**Table of Contents**

1. Problem Description
2. Literature Review
3. Descriptive analysis and visualizations
   1. Statistical summary
   2. Distributive analysis through visualizations
4. Implementing machine learning models
   1. Split the dataset
   2. Applying Logistic regression,Random forest and SVM
   3. Comparing model perpormance and selecting best model
5. Evaluation Metrics and Data Splitting Techniques
   1. Accuracy, precision, recall, F1-score, ROC-AUC
   2. The data splitting technique
6. Error Analysis and Recommendations
   1. Error analysis by examining misclassified instances or prediction errors
   2. Common reasons for misclassification
   3. Propose recommendations for improving the models
7. References
8. **Problem Description**

Diabetes is a contagious metabolic disorder that can be characterized by an increase in blood sugar level, it is causing a significant global health challenges and from past few years it is been one of the severe disease, millions of people are effected world wide and it leads to severe complications like cardiovascular disease, renal failure, vision impairment and many more diseases. The early identification of this disease may play a vital role in health care industry to save many lives and as well as save people from expensive treatments. Effective prediction of diabetes risk entails a comprehensive analysis of diverse factors, including demographic attributes, lifestyle data, medical history and biomarkers like glucose level, insulin sensitivity and body mass index as the data or information for every organism varies because real world data is always nonlinear. Traditional risk assessment techniques rely on the clinical symptoms only with no proper medical history and life style data this causes assessment techniques have limitations in accuracy and scalability. The integration of machine learning and artificial intelligence can cause enhancement in the accuracy and scalability in diabetes risk prediction. As the machine learning techniques are advance techniques to gain best accuracy performance in classification and regression tasks. Using an extensive dataset and analyzing statistical patterns within them, machine learning algorithms can learn this patterns and train a predictive model over that dataset and prepare a predictive model for the prediction of early diabetes risk. The purpose of integration of machine learning techniques into diabetes risk prediction is not only to increase the efficacy of predictions but this approach may helps in decision making for the health care industry to take steps to make more advancement by generating more complex and extensive data of patients medical history that may relate to their genetic so that risk assessment will be even more advanced.

**2. Literature Review**

A vast body of research explores the use of machine learning (ML) techniques for predicting the early risk of diabetes, a condition often triggered by extreme metabolic imbalances. Past studies have delved into a wide range of methodologies, aiming to achieve the highest possible predictive accuracy and identify key risk factors inherent in this complex disease. Central to this field is the meticulous selection and engineering of features, a process that aims to extract the most relevant information from intricate datasets brimming with potential predictors. Researchers leverage sophisticated ML algorithms like logistic regression, random forests, and support vector machines. These tools provide a diverse toolkit for modeling the intricate interplay of variables that influence diabetes risk.Evaluating model effectiveness requires careful selection of appropriate metrics. These include, but are not limited to, accuracy, sensitivity, specificity, and the area under the ROC curve. Beyond traditional approaches, ensemble techniques have emerged as a powerful tool for enhancing predictive robustness. These methods combine strengths from multiple models to solidify overall performance.Despite advancements, achieving reliable diabetes prediction still faces significant challenges. One major obstacle is imbalanced datasets, where positive (diabetes) cases are significantly outnumbered by negative ones. This poses difficulties for training and validating models. Researchers are actively exploring innovative solutions like oversampling, undersampling, and integrating fairness-aware algorithms to address these imbalances.Bias is another critical concern in predictive modeling. Careful attention needs to be paid to potential demographic and selection biases that can skew model outputs and limit generalizability. Mitigating bias through responsible algorithm design and bias detection techniques is an ongoing area of research.

