University of Essex

CE802-7-SP Machine Learning

Assignment Report

Total Word Count-1414

Priyanshu Banerjee

MSc AI

2211886

Content

Introduction

In this study, we investigate the performance of various machine learning procedures on two datasets for predicting diabetes risk and estimating average blood glucose levels. Our goal is to assess the relative strengths and weaknesses of the methods under different conditions. The datasets consist of several features, such as demographic information, medical history, and lifestyle factors, that are known to be associated with the risk of developing diabetes and elevated blood glucose levels.

Task A involves predicting the risk of developing diabetes using classification techniques, while Task B focuses on estimating average blood glucose levels, specifically identifying instances where the average blood glucose level exceeds the diagnostic threshold, using regression techniques.

We perform experiments to fill in missing data, apply Principal Component Analysis (PCA), and implement different machine learning techniques. Our evaluation metrics for Task A (classification) include accuracy, F1-score, recall, and precision. For Task B (regression), we use evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R2) score.

Throughout this study, we aim to identify the most effective machine learning models for predicting diabetes risk and estimating average blood glucose levels, as well as provide insights into the impact of data pre-processing techniques, such as filling in missing data and applying PCA, on the performance of the models. Our findings will be helpful in guiding future applications of these methods on similar datasets and providing a better understanding of their strengths and limitations, ultimately contributing to more accurate and efficient diabetes risk prediction and blood glucose level estimation models.

Methods

- **Task A Methods**
  In Task A, our goal was to predict the risk of developing diabetes using classification techniques. We began our investigation by filling in the missing data with zeros as a baseline approach. We then applied the following machine learning models to the dataset: Random Forest, Gaussian Naive Bayes, XGBoost, SVM, and KNN.
  Next, we used KNN imputation to fill in the missing data, providing a more sophisticated approach for handling missing values. We re-evaluated the performance of the same machine learning models with the KNN-imputed dataset to compare their effectiveness in predicting diabetes risk under different data pre-processing conditions.

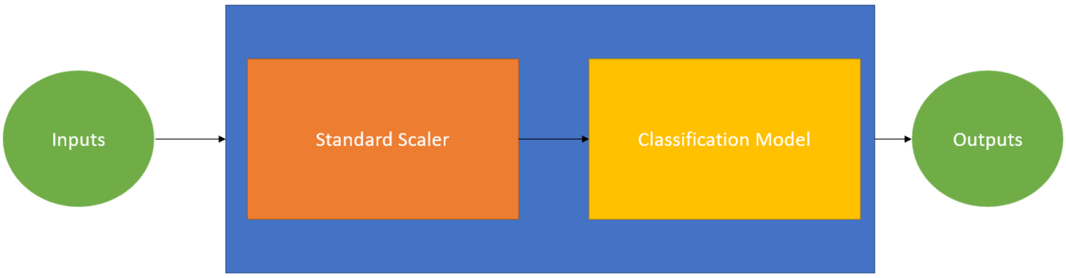

Figure 1 Model Pipeline for Task A

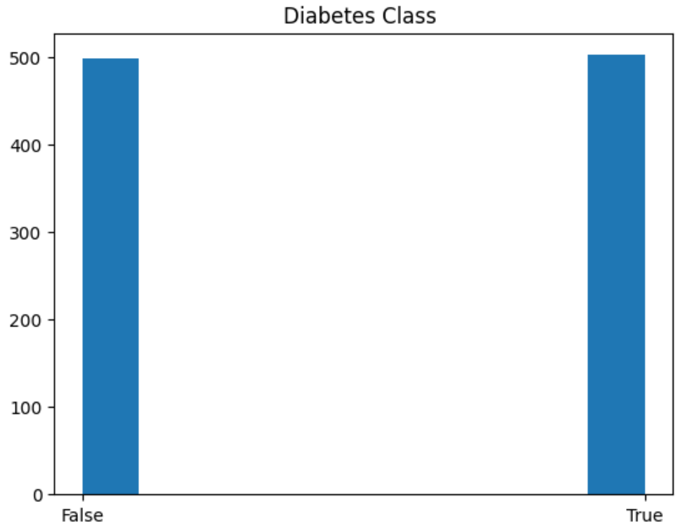

Figure 2 Class Distribution (Balanced dataset)

