**Breakout Activity**

Create a Jupyter Notebook or Quarto document that solves a regression problem using a supervised learning algorithm (e.g., Linear Regression).

The dataset you use should be related to healthcare (e.g., predicting weight based on height, age, and exercise level). Downloadable from Kaggle or UCI repository

Your notebook should include:

1.Data loading and preprocessing

2.Model training

3.Evaluation using Mean Squared Error (MSE)

4.Reflection on the problem and solution

**Code:**

# Import necessary libraries

import pandas as pd

from sklearn.model\_selection import train\_test\_split

from sklearn.linear\_model import LinearRegression

from sklearn.metrics import mean\_squared\_error

# Load the dataset

file\_path = 'diabetes.csv' # Update the path if necessary

data = pd.read\_csv(file\_path)

# Replace zero values in specific columns with the median

columns\_to\_replace\_zero = ['Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI']

for column in columns\_to\_replace\_zero:

data[column] = data[column].replace(0, data[column].median())

# Define features (X) and target (y)

X = data.drop(columns=['Glucose', 'Outcome']) # Exclude 'Glucose' (target) and 'Outcome' (classification label)

y = data['Glucose']

# Split the data into training and testing sets (80% train, 20% test)

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

# Initialize the Linear Regression model

model = LinearRegression()

# Train the model on the training data

model.fit(X\_train, y\_train)

# Make predictions on the test data

y\_pred = model.predict(X\_test)

# Evaluate the model using Mean Squared Error (MSE)

mse = mean\_squared\_error(y\_test, y\_pred)

print(f"Mean Squared Error (MSE): {mse}")

**output & result:**

A screenshot of a computer

Description automatically generated

The mean squared error is:834.03948

**Reflection:**

The dataset used for this project relates to diabetes and includes features such as Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, and Age. The target variable we aim to predict is Glucose, a key metric in diagnosing diabetes.

This project aimed to predict the glucose level in patients using a variety of health-related features such as age, body mass index, and insulin levels. This type of predictive model could be useful in healthcare for screening or managing diabetes, allowing for early interventions.

One of the challenges I faced was the presence of biologically implausible values (e.g., zero values for blood pressure or insulin). Replacing these values with the median was one way to address the issue, but a more robust solution could involve gathering additional domain-specific knowledge to impute missing values more accurately.

While executing the code I faced some bugs but after correcting the code I get the proper result, and it give correct answer, and I get bugs in zero values in specific columns in the median.

**Conclusion**: This project showcased the steps required to build a regression model to predict glucose levels in patients based on health-related data. Linear Regression served as a starting point, and the use of MSE allowed for a quantitative evaluation of model performance. Further enhancements could be made through advanced modeling techniques and more robust data preprocessing.