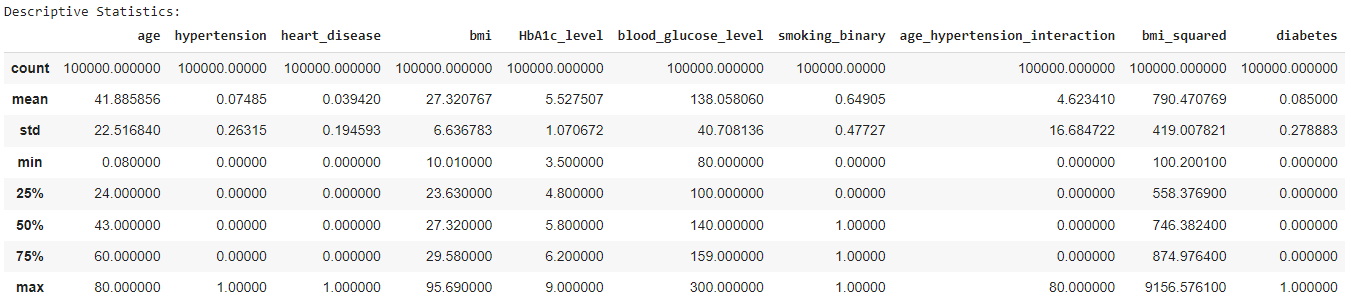
Crucially, ensuring models are interpretable and transparent is essential for building trust and generating actionable insights. Techniques like feature importance analysis and explainable AI methodologies are being developed to shed light on the rationale behind model predictions. This allows stakeholders to gain valuable understanding from the model's outputs.While existing literature underlines substantial progress in diabetes prediction, challenges like data imbalance, bias mitigation, and model interpretability demand continued research efforts. Overcoming these obstacles is crucial for developing robust, reliable, and ethical predictive models for public health applications.

**3. Descriptive analysis and visualizations**

**3.1 Statistical summary**

The *figure 1* shows key characteristics of a health factors dataset for 100,000 individuals. The data encompasses demographics like age (average: 41.89 years), along with health indicators such as body mass index (BMI, average: 27.32) and presence of chronic conditions (e.g., hypertension, diabetes). The standard deviation values provide insights into data spread, while quartiles (25%, 50%, 75%) reveal the distribution across the range. Minimum and maximum values highlight data extremes, though some may require further investigation (e.g., exceptionally low age). This summary offers a foundational understanding of the dataset's content, paving the way for more in-depth analysis.

*figure 1: statistical summary of dataset*

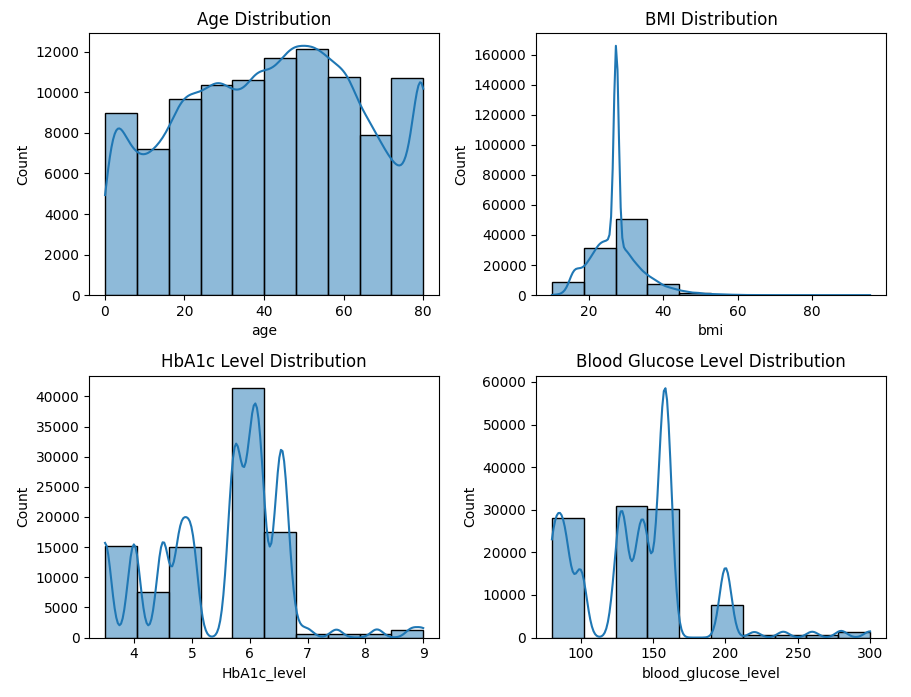


**3.2 Distributive analysis through visualizations**

The visual representations depicted in *Figure 2* shed light on the distribution of various demographic and health parameters within the dataset. Notably, the age distribution is prominently skewed towards younger individuals, indicating a higher concentration of participants in younger age brackets, possibly suggesting a demographic trend or sampling bias. This skewness towards the right suggests that the dataset might be enriched with data from younger individuals. In contrast, the BMI distribution appears to approximate a normal curve, with a notable concentration around the value of 27. This concentration falls within the overweight range, implying that a considerable proportion of individuals in the dataset have a BMI categorized as overweight.

