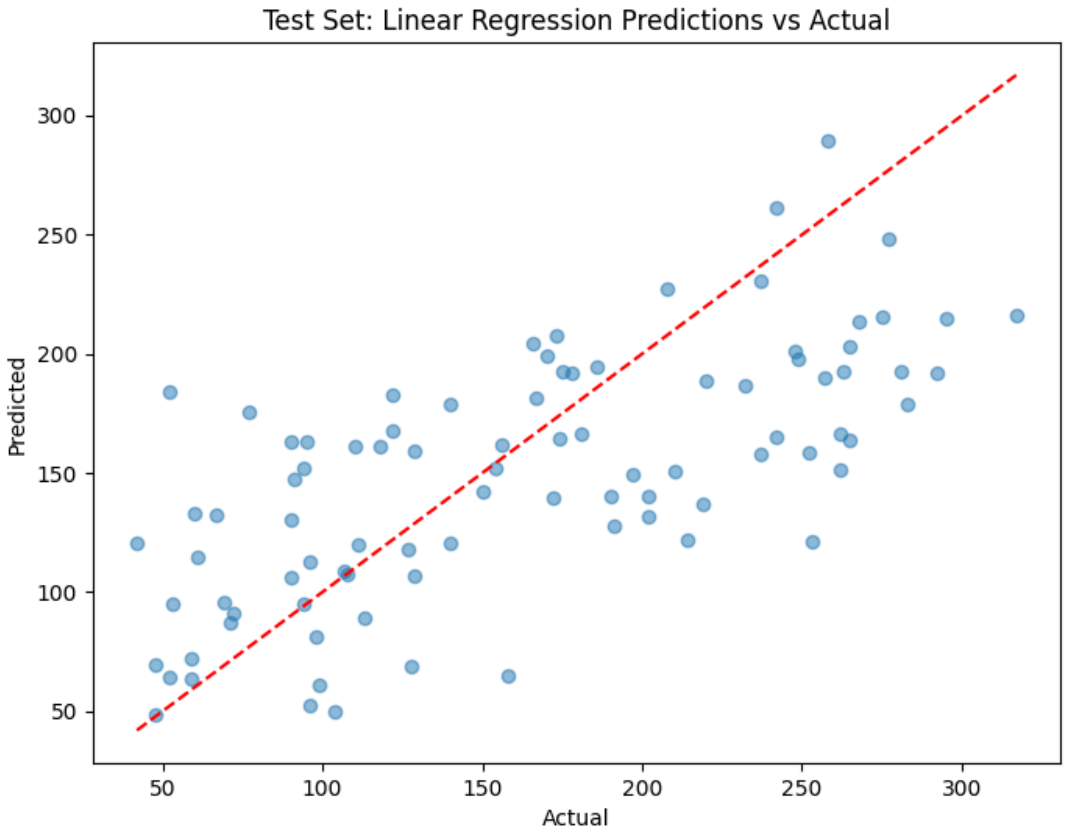
REFERENCES

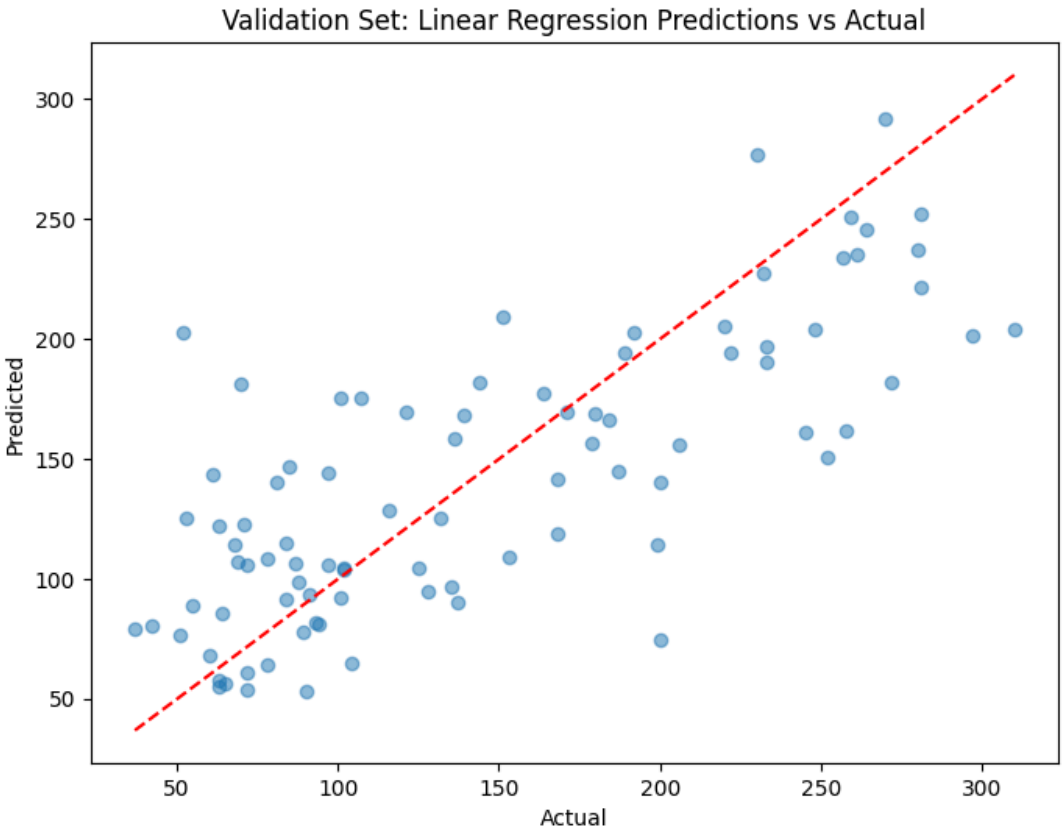
1. American Diabetes Association. (2020). **Standards of Medical Care in Diabetes - 2020**. *Diabetes Care, 43*(Supplement 1): S1–S212. doi:10.2337/dc20-SINT.
2. Hastie, T., Tibshirani, R., & Friedman, J. (2009). **The Elements of Statistical Learning: Data Mining, Inference, and Prediction** (2nd ed.). Springer Science & Business Media.
3. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). **Scikit-learn: Machine Learning in Python**. *Journal of Machine Learning Research, 12*, 2825–2830.
4. Zou, H., & Hastie, T. (2005). **Regularization and Variable Selection via the Elastic Net**. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301-320.

PLOTS


Test Set: SGD (Alpha=0.01) Predictions vs Actual
Predicted
Actual
0
500
1000
1500
2000
50
100
150
200
250
300


Validation Set: SGD (Alpha=0.01) Predictions vs Actual
Predicted
Actual
50
100
150
200
250
300


Test Set: Linear Regression Predictions vs Actual
Predicted
Actual
50
100
150
200
250
300


Validation Set: Linear Regression Predictions vs Actual
Predicted
Actual
300
250
200
150
100
50
50
100
150
200
250
300