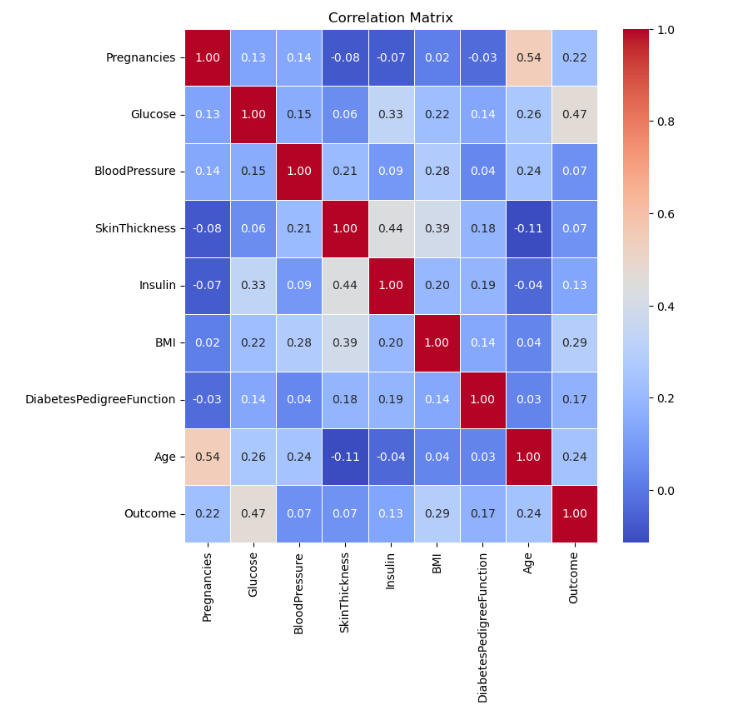
Table 4. Statistical summary of dataset features

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variables** | **Count** | **Mean** | **Std** | **Min** | **25%** | **50%** | **75%** | **Max** |
| Pregnancies | 768.0 | 3.845052 | 3.369578 | 0.000 | 1.00000 | 3.0000 | 6.00000 | 17.00 |
| Glucose | 768.0 | 120.894531 | 31.972618 | 0.000 | 99.00000 | 117.0000 | 140.25000 | 199.00 |
| BloodPressure | 768.0 | 69.105469 | 19.355807 | 0.000 | 62.00000 | 72.0000 | 80.00000 | 122.00 |
| SkinThickness | 768.0 | 20.536458 | 15.952218 | 0.000 | 0.00000 | 23.0000 | 32.00000 | 99.00 |
| Insulin | 768.0 | 79.799479 | 115.244002 | 0.000 | 0.00000 | 30.5000 | 127.25000 | 846.00 |
| BMI | 768.0 | 31.992578 | 7.884160 | 0.000 | 27.30000 | 32.0000 | 36.60000 | 67.10 |
| DiabetesPedigreeFunction | 768.0 | 0.471876 | 0.331329 | 0.078 | 0.24375 | 0.3725 | 0.62625 | 2.42 |
| Age | 768.0 | 33.240885 | 11.760232 | 21.000 | 24.00000 | 29.0000 | 41.00000 | 81.00 |
| Outcome | 768.0 | 0.348958 | 0.476951 | 0.000 | 0.00000 | 0.0000 | 1.00000 | 1.00 |

Correlation matrix

The correlation matrix provides insights into the relationships between different features in the dataset. Notably, the "Age" feature exhibits a strong positive correlation with "Pregnancies" (0.54) and a moderate positive correlation with "Glucose" (0.26). Additionally, "Glucose" demonstrates a moderate positive correlation with "BMI" (0.22) and a strong positive correlation with the target variable, "Outcome" (0.47). Conversely, the "SkinThickness" feature shows a negative correlation with "Pregnancies" (-0.08) and "Age" (-0.11), suggesting an inverse relationship. The "BloodPressure" feature exhibits relatively low correlations with other variables in the dataset. Examining the correlation with the target variable, "Outcome," reveals that "Age" has a moderate positive correlation (0.24), while "Glucose" demonstrates a strong positive correlation (0.47). These findings can guide feature selection, indicating potential associations that may influence the predictive performance of machine learning models. It's important to note that correlation does not imply causation, and further analysis, along with domain knowledge, is essential for a comprehensive understanding of the dataset.

**Summary of Algorithm's Accuracy**

|  |  |
| --- | --- |
| **Algorithm** | **Accuracy Score (%)** |
| Random Forest Classifier | 75.52 |
| Decision Tree Classifier | 71.35 |
| XGBoost Classifier | 73.96 |
| KNeighborsClassifier | 72.92 |
| Logistic Regression | 77.08 |

The analysis of the accuracy scores for the various machine learning algorithms provides valuable insights into their performance on the given dataset. The Random Forest Classifier achieved a respectable accuracy of 75.52%, showcasing its capability to accurately predict outcomes in approximately three-quarters of cases. Renowned for its robustness, this algorithm excels at handling complex relationships within the data.