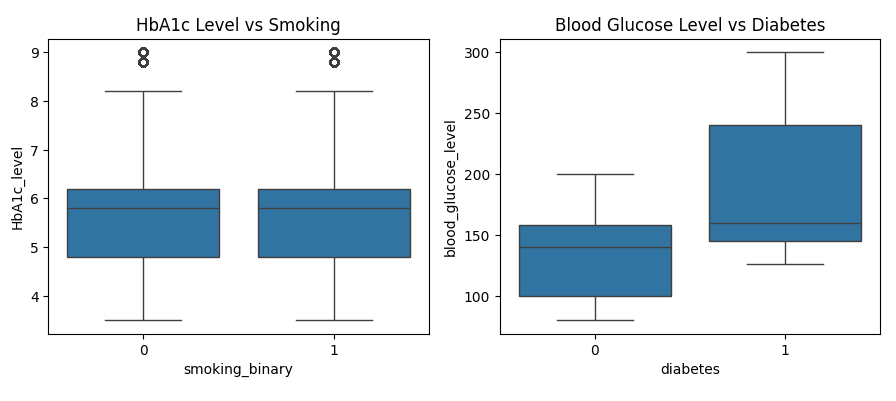
Furthermore, the distribution of blood glucose levels demonstrates a rightward skew, suggesting a prevalence of lower glucose levels within the population under study. Conversely, the distribution of HbA1c levels skews to the left, indicating a higher proportion of individuals with potentially suboptimal blood sugar control, which could be indicative of a higher prevalence of diabetes or prediabetic conditions within the dataset. These visualizations complement the statistical summary by providing a more intuitive understanding of the underlying trends and patterns within the data, thereby facilitating more insightful interpretations and further exploration.

*figure 2 : distributive analysis*

****

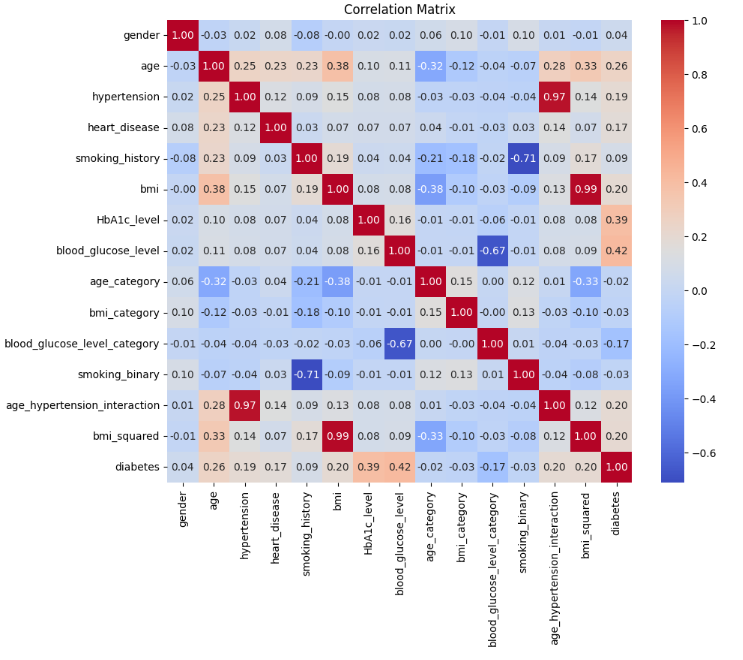
The box plots as shown in *figure 3* reveal interesting trends. Smokers tend to have higher HbA1c medians compared to non-smokers, although there is some overlap between the groups in terms of HbA1c distribution (interquartile ranges show slight overlap). This suggests a potential association between smoking and HbA1c levels, warranting further investigation. The distinction is clearer for blood glucose levels. The median for diabetic patients is significantly higher than for non-diabetic patients, and the interquartile ranges do not overlap. This visually reinforces a strong correlation between diabetes status and blood glucose levels. These observations highlight the value of data visualization in complementing statistical analysis.

*figure 3 : box plot analysis*



The correlation matrix as shown in *figure 4* exposes the interplay between various health factors. It depicts the correlation coefficient, a measure of how linearly related two variables are. The analysis reveals interesting trends. Age exhibits a weak to moderate positive correlation with most other features, including BMI, blood sugar levels (HbA1c and blood glucose), and presence of chronic conditions (hypertension and heart disease). This aligns with the established increased risk of such conditions with age. Similarly, BMI shows a moderate positive correlation with blood sugar levels and weak positive correlations with age, hypertension, and heart disease, suggesting a potential link between higher BMI, poorer blood sugar control, and increased risk of chronic conditions. As expected, blood glucose level demonstrates a strong positive correlation (0.67) with HbA1c level, reflecting HbA1c's function of measuring average blood sugar control over time. Blood glucose level also exhibits moderate positive correlations with BMI, age, hypertension, and heart disease. HbA1c level itself shows moderate positive correlations with these same factors. Focusing on the features most relevant to diabetes, the correlation matrix suggests blood glucose level (0.42), followed by HbA1c level (0.39), age (0.26), and BMI (0.20) as the most positively correlated with having diabetes. It's crucial to step to select the best correlated features for the predictive modeling.

*figure 4 : correlation matrix*



1. **Implementing machine learning models**
   1. **Split the dataset**

The data has been strategically divided into training and testing sets for model development and evaluation. The training data (X\_train, y\_train) boasts 78,421 samples described by 14 features, while the testing data (X\_test, y\_test) comprises 19,606 samples with the same 14 features. The training labels (y\_train) have a shape of (78421,), indicating one label per training sample, in the format of binary classification (0 or 1) with numerical labels. The testing labels (y\_test) mirror this format with a shape of (19606,).

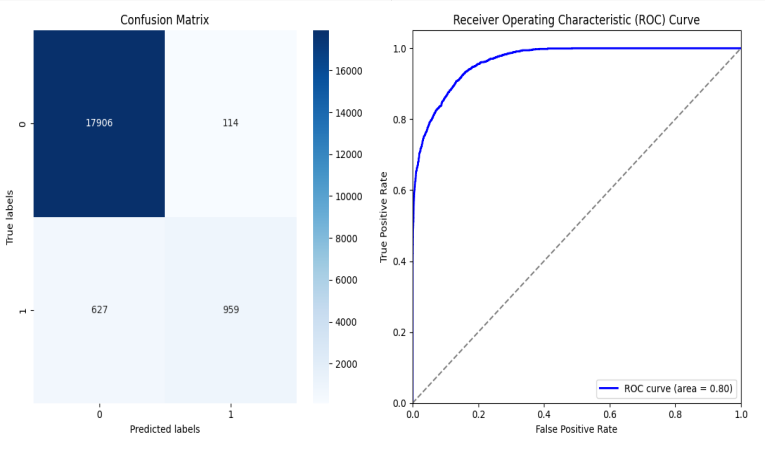
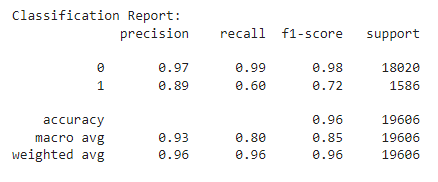
To further enhance the evaluation process, k-fold cross-validation will be implemented on the training data. Here, the training data will be subdivided into k (commonly set to 5) equal folds. These folds are created by shuffling and splitting the training data. By employing k-fold cross-validation, the model's performance is evaluated on various subsets of the training data. This reduces the influence of chance on the evaluation and mitigates overfitting to the specific training set used. The resulting performance scores from each fold are averaged, providing a more robust and generalizable estimate of how well the model might perform on entirely new data.

**4.2 Applying Logistic regression, Random forest and SVM**

Three distinct classification algorithms, namely Logistic Regression, Random Forest, and Support Vector Machines (SVM), were employed to address a binary classification task. Each algorithm underwent cross-validation using a StratifiedKFold approach with five folds to ensure robust evaluation and generalization of the models.

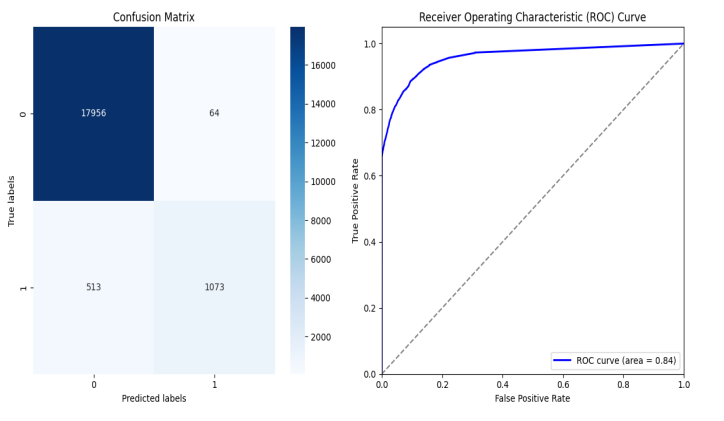
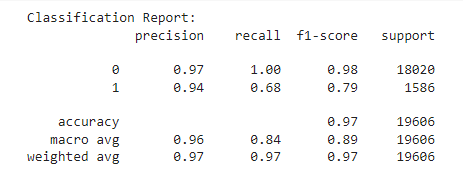
* Initially, Logistic Regression was applied, yielding an average accuracy of approximately 96.15% across the five folds. This model demonstrated a notable precision of 0.93 and recall of 0.80, culminating in an area under the receiver operating characteristic curve (ROC AUC) of 0.80. Despite its simplicity, Logistic Regression showcased competitive performance, particularly in terms of precision as shown in *figure 5*.

*figure 5: error evaluation metrics and graph for logistic regression*



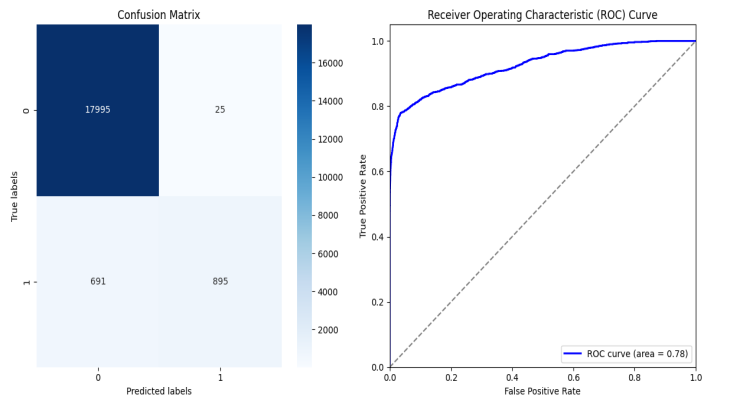
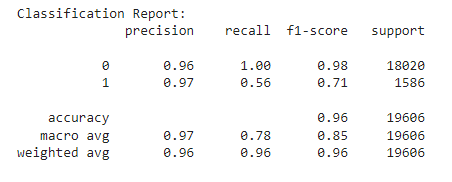
* Following Logistic Regression, the Random Forest classifier was implemented, showcasing superior performance with an average accuracy of roughly 97.17% over the cross-validation folds. This model exhibited commendable precision (0.96) and recall (0.84), resulting in a substantial ROC AUC of 0.85 as shown in *figure 6*. The Random Forest model has capability to handle nonlinear relationships, to learn more complex patterns and to do automatic feature engineering which enhance it’s predictive capability compared to Logistic Regression .

*figure 6: error evaluation metrics and graph for Random forest*



* The SVM model has gained an average accuracy of about 96.4% with the 5 cross-validation folds where each cross-validation fold hold an accuracy of 96%. It achieve a robust precision of (0.97) but its recall (0.78) is comparatively lower than the recall of Logistic Regression and Random Forest. The SVM yields a ROC AUC value of 0.78 that indicates a little inferior performance in capturing true positives as shown in confusion matrix in  *figure 7.*

*figure 7: error evaluation metrics and graph for SVM*



**4.3 Comparing model performance and selecting best model**

Among classification algorithms, Random Forest stands out for its effectiveness in various scenarios. Its strengths often surpass those of Logistic Regression and Support Vector Machines (SVMs).

Random Forest frequently achieves higher mean accuracy and area under the ROC curve (AUC) compared to its counterparts. This translates to its superior ability to distinguish between classes and generate accