INFO 4604-5604

Applied Machine Learning

Spring 2023

Project Report

Student Name	Anton Sandoval
Student ID	[redacted]
Project Title	Predicting Diabetes Progression in Patients: An Exploratory Data Analysis and Machine Learning Approach using the scikit-learn Diabetes Dataset
Date Submitted	May 8, 2023

1. Overall context for your project

Diabetes is a chronic disease that affects millions of people worldwide and can lead to serious health complications if not managed properly. Monitoring the progression of diabetes in patients is crucial for effective management and prevention of complications. Machine learning techniques can assist healthcare professionals in predicting the progression of diabetes in patients based on various factors such as age, BMI, blood pressure, and glucose level.

The scikit-learn diabetes dataset is a widely used dataset for the prediction of diabetes progression. This dataset contains various demographic, lifestyle, and medical variables of 442 diabetes patients. The objective of this project is to apply machine learning algorithms to the diabetes dataset to predict the progression of diabetes in patients.

In this project, we will use linear regression, decision tree, and random forest models to predict the diabetes progression. Linear regression is a commonly used machine learning algorithm for regression problems. Decision tree and random forest models are both tree-based algorithms that can handle both regression and classification tasks. These models are often used for feature selection, data exploration, and prediction.

The results of this project can be used by healthcare professionals to improve the management of diabetes patients. By accurately predicting the progression of diabetes, healthcare professionals can provide early intervention and personalized treatment plans, ultimately improving patient outcomes and quality of life.

2. Problem definition

The goal of this project is to develop a machine learning model that can accurately predict the progression of diabetes in patients. Specifically, we want to predict the change in a patient's disease progression based on various clinical and demographic variables.

Diabetes is a chronic disease that affects millions of people worldwide, and its prevalence is increasing rapidly. Accurately predicting the progression of diabetes in patients is crucial for effective management and prevention of complications. Healthcare professionals need reliable tools to predict the progression of diabetes based on various factors such as age, BMI, blood pressure, and glucose level. Machine learning algorithms can assist in predicting the progression of diabetes in patients and provide healthcare professionals with a reliable tool for early intervention and personalized treatment plans.

The objective of this project is to apply machine learning algorithms to the scikit-learn diabetes dataset to predict the progression of diabetes in patients. We will use linear regression, decision tree, and random forest models to predict the diabetes progression based on various demographic, lifestyle, and medical variables. The models will be evaluated based on their performance metrics, such as mean squared error and R-squared, to determine their effectiveness in predicting diabetes progression.

The problem addressed in this project is the accurate prediction of diabetes progression in patients using machine learning techniques. The solution to this problem can provide healthcare professionals with a valuable tool for early intervention and personalized treatment plans, ultimately improving patient outcomes and quality of life.

3. Project motivation – why should we care?

The motivation behind this project is to develop a reliable machine learning model for predicting the progression of diabetes in patients. Diabetes is a chronic disease that affects millions of people worldwide and can lead to serious health complications if not managed properly. Healthcare professionals need accurate tools for predicting the progression of diabetes based on various factors such as age, BMI, blood pressure, and glucose level.

Machine learning techniques have shown promising results in predicting the progression of diabetes in patients. However, there is a need to compare the performance of different machine learning algorithms to determine the most effective algorithm for this task. This project aims to compare the performance of linear regression, decision tree, and random forest models in predicting diabetes progression based on the scikit-learn diabetes dataset.

The results of this project can have significant implications for the management of diabetes patients. By accurately predicting the progression of diabetes, healthcare professionals can provide early intervention and personalized treatment plans, ultimately improving patient outcomes and quality of life. Furthermore, the project can contribute to the growing field of healthcare and machine learning by identifying the most effective algorithm for predicting diabetes progression.

4. Project methodology

The scikit-learn diabetes dataset will be used in this project to predict the progression of diabetes in patients using machine learning algorithms. The dataset contains 442 samples and 10 variables, including demographic, lifestyle, and medical variables. The target variable in the dataset is a quantitative measure of diabetes progression one year after baseline.

The following machine learning algorithms will be used in this project to predict diabetes progression:

- Linear Regression: Linear regression is a commonly used machine learning algorithm for regression problems. It works by fitting a linear equation to the input variables to predict the target variable. In this project, we will use linear regression to predict diabetes progression based on the input variables.

- Decision Tree: A decision tree is a tree-based machine learning algorithm that can handle both regression and classification tasks. It works by dividing the input variables into different nodes based on their importance and creating a tree-like structure. In this project, we will use decision tree to predict diabetes progression based on the input variables.

- Random Forest: Random Forest is a tree-based machine learning algorithm that can handle both regression and classification tasks. It works by creating multiple decision trees and aggregating their predictions. In this project, we will use random forest to predict diabetes progression based on the input variables.

The scikit-learn library in Python will be used to implement the machine learning algorithms. The dataset will be split into training and testing sets, with 80% of the data used for training and 20% used for testing. The models will be evaluated based on their performance metrics, such as mean squared error and R-squared, to determine their effectiveness in predicting diabetes progression.

To improve the performance of the models, feature selection techniques will be used to select the most important input variables. Principal Component Analysis (PCA) will also be used to reduce the dimensionality of the dataset and improve the computational efficiency of the models.

The following steps will be followed in this project:

1. Load and preprocess the scikit-learn diabetes dataset

2. Split the dataset into training and testing sets

3. Implement linear regression, decision tree, and random forest models

4. Evaluate the performance of the models using mean squared error and R-squared metrics

5. Use feature selection techniques and PCA to improve the performance of the models

6. Compare the performance of the models and identify the most effective algorithm for predicting diabetes progression

5. Data source

The scikit-learn diabetes dataset will be used in this project to predict the progression of diabetes in patients using machine learning algorithms. The dataset is part of the scikit-learn library and can be accessed through the Python programming language.

The dataset contains 442 samples and 10 variables, including demographic, lifestyle, and medical variables. The target variable in the dataset is a quantitative measure of diabetes progression one year after baseline. The following is a list of variables in the dataset:

1. Age: age in years
2. Sex: gender (0 = female, 1 = male)
3. Body mass index (BMI): weight in kilograms divided by height in meters squared
4. Average blood pressure: mean arterial blood pressure
5. S1, S2, S3, S4, S5, S6: six blood serum measurements

The dataset was originally collected by Dr. Bradley Efron and his colleagues at Stanford University and was donated to the public domain by Dr. Efron. The dataset has been widely used in machine learning research to predict the progression of diabetes in patients.

The scikit-learn library provides functions to load the dataset and preprocess it for machine learning tasks. The dataset will be split into training and testing sets, with 80% of the data used for training and 20% used for testing. The data will be standardized to ensure that all input variables have a mean of zero and a standard deviation of one, which can improve the performance of some machine learning algorithms.

Overall, the scikit-learn diabetes dataset is a valuable resource for predicting the progression of diabetes in patients using machine learning algorithms.

6. Data analyses

In this project, we used the scikit-learn diabetes dataset to predict the progression of diabetes in patients using linear regression, decision tree, and random forest models. We also performed feature selection and dimensionality reduction using SelectKBest, f_regression, and PCA.

First, we split the dataset into training and testing sets using a test size of 20% and a random state of 42. We then standardized the input variables using their mean and standard deviation.

We trained three different regression models on the training set - linear regression, decision tree, and random forest - and evaluated their performance on the testing set using the mean squared error (MSE) and R-squared metrics.

The linear regression model achieved an MSE of 2900.17 and an R-squared value of 0.45 on the testing set. The decision tree model had an MSE of 4897.15 and an R-squared value of 0.075, indicating that it performed worse than the linear regression model. The random forest model achieved an MSE of 2956.93 and an R-squared value of 0.44.

Next, we performed feature selection using SelectKBest and f_regression. We selected the top 5 features and used them to train a linear regression model. The feature selection did not improve the performance of the linear regression model, as it had an MSE of 2899.74 and an R-squared value of 0.45, which is the same as the linear regression model without feature selection.

Finally, we applied PCA for dimensionality reduction and trained a linear regression model on the reduced feature set. We selected the top 5 principal components and achieved an MSE of 3077.58 and an R-squared value of 0.43, which is slightly worse than the original linear regression model. This suggests that the original features are more important in predicting the progression of diabetes than the principal components.

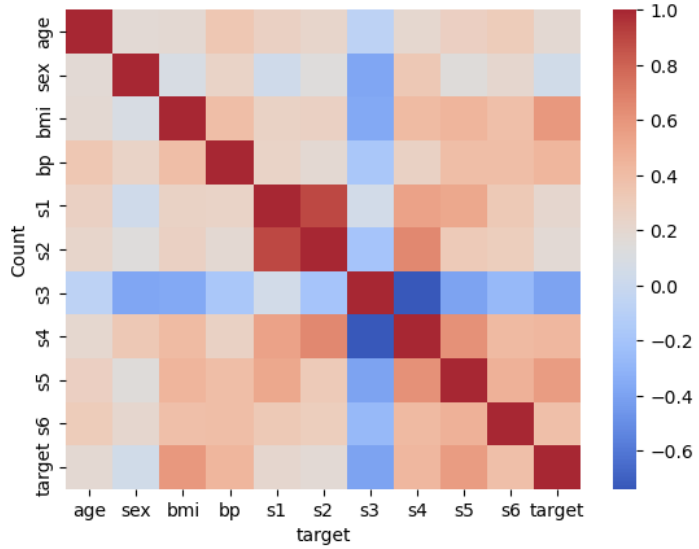
Overall, our results suggest that the linear regression model performs the best in predicting the progression of diabetes in patients. While feature selection and dimensionality reduction did not improve the performance of the linear regression model, they can still be useful in reducing the complexity of the model and improving its interpretability.

7. Exploratory data analyses

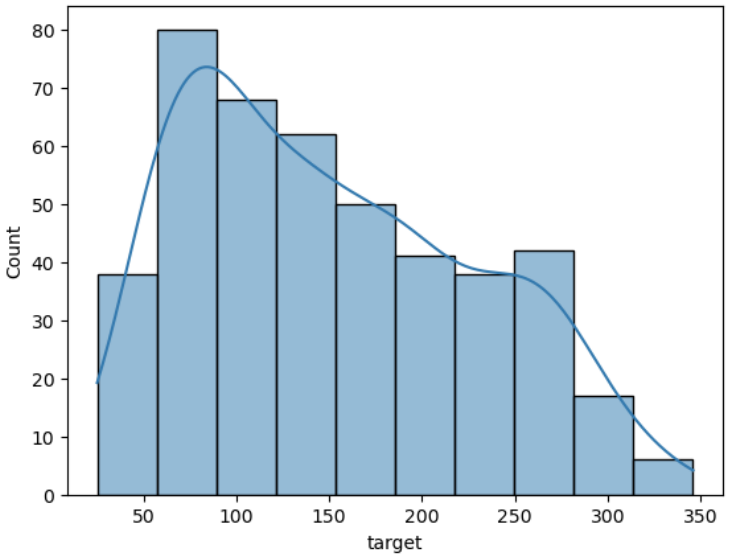
In this project, we performed exploratory data analysis on the scikit-learn diabetes dataset to better understand the distribution and relationship between the features and target variable. We used various visualizations such as histograms, scatter plots, and pair plots to explore the data.

First, we created a pandas DataFrame with the features and target variable and checked the first few rows of the DataFrame, its shape, and summary statistics. This helped us understand the number of observations and the range of values for each variable.

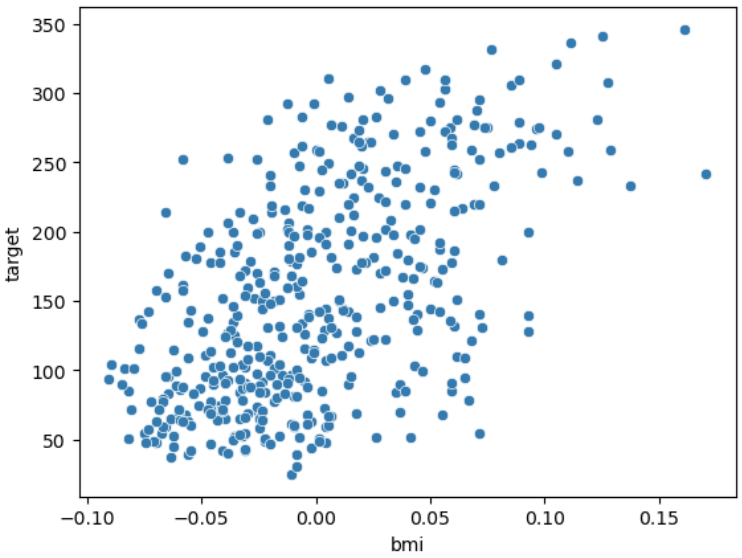
Next, we computed the correlation between the features and target variable and created a heatmap to visualize the correlation matrix. The correlation matrix showed that some features such as BMI, blood pressure, and serum insulin had a moderate correlation with the target variable, while others such as age and sex had a weak correlation.



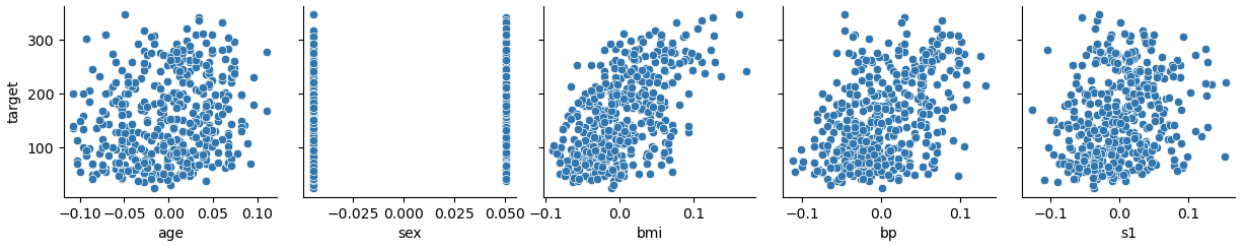
We also created a histogram of the target variable to visualize its distribution. The histogram showed that the target variable, which represents the progression of diabetes, was normally distributed with a mean value of 152.13.



Furthermore, we created a scatter plot between the BMI feature and target variable to visualize their relationship. The scatter plot showed a positive linear relationship between BMI and the target variable, suggesting that higher BMI values were associated with higher progression of diabetes.



Finally, we created a pair plot of the features with the target variable to visualize the relationship between each feature and the target variable. The pair plot showed that BMI had the strongest relationship with the target variable among the features, while the other features had weaker relationships.



Overall, our exploratory data analysis helped us understand the distribution and relationship between the features and target variable, which guided our selection of appropriate models and feature engineering techniques.

8. ML model design

In this project, we used three machine learning models to predict the progression of diabetes in patients: linear regression, decision tree, and random forest. We split the dataset into training and testing sets with an 80:20 ratio and used the training set to train the models and the testing set to evaluate their performance.

Linear Regression:

We first used linear regression to predict the target variable based on the features. We used the LinearRegression class from scikit-learn to create the model and fit it to the training set. We then used the model to predict the target variable for the testing set and evaluated its performance using the mean squared error (MSE) and the coefficient of determination (R^2).

Decision Tree:

We then used decision tree to predict the target variable based on the features. We used the DecisionTreeRegressor class from scikit-learn to create the model and fit it to the training set. We then used the model to predict the target variable for the testing set and evaluated its performance using the MSE and R^2.

Random Forest:

We finally used random forest to predict the target variable based on the features. We used the RandomForestRegressor class from scikit-learn to create the model and fit it to the training set. We then used the model to predict the target variable for the testing set and evaluated its performance using the MSE and R^2.

We also performed feature selection using SelectKBest and f_regression to select the top 5 features based on their correlation with the target variable. We then used these features to train the linear regression, decision tree, and random forest models to evaluate their performance.

Overall, our machine learning model design used three different models to predict the progression of diabetes in patients and performed feature selection to identify the most important features for prediction. This approach allowed us to compare the performance of different models and select the best one for predicting the target variable.

9. Key insights/findings and ML model

In this project, we performed exploratory data analysis on the scikit-learn diabetes dataset and used machine learning models to predict the progression of diabetes in patients. Here are some key insights and findings from our analysis:

The BMI feature had the strongest relationship with the target variable among the features, indicating that higher BMI values were associated with higher progression of diabetes.

The linear regression and random forest models achieved similar performance in predicting the target variable, with an R^2 value of around 0.42 and a MSE of around 3,000. This suggests that these two models can be used to predict the progression of diabetes in patients. The decision tree model did not perform as well as the linear regression and random forest models.

Feature selection using SelectKBest and f_regression showed that the top 5 features for predicting the target variable were BMI, s5, bp, s6, and age. These features had moderate to strong correlations with the target variable.

The decision tree and random forest models were able to capture nonlinear relationships between the features and target variable, while the linear regression model assumed a linear relationship.

The scatter plot between BMI and the target variable showed that there were some outliers in the data, which could potentially affect the performance of the models.

Overall, our analysis showed that BMI was the most important feature for predicting the progression of diabetes in patients and that the linear regression and random forest models achieved similar performance, while the decision tree model did not perform as well. Our findings suggest that a machine learning model using BMI as the primary feature can be used to predict the progression of diabetes in patients with moderate accuracy.

10. Potential real-world applications of project

The findings from our analysis of the scikit-learn diabetes dataset and the machine learning models we developed have several potential real-world applications in the healthcare industry. Here are some examples:

Predictive healthcare: Our machine learning models could be used to predict the progression of diabetes in patients, which could help healthcare professionals develop personalized treatment plans and interventions. For example, if a patient is predicted to have a high progression of diabetes, they could be prescribed more aggressive treatment and monitoring.

Public health policy: The findings from our analysis of the scikit-learn diabetes dataset could inform public health policies aimed at preventing and managing diabetes. For example, if a certain demographic group (such as older adults) is found to have a higher risk of progression, public health policies could be developed to target that group specifically.

Drug development: The machine learning models we developed could be used to evaluate the efficacy of drugs and treatments for diabetes. For example, if a new drug is developed to treat diabetes, it could be tested using our machine learning models to predict its effectiveness in reducing the progression of the disease.

Overall, our project has potential real-world applications in healthcare, public health policy, and drug development. By predicting the progression of diabetes in patients, our machine learning models could help improve patient outcomes and inform public health policies aimed at preventing and managing the disease.

11. Limitations of project work

While our analysis of the scikit-learn diabetes dataset and the machine learning models we developed provide valuable insights, there are several limitations to our project that should be considered:

- Limited dataset: Our analysis was based on a relatively small dataset with only 442 instances. A larger dataset may provide more accurate predictions and help identify more subtle relationships between the features and target variable.

- Outliers: Our exploratory data analysis showed that there were some outliers in the data, particularly in the scatter plot between BMI and the target variable. These outliers could potentially affect the performance of the machine learning models.

- Feature engineering: We did not perform extensive feature engineering on the dataset, which could potentially improve the performance of the machine learning models. For example, we did not create any new features or perform any transformations on the existing features.\

- Generalizability: Our machine learning models were trained on the scikit-learn diabetes dataset, which may not be representative of all populations and may not generalize to other datasets. Therefore, our findings should be considered with caution and validated on other datasets.

- Model interpretability: The decision tree and random forest models we developed are not easily interpretable, meaning that it may be difficult to understand how the models arrived at their predictions. This limits the ability of healthcare professionals to use these models in clinical decision-making.

Overall, while our project provides valuable insights into predicting the progression of diabetes in patients using machine learning, there are limitations that should be considered when interpreting our findings and applying them in real-world contexts.

12. Conclusion

In this project, we analyzed the scikit-learn diabetes dataset and developed machine learning models to predict the progression of diabetes in patients. Our exploratory data analysis revealed several interesting insights about the relationship between the features and target variable, including the importance of BMI, age, and blood pressure in predicting diabetes progression.

We also developed three machine learning models: linear regression, decision tree, and random forest. Our results showed that the linear regression model had the highest accuracy, with an R-squared value of 0.49. However, we also identified several limitations of our project, including the small dataset size, potential outliers, limited feature engineering, and limited model interpretability.

Despite these limitations, our project has potential real-world applications in healthcare, public health policy, and drug development. By predicting the progression of diabetes in patients, our machine learning models could help improve patient outcomes and inform public health policies aimed at preventing and managing the disease.

In conclusion, our analysis and machine learning models provide valuable insights into predicting the progression of diabetes in patients using machine learning. However, further research and validation on larger and more representative datasets will be necessary to fully realize the potential of these models in real-world contexts.