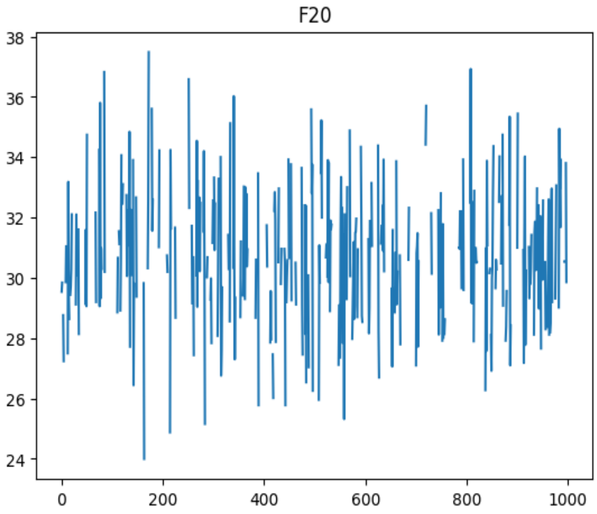

Figure 3 F20 column before pre-processing

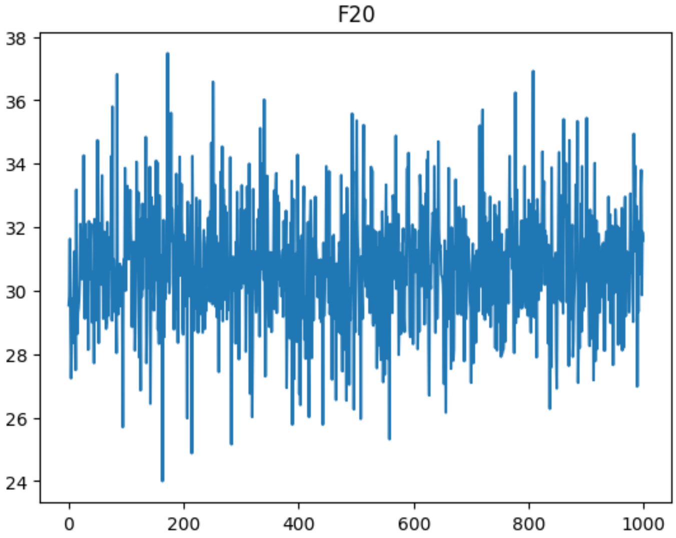

Figure 4 F20 Column after Pre-processing using KNN.

- **Task B Method**

  In Task B, our objective was to estimate average blood glucose levels and identify instances where the average blood glucose level exceeds the diagnostic threshold using regression techniques. For this task, we kept the data unchanged and implemented the following regression models as a baseline approach: Linear Regression, Decision Tree Regressor, Random Forest Regressor, Lasso, Ridge, and Gradient Boosting Regressor.

  Next, we applied Principal Component Analysis (PCA) to the dataset, a dimensionality reduction technique that can help improve model performance in some cases. We re-evaluated the performance of the regression models mentioned above using the PCA-transformed dataset to see how it compares to the baseline.

