DATA SCIENCE

DIABETES CLASSIFICATION

Created & Presented by:

Rinaldi Nurhardiansyah

27 Feb 2025

ibimbing

INTRODUCTION

This project aims to predict the progression of diabetes in patients using a Random Forest Regressor. The dataset used is the Diabetes dataset from Scikit-learn, which contains 10 baseline variables (features) collected from 442 diabetes patients. The target variable is a quantitative measure of disease progression one year after baseline.

`load_diabetes` (*[, return_X_y, as_frame, scaled])	Load and return the diabetes dataset (regression).

GOALS

01.

UNDERSTAND THE DATASET:

Explore the distribution of features and their relationships with the target variable.

02.

BUILD AND TRAIN A REGRESSION MODEL:

Use a Random Forest Regressor to predict diabetes progression

03.

EVALUATE THE MODEL'S PERFORMANCE:

Use metrics like Mean Squared Error (MSE) and R-squared (R^2) to assess the model's accuracy

04.

VISUALIZE THE RESULTS:

Plot the actual vs. predicted values to understand the model's performance.

TOOLS

Write and manage the code

pandas

Data manipulation and analysis

scikit learn

Machine learning tools and datasets

seaborn

Advanced statistical visualizations.

NumPy

Numerical computations.

matplotlib

Data visualization.

DATASET

```python
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn import datasets
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error, r2_score

# Load the Diabetes dataset from scikit-learn
diabetes = datasets.load_diabetes()

# Convert to DataFrame for easier understanding
df = pd.DataFrame(data=diabetes.data, columns=diabetes.feature_names)
df['target'] = diabetes.target

# Display the first 5 rows of the dataset
print("First 5 Rows of the Dataset:")
print(df.head())

# Check the number of rows and columns
print("\nNumber of Rows and Columns:", df.shape)
```

```
First 5 Rows of the Dataset:
        age       sex       bmi        bp        s1        s2        s3        s4        s5        s6  target
0  0.038076  0.050680  0.061696  0.021872 -0.044223 -0.034821 -0.043401 -0.002592  0.019907 -0.017646   151.0
1 -0.001882 -0.044642 -0.051474 -0.026328 -0.008449 -0.019163  0.074412 -0.039493 -0.068332 -0.092204    75.0
2  0.085299  0.050680  0.044451 -0.005670 -0.045599 -0.034194 -0.032356 -0.002592  0.002861 -0.025930   141.0
3 -0.089063 -0.044642 -0.011595 -0.036656  0.012191  0.024991 -0.036038  0.034309  0.022688 -0.009362   206.0
4  0.005383 -0.044642 -0.036385  0.021872  0.003935  0.015596  0.008142 -0.002592 -0.031988 -0.046641   135.0

Number of Rows and Columns: (442, 11)
```

The dataset consists of 11 columns and 442 rows.

FEATURES:

The Diabetes dataset consists of 442 samples and 10 features. Each feature represents a medical measurement, and the target variable is a quantitative measure of disease progression.

- Age: Age of the patient (scaled).
- Sex: Gender of the patient (scaled).
- BMI: Body Mass Index (scaled).
- BP: Average blood pressure (scaled).
- S1: Total serum cholesterol (scaled).
- S2: Low-density lipoproteins (scaled).
- S3: High-density lipoproteins (scaled).
- S4: Total cholesterol / HDL ratio (scaled).
- S5: Log of serum triglycerides level (scaled).
- S6: Blood sugar level (scaled).

EXPLATORY DATA ANALYSIS (EDA)

```
# Check dataset information
print("\nDataset Information:")
print(df.info())
```

```
Dataset Information:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 442 entries, 0 to 441
Data columns (total 11 columns):
 #   Column  Non-Null Count  Dtype  
---  ------  --------------  -----  
 0   age     442 non-null    float64
 1   sex     442 non-null    float64
 2   bmi     442 non-null    float64
 3   bp      442 non-null    float64
 4   s1      442 non-null    float64
 5   s2      442 non-null    float64
 6   s3      442 non-null    float64
 7   s4      442 non-null    float64
 8   s5      442 non-null    float64
 9   s6      442 non-null    float64
 10  target  442 non-null    float64
dtypes: float64(11)
memory usage: 38.1 KB
None
```

DATASET OVERVIEW:

- The dataset contains 442 rows and 11 columns (10 features + 1 target).
- There are no missing values in the dataset.
- All features are numeric and have been scaled.

