

Reflection:

In this section, steps involved in loading data and data preprocessing are discussed.

Dataset Loading: Reading the dataset is done with the help of the pandas library. This is the required file path for the dataset: The file path for this data is then assigned and the data imported into a DataFrame using `pd.read_csv(file_path)`.

Features and Target Definition: The first feature set denoted by X signifies the removal of the Outcome column from the dataset, the second element, denoted by y , represents the Outcome column which corresponds to the predicted variable or the class attribute.

Data Splitting: CROSS VALIDATION The data set is split the data in to training and test data using `train_test_split` method where 80% of the data is used for training and 20% is used for testing. We set the random seed to 42 for reproducibility of the experiments.

Model Training

Model Selection: For training, a LinearRegression model is chosen.

Training Process: The model uses the training data to train for each layer with the fit method using `X_train` and `y_train` data set. Mean Square Error (MSE) will be in used for evaluation.

Prediction: Following on from the training prediction is made on the test set using the command `model.predict(X_test)`.

Evaluation: In this research, the effectiveness of the developed model is analyzed based on Mean Squared Error (MSE). The calculations for the MSE can be done using the default Python function with `mean_squared_error(y_test, y_pred)`. Further, the adjusted R-squared ($\text{Adj } R^2$) is calculated using the formula $1 - [(SSE / (n - p - 1)) / (SST / (n - 1))]$ in order to determine the fitness of the model to the data, in this case using the `r2_score(y_test, y_pred)`.

Analysis of Data Collected on the Problem and Solution (README)

Problem: The work requires making an assessment of the diabetes prognosis considering parameters of ones health, including glucose level, blood pressure, and body mass index.

Solution: Linear regression is employed to model the outcome in this case. The model accuracy is measured with MSE and indicates an error of approximately 0.171, with the R^2 score of 0.255, which calculates how many portions of the variance in the result variable can be expoun in reference to the characterize features. **Future Considerations:** Low ' R^2 '

indicates that a better fit model or feature engineering might help improve the model prediction.