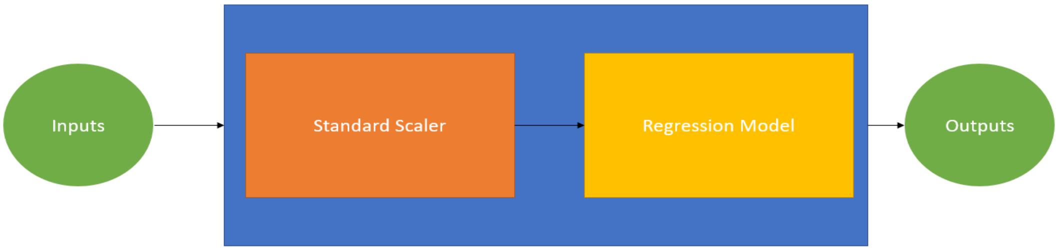

Figure 5 Model Pipeline for Task B

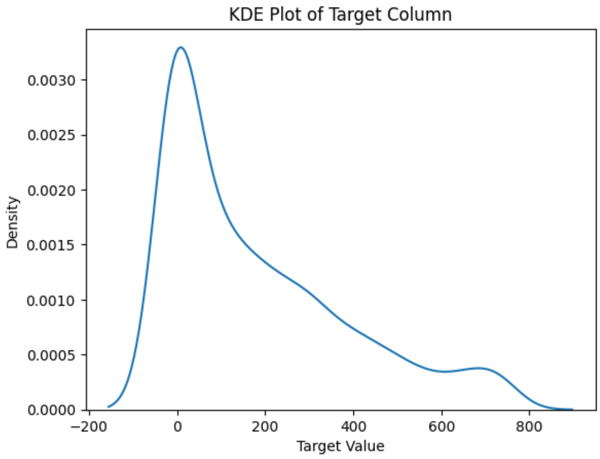

Figure6 KDE Plot

By comparing the performance of various machine learning models under different data pre-processing techniques, we aim to identify the most effective models and pre-processing strategies for predicting diabetes risk (Task A) and estimating average blood glucose levels (Task B).

Results

Task A

From the above figure, we can see that we have a balanced dataset, which helps ensure a more accurate representation of the real-world problem. In such cases, accuracy is an appropriate metric to evaluate the performance of the models, as it represents the proportion of correct predictions out of the total predictions.

We have selected accuracy as the main evaluation metric for our models due to the following reasons:

• Balanced dataset: In our dataset, the classes are not biased or unbalanced, and all labels have equal weight. This means that there is no significant difference in the number of instances for each class. In such cases, accuracy serves as an appropriate metric to evaluate the model's performance, as it represents the proportion of correct predictions out of the total predictions.

• Simple and intuitive: Accuracy is a straightforward and easily interpretable metric. It gives a clear idea of how well the model is performing by providing the percentage of correct predictions.

• Suitable for classification problems: Accuracy is a widely used metric in classification problems where the objective is to assign an instance to one of the given classes. In our case, we are dealing with a binary classification problem, and accuracy helps us understand the overall effectiveness of the model in classifying the data correctly.

Even though our dataset is balanced, examining the F1-score can still give us valuable insights into the performance of our models. F1-score is the harmonic mean of precision and recall, and it provides a single metric that considers both false positives and false negatives. By examining the F1-score alongside accuracy, we can gain a more comprehensive understanding of our models' performance and their ability to correctly predict high-risk cases and identify instances where average blood glucose levels exceed the diagnostic threshold.

Table 1

Model	Accuracy	F1-score	Recall	Precision
Random Forest Model	90.00	0.907407	0.907407	0.907407
Gaussian Naive Bayes	61.00	0.723404	0.944444	0.586207
XGBoost Model	92.00	0.926606	0.935185	0.918182
SVM Model	78.50	0.790244	0.750000	0.835052
KNN Model	69.50	0.710900	0.694444	0.728155

Table 2

Model	Accuracy	F1-score	Recall	Precision
Random Forest Model (Baseline)	91.00	0.917431	0.925926	0.909091
Gaussian Naive Bayes(baseline)	52.00	0.671233	0.907407	0.532609
XGBoost Model (Baseline)	92.50	0.930876	0.935185	0.926606
SVM Model (Baseline)	74.00	0.745098	0.703704	0.791667
KNN Model (Baseline)	69.50	0.710900	0.694444	0.728155

Based on the results, we can observe the following:

- XGBoost model consistently performs the best in terms of accuracy, F1-score, recall, and precision when applied to the given dataset, irrespective of the missing data imputation technique.

- The Gaussian Naive Bayes model shows the lowest performance for both imputation methods. This may be due to the assumption of feature independence, which might not hold for this dataset.

- The performance of the other models (Random Forest, SVM, and KNN) varies depending on the imputation technique, with the Random Forest model showing relatively stable performance across both methods.

Overall, it is important to note that the choice of data pre-processing, such as missing data imputation, can have a significant impact on the performance of machine learning models.

Task B

For the regression task, we evaluated the performance of the models using MAE, MSE, RMSE, and R2. Below are the results for both the baseline (unchanged data) and the PCA-applied dataset:

Table 3

Model	MAE	MSE	RMSE	R2
Linear Regression (PCA)	131.881283	32995.421323	181.646418	0.237620
Decision Tree Regressor (PCA)	190.109679	70248.916702	265.045122	-0.623145
Random Forest Regressor (PCA)	135.427045	33069.722412	181.850825	0.235904
Lasso (PCA)	131.875304	32866.498692	181.291199	0.240599
Ridge (PCA)	131.883062	32993.022808	181.639816	0.237676
Gradient Boosting Regressor (PCA)	136.785859	34887.012448	186.780653	0.193914

Table 4

Model	MAE	MSE	RMSE	R2
Linear Regression (Baseline)	118.396816	22921.541439	151.398618	0.470384
Decision Tree Regressor (Baseline)	115.081464	26983.882053	164.267715	0.376521
Random Forest Regressor (Baseline)	88.694853	14640.327259	120.997220	0.661726
Lasso (Baseline)	117.676838	22692.254373	150.639485	0.475681
Ridge (Baseline)	118.363558	22915.984017	151.380263	0.470512
Gradient Boosting Regressor (Baseline)	68.962375	8664.529614	93.083455	0.799801

Based on the results, we can observe the following:

Gradient Boosting Regressor performs the best in terms of MAE, MSE, RMSE, and R2 when applied to the baseline dataset.

Applying PCA to the dataset negatively impacts the performance of all regression models, as indicated by increased MAE, MSE, and RMSE, and decreased R2 values.

The performance of the models varies significantly depending on the data pre-processing method (PCA or unchanged data). This highlights the importance of feature selection and dimensionality reduction techniques in the model development process.

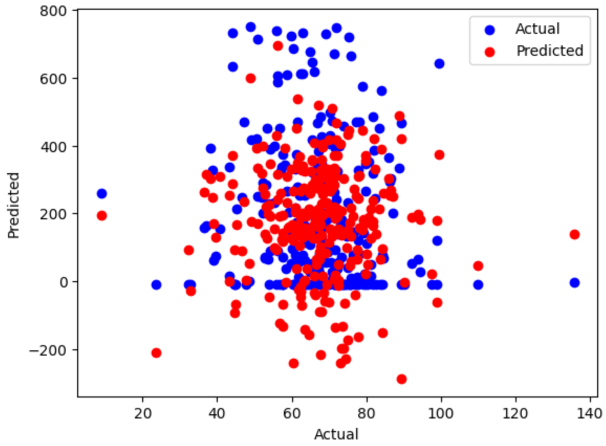

Figure 7 Linear Regression (Baseline) Actual vs predictions

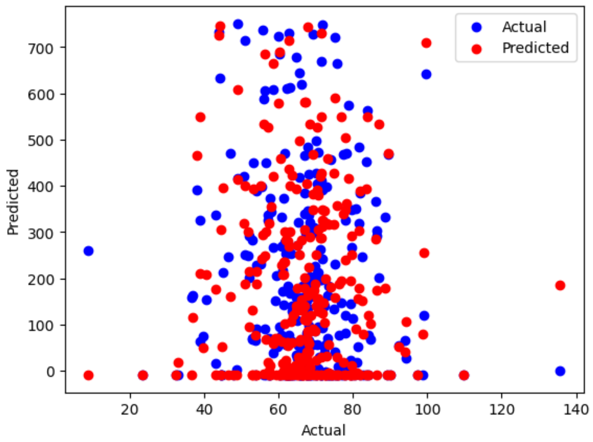

Figure 8 Decision Tree Regressor (Baseline) Actual vs predictions

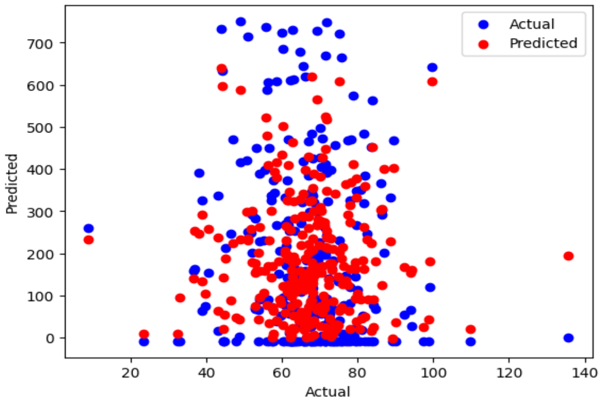

Figure 9 Random Forest Regressor (Baseline) Actual vs predictions

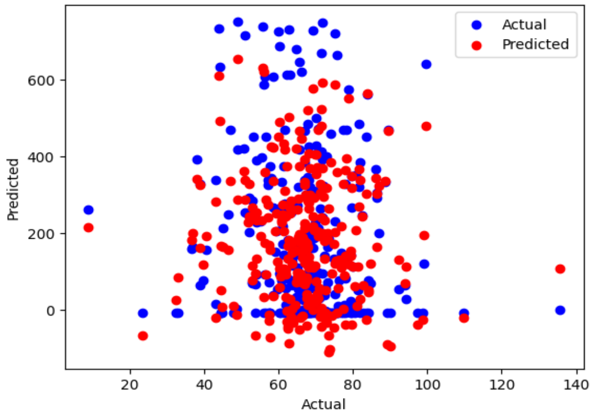

Figure 10 Lasso (Baseline) Actual vs predictions

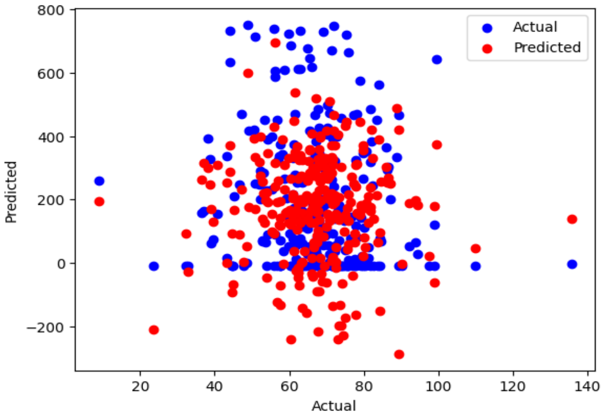

Figure 11 Ridge (Baseline) Actual vs predictions

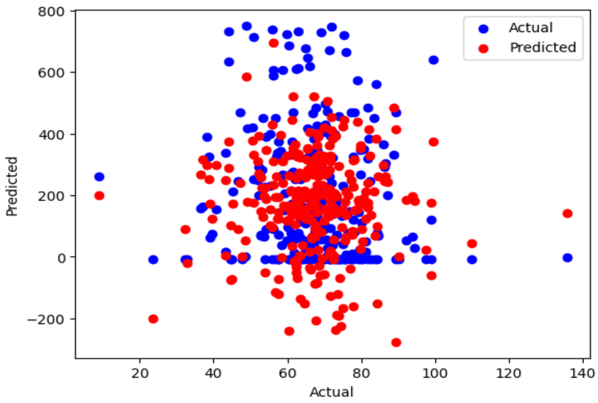

Figure 12 Gradient Boosting Regressor (Baseline) Actual vs predictions

Conclusion.

In conclusion, the choice of pre-processing techniques and machine learning models plays a critical role in determining the performance of predictive models. Considering the specific dataset and problem at hand is essential when selecting suitable data pre-processing techniques and machine learning models to ensure optimal performance.

For Task A, our experiments showed excellent results, with the XGBoost model consistently performing the best. However, there is always room for improvement, and future studies can explore other feature selection methods, pre-processing techniques, and ensemble learning methods to further enhance the model's performance.

For Task B, the results indicate that there is more room for improvement in the regression models. In future studies, we can focus on exploring other feature selection methods, pre-processing techniques, and ensemble learning methods to further improve the performance of the predictive models. Moreover, we can investigate alternative regression models and techniques to better capture the underlying patterns in the data, which could lead to more accurate predictions for average blood glucose levels that exceed the diagnostic threshold.