EXPLATORY DATA ANALYSIS (EDA)

```python
# Check descriptive statistics
print("\nDescriptive Statistics:")
print(df.describe())
```

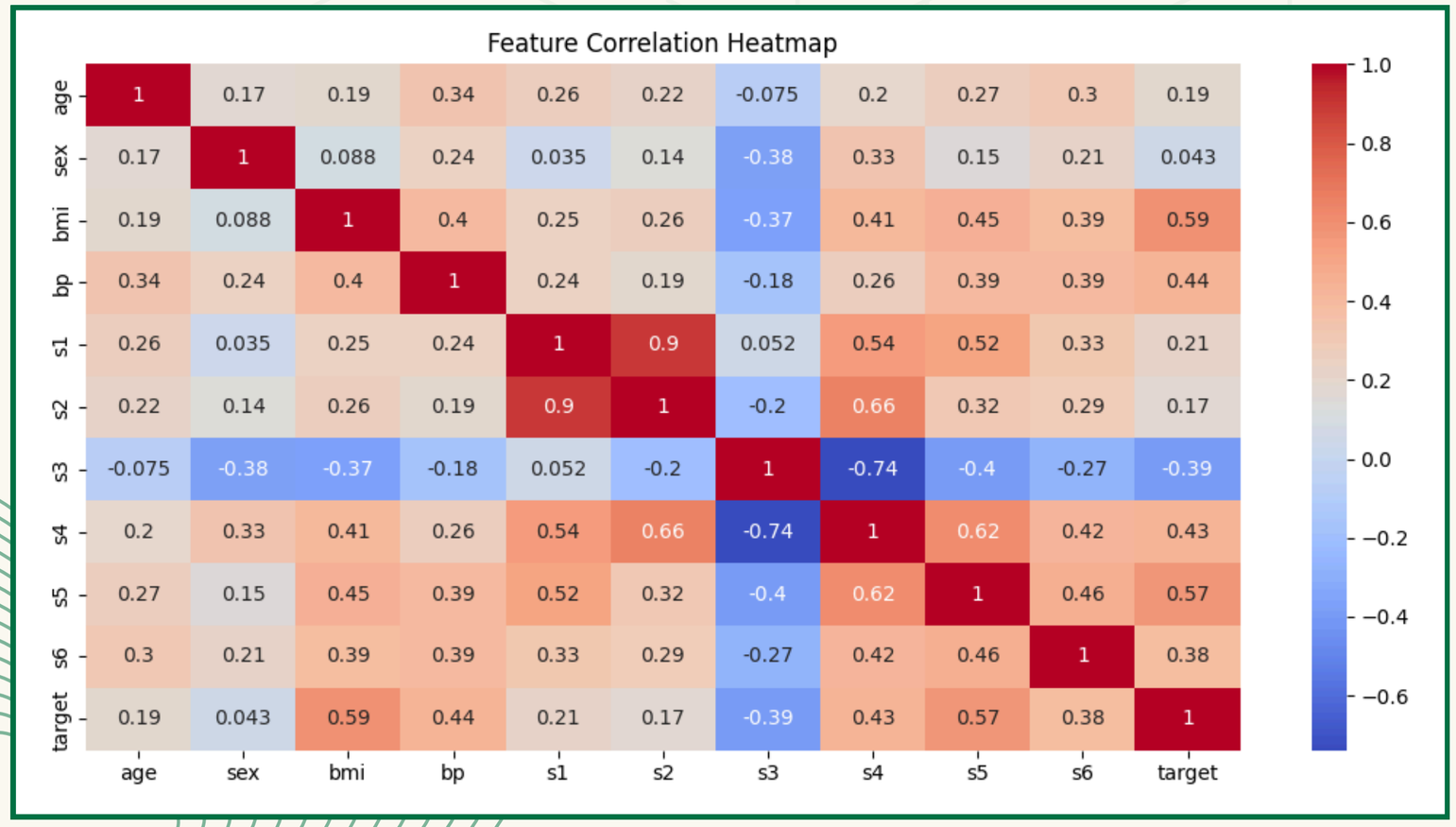
```
Descriptive Statistics:
                age           sex           bmi            bp            s1            s2            s3            s4            s5            s6        target
count  4.420000e+02  4.420000e+02  4.420000e+02  4.420000e+02  4.420000e+02  4.420000e+02  4.420000e+02  4.420000e+02  4.420000e+02  4.420000e+02  442.000000
mean  -2.511817e-19  1.230790e-17 -2.245564e-16 -4.797570e-17 -1.381499e-17  3.918434e-17 -5.777179e-18 -9.042540e-18  9.293722e-17  1.130318e-17  152.133484
std    4.761905e-02  4.761905e-02  4.761905e-02  4.761905e-02  4.761905e-02  4.761905e-02  4.761905e-02  4.761905e-02  4.761905e-02  4.761905e-02   77.093005
min   -1.072256e-01 -4.464164e-02 -9.027530e-02 -1.123988e-01 -1.267807e-01 -1.156131e-01 -1.023071e-01 -7.639450e-02 -1.260971e-01 -1.377672e-01   25.000000
25%   -3.729927e-02 -4.464164e-02 -3.422907e-02 -3.665608e-02 -3.424784e-02 -3.035840e-02 -3.511716e-02 -3.949338e-02 -3.324559e-02 -3.317903e-02   87.000000
50%    5.383060e-03 -4.464164e-02 -7.283766e-03 -5.670422e-03 -4.320866e-03 -3.819065e-03 -6.584468e-03 -2.592262e-03 -1.947171e-03 -1.077698e-03  140.500000
75%    3.807591e-02  5.068012e-02  3.124802e-02  3.564379e-02  2.835801e-02  2.984439e-02  2.931150e-02  3.430886e-02  3.243232e-02  2.791705e-02  211.500000
max    1.107267e-01  5.068012e-02  1.705552e-01  1.320436e-01  1.539137e-01  1.987880e-01  1.811791e-01  1.852344e-01  1.335973e-01  1.356118e-01  346.000000
```

DESCRIPTIVE STATISTICS:

- The target variable (disease progression) ranges from 25 to 346, with a mean of 152.13.
- Features like bmi and s5 show significant variability, which may indicate their importance in predicting the target.

EXPLATORY DATA ANALYSIS (EDA)

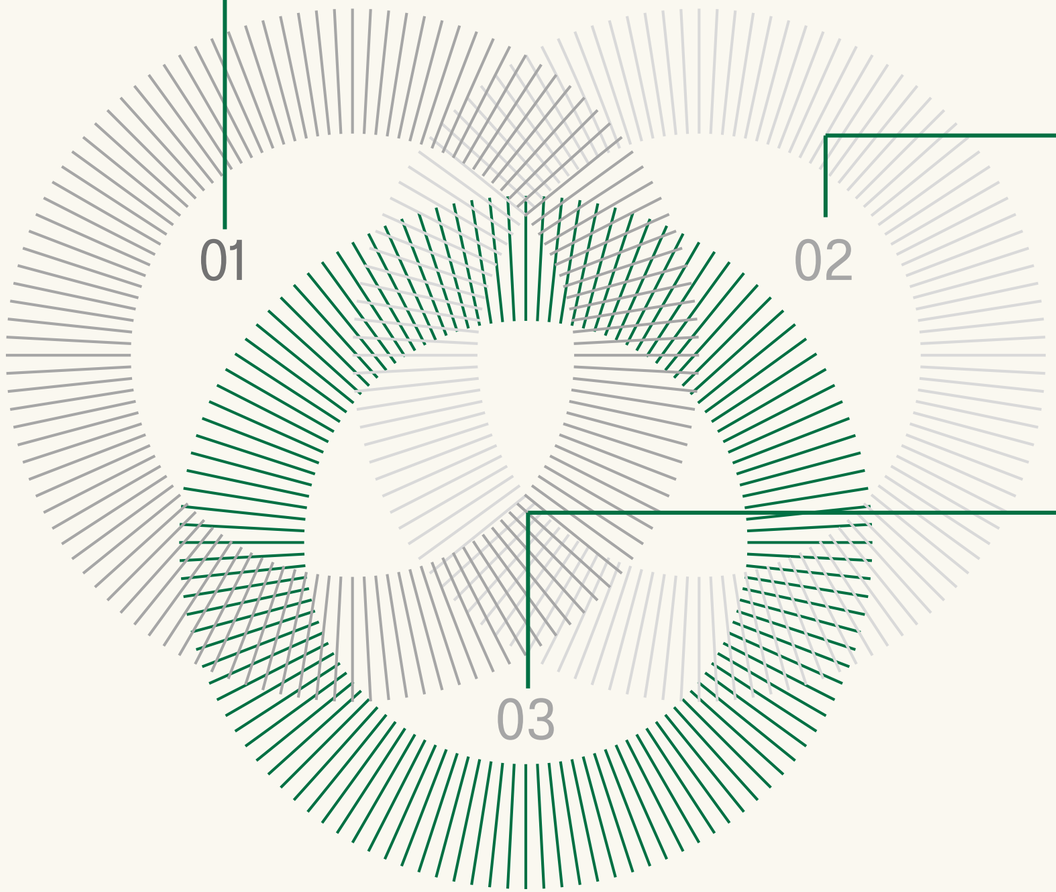
```python
# Heatmap to visualize feature correlations
plt.figure(figsize=(10, 8))
sns.heatmap(df.corr(), annot=True, cmap='coolwarm')
plt.title("Feature Correlation Heatmap")
plt.show()
```



CORRELATION ANALYSIS:

- A heatmap was created to visualize the correlation between features and the target.
- Features like bmi and s5 show moderate positive correlations with the target, suggesting they are important predictors.

ALGORITHM



Random Forest Regressor:

An ensemble learning method that combines multiple decision trees to improve prediction accuracy and reduce overfitting.qua.

Why Random Forest Regressor?

- Handles Non-Linearity
- Robust to Overfitting
- Feature Importance

Steps

- Split the Data: The dataset was split into training (80%) & testing (20%) sets.
- Train the Model: A Random Forest Regressor with 100 trees (n_estimators=100) was trained on the training data.
- Make Predictions: The model was used to predict the target values for the test set.
- Evaluate the Model: Performance was evaluated using Mean Squared Error (MSE) and R-squared (R^2).

MODEL PERFORMANCE

```
# Separate features (X) and target (y)
X = diabetes.data
y = diabetes.target

# Split the data into training (80%) and testing (20%) sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Create a Random Forest model
model = RandomForestRegressor(n_estimators=100, random_state=42)

# Train the model with the training data
model.fit(X_train, y_train)

# Predict on the testing data
y_pred = model.predict(X_test)

# Evaluate the model
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print(f"\nMean Squared Error (MSE): {mse:.2f}")
print(f"R-squared (R2): {r2:.2f}")
```

```
Mean Squared Error (MSE): 2952.01
R-squared (R2): 0.44
```

EVALUATION METRICS:

Mean Squared Error (MSE): 2859.69

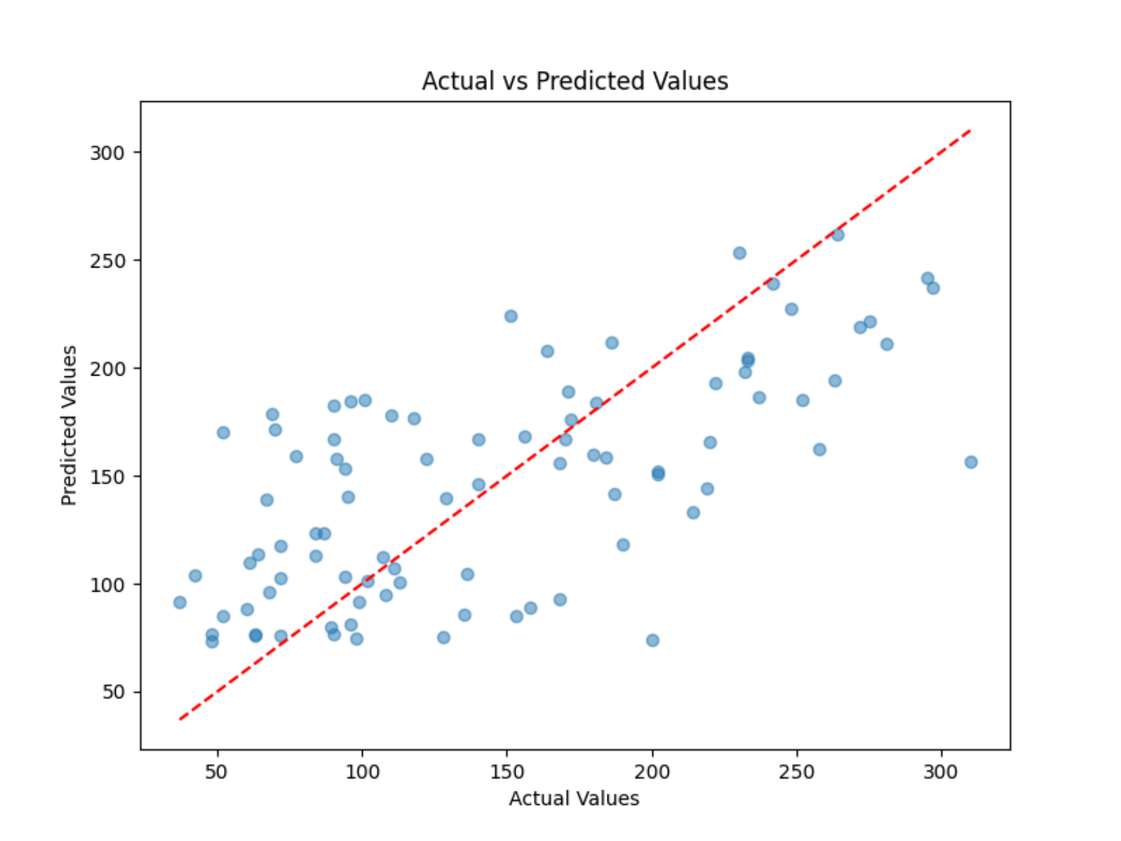
This measures the average squared difference between the actual and predicted values. Lower values indicate better performance.

R-squared (R^2): 0.44

This measures the proportion of variance in the target variable that is explained by the model. An R^2 of 0.44 means the model explains 44% of the variance.

MODEL PERFORMANCE

```python
# Visualize actual vs predicted values
plt.figure(figsize=(8, 6))
plt.scatter(y_test, y_pred, alpha=0.5)
plt.plot([min(y_test), max(y_test)], [min(y_test), max(y_test)], color='red', linestyle='--')
plt.xlabel("Actual Values")
plt.ylabel("Predicted Values")
plt.title("Actual vs Predicted Values")
plt.show()
```



VISUALIZATION:

- A scatter plot was created to compare the actual vs. predicted values.
- The plot shows that the model performs reasonably well, but there is room for improvement, especially for higher target values.

CONCLUSION

FEATURE IMPORTANCE:	Features like bmi and s5 are the most significant predictors of diabetes progression.
MODEL PERFORMANCE:	The Random Forest Regressor achieved moderate performance with an R^2 of 0.44 and an MSE of 2859.69.
VISUALIZATION:	The scatter plot of actual vs. predicted values confirms that the model performs reasonably well but struggles with higher target values.

ibimbing

THANK YOU

Diabetes is not a choice, but we can choose to fight it with courage and resilience 🔥

ABOUT ME

Hello! My name is Rinaldi Nurhardiansyah, a recent Industrial Engineering graduate with a passion for supply chain management and data analytics with a drive to optimize processes and deliver data-driven solutions. My academic background has equipped me with a strong foundation in core industrial engineering principles and statistics, while my self-driven learning journey has allowed me to develop practical skills in data field.

LET'S CONNECT ^_^

rinaldinur

@rinaldinur_14

Rinaldinur-14

+62 882-2400-4553

nurhardiansyah14@gmail.com