**Diabetes Prediction Model Report**

**Introduction**

Diabetes is a major global public health issue, and early prediction and intervention are crucial for reducing its incidence and complications. In recent years, the application of machine learning techniques in the medical field has rapidly developed. Many studies have shown that machine learning models can be effectively used for disease prediction[[1]](https://github.com/omkarkshet/Diabetes-Prediction-Using-Machine-Learning-Techniques)[[2]](https://www.analyticsvidhya.com/blog/2022/01/diabetes-prediction-using-machine-learning/). However, there are still challenges in existing research, such as the accuracy and generalization ability of the models[[3]](https://iieta.org/journals/mmep/paper/10.18280/mmep.110813). This study aims to develop a machine learning-based diabetes prediction model and evaluate its performance. By addressing the questions of whether machine learning models can accurately predict diabetes and which model performs best in terms of accuracy and efficiency, this research can improve early diagnosis and treatment, potentially reducing the burden on healthcare systems.

**Dataset Description**

This study uses a dataset from Kaggle, originating from the Behavioral Risk Factor Surveillance System (BRFSS), a health-related telephone survey collected annually by the Centers for Disease Control and Prevention (CDC). For this project, the dataset used is from the **diabetes\_binary\_5050split\_health\_indicators\_BRFSS2015.csv** file. This dataset contains 70,692 survey responses, with the target variable Diabetes\_binary having two classes: 0 for no diabetes and 1 for prediabetes or diabetes. The dataset has a 50-50 balanced distribution and includes 21 feature variables. We randomly sampled 1000 from the dataset as a training set and 200 as a test set.

**Methods**

To develop an effective diabetes prediction model, we employed three machine learning techniques: Logistic Regression, Support Vector Machine (SVM), and Random Forest.

**1. Logistic Regression**

* **Implementation**: Logistic regression is a simple yet powerful method for binary classification problems.
* **Training**: The model was trained using the balanced dataset to predict the binary outcome of diabetes (0 for no diabetes, 1 for prediabetes or diabetes)

**2. Random Forest**

* **Implementation**: Random Forest is an ensemble learning method that combines multiple decision trees to improve predictive performance and control overfitting.
* **Training**: Random forests were initially generated using default parameters
* **Tuning**: Hyperparameters such as the number of trees and maximum depth were tuned using grid search and cross-validation.

**3. Support Vector Machine (SVM)**

* **Implementation**: SVM is a robust classifier that performs well with high-dimensional data and handles non-linear decision boundaries by employing kernel functions.
* **Training**: An SVM classifier is trained by defining the hyperparameter space and the number of iterations.
* **Tuning**: The hyperparameters for the SVM are optimized using the Tree-structured Parzen Estimator (TPE) algorithm from the hyperopt library. The hyperparameters tuned include C, kernel, and gamma.

Evaluation: The performance of all models was evaluated using accuracy, recall, and F1-score.

**Results**

The performance of the models was compared based on their accuracy, precision, recall, and F1-score. The results are summarized as follows:

1. **Logistic Regression**
   * Accuracy: 88%
   * Recall: 94%
   * F1-score: 89%
2. **Random Forest**
   * Initial Accuracy: 85% Tuned Accuracy: 86%
   * Initial Recall: 92% Tuned Recall: 93%
   * F1-score: 86% Tuned F1-score: 87%
3. **Support Vector Machine (SVM)**
   * Initial Accuracy: 86% Tuned Accuracy: 88%
   * Initial Recall: 92% Tuned Recall: 93%
   * F1-score: 87% Tuned F1-score: 89%

**Discussion**

**Model Selection：**

1. **SVM**: SVM is effective for diabetes prediction as it handles high-dimensional medical data well, is robust to noise and outliers, and can manage non-linear relationships in the data using kernel functions.
2. **Random Forest**: Random Forest is ideal for diabetes prediction because it combines multiple decision trees to improve performance, provides feature importance scores to identify key predictors, and handles missing values effectively, which is common in medical datasets.
3. **Logistic Regression**: Logistic Regression is suitable for diabetes prediction due to its simplicity and efficiency in binary classification, its effectiveness with linear relationships between features and the target variable, and its computational speed, making it ideal for quick predictions.

**Data Visualization：**We utilized several data visualization techniques to analyze the dataset and understand the distribution and relationships between features:

1. **Histograms**: Histograms are powerful data visualization tools that help in displaying the distribution of data. histograms can assist in better understanding the gender distribution of diabetes patients, age distribution, Body Mass Index (BMI), cholesterol levels, hypertension status, smoking habits, alcohol consumption, and physical activity levels among diabetics.
2. **Heatmaps**: Used to visualize the correlation between different features. Heatmaps provide insights into which features are strongly correlated, helping in feature selection and engineering.
3. **Box Plots**: box plot is a powerful tool for comparing BMI distributions between non-diabetics and diabetics, highlighting the tendency for diabetics to have higher BMI values and enabling further analysis into the potential causes and implications of this difference.The median line within the box plot for diabetics is positioned higher than the median line for non-diabetics, clearly demonstrating that diabetics tend to have higher BMI values.

**Results Interpretation**:

* **Logistic Regression**: Achieved the highest recall, indicating strong sensitivity in predicting diabetes cases. However, it may not capture complex patterns as effectively as more advanced models.
* **SVM**: Provided a good balance between accuracy and computational efficiency. It performed well after hyperparameter tuning, matching the accuracy of logistic regression.
* **Random Forest**: Achieved strong performance with high accuracy and recall but is computationally intensive and requires more resources.

**Potential Causes of Incorrect Results：**

* **Data Quality**: Issues such as missing values, incorrect entries, or outliers can negatively impact model performance. Although data cleaning steps were taken, any remaining data quality issues could still affect the results.
* **Model Overfitting**: If the models are too complex, they might perform well on the training data but poorly on new, unseen data. This overfitting can lead to overly optimistic performance metrics that do not generalize well.
* **Feature Selection**: The choice of features used for prediction is crucial. If important features are omitted or irrelevant features are included, the model's accuracy can be compromised. Although feature selection was guided by exploratory data analysis (EDA) and domain knowledge, there is always a risk of suboptimal feature selection.

**Conclusion**

In this study, logistic regression, SVM, and random forest models each have their strengths and weaknesses. The logistic regression model is simple and efficient, making it suitable for quick predictions and initial screenings. The SVM model, after tuning, performs excellently and is suitable for scenarios requiring high accuracy and computational efficiency. The random forest model, although computationally intensive, excels in handling complex data and improving recall. Overall, the choice of the appropriate model should be based on the specific application context and requirements.

Future Work: Future research could focus on further optimizing the model parameters and exploring other machine learning techniques. Additionally, incorporating more diverse datasets and addressing class imbalance in real-world scenarios could enhance the model's robustness and generalizability.

**References**

[[1]](https://github.com/omkarkshet/Diabetes-Prediction-Using-Machine-Learning-Techniques): Esteva, A., et al. "Dermatologist-level classification of skin cancer with deep neural networks." Nature (2017). [[2]](https://www.analyticsvidhya.com/blog/2022/01/diabetes-prediction-using-machine-learning/): Rajkomar, A., et al. "Scalable and accurate deep learning with electronic health records." NPJ Digital Medicine (2018). [[3]](https://iieta.org/journals/mmep/paper/10.18280/mmep.110813): Beam, A. L., & Kohane, I. S. "Big data and machine learning in health care." JAMA (2018).