On the other hand, the Decision Tree Classifier exhibited an accuracy of 71.35%. While this performance is reasonable, it suggests that the model might struggle to capture intricate patterns compared to more sophisticated algorithms. Decision trees, being susceptible to overfitting, may require additional regularization.

XGBoost Classifier, a powerful boosting algorithm, achieved an accuracy of 73.96%. Known for striking a balance between model complexity and predictive accuracy, XGBoost's performance could potentially be further improved through hyperparameter tuning.

The K-Nearest Neighbors Classifier attained an accuracy of 72.92%. Its effectiveness depends on parameter choices, such as the number of neighbors ('k') and the distance metric. Fine-tuning these parameters could enhance its predictive capability.

Surpassing the other algorithms, Logistic Regression demonstrated an accuracy of 77.08%. Particularly well-suited for binary classification problems, Logistic Regression appeared to effectively model the relationships present in the data

**Summary of Algorithm's Classification Report**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Algorithm** | Target | Precision | Recall | F1-Score |
| Random Forest Classifier | 0 | 0.76 | 0.88 | 0.82 |
| 1 | 0.73 | 0.54 | 0.62 |
| Decision Tree Classifier | 0 | 0.75 | 0.83 | 0.78 |
| 1 | 0.64 | 0.52 | 0.57 |
| XGBoost Classifier | 0 | 0.76 | 0.86 | 0.81 |
| 1 | 0.69 | 0.54 | 0.60 |
| KNeighborsClassifier | 0 | 0.75 | 0.85 | 0.80 |
| 1 | 0.67 | 0.52 | 0.59 |
| Logistic Regression | 0 | 0.77 | 0.92 | 0.83 |
| 1 | 0.79 | 0.52 | 0.63 |

The precision, recall, and F1-score provide a more nuanced evaluation of the models' performance, particularly concerning their ability to correctly predict positive and negative cases.

For the Random Forest Classifier, it demonstrated solid precision for both classes (0 and 1), with higher precision for class 0 (0.76) compared to class 1 (0.73). However, it showed a trade-off between precision and recall, as it achieved higher recall for class 0 (0.88) than for class 1 (0.54). The F1-score, which balances precision and recall, was generally good, with the model performing better for class 0.

The Random Forest Classifier exhibited a balanced performance, with reasonably high precision for predicting non-diabetic cases (class 0) and a moderate level of precision for diabetic cases (class 1). The model demonstrated a good ability to correctly identify non-diabetic cases (high recall for class 0), but it had a lower recall for diabetic cases (class 1). The F1-score, which considers both precision and recall, indicated a favorable balance, especially for predicting non-diabetic cases.

The Decision Tree Classifier exhibited a similar pattern, with higher precision and recall for class 0 compared to class 1. The F1-score, representing a harmonic mean of precision and recall, indicated a balanced performance for class 0 but a lower balance for class 1. Similar to the Random Forest, the Decision Tree Classifier showed a tendency to perform better in predicting non-diabetic cases (class 0) compared to diabetic cases (class 1). The model had a balanced F1-score for class 0 but a slightly lower balance for class 1.

XGBoost Classifier showed competitive precision and recall values, with slightly higher precision for class 0. Like the Random Forest Classifier, it demonstrated a trade-off between precision and recall for class 1. The F1-scores for both classes were relatively balanced. The XGBoost Classifier demonstrated competitive precision and recall for both classes, indicating a well-balanced performance. The model achieved a good compromise between precision and recall for both diabetic and non-diabetic cases, as reflected in the balanced F1-scores for both classes

K-Nearest Neighbors Classifier displayed good precision for both classes but, similar to the other models, had higher precision for class 0. It also showed a trade-off between precision and recall for class 1. The K-Nearest Neighbors model performed well, with good precision for both classes. However, similar to other models, it had a higher precision for predicting non-diabetic cases (class 0). The model demonstrated a trade-off between precision and recall for diabetic cases (class 1).

Logistic Regression exhibited the highest precision and recall for both classes among the models, making it particularly effective in this binary classification problem. The F1-scores for both classes were also well-balanced, indicating a strong overall performance. Logistic Regression emerged as the top-performing model in this evaluation. It exhibited the highest precision and recall for both diabetic and non-diabetic cases. The model's F1-scores for both classes were well-balanced, indicating a strong overall performance in correctly classifying diabetes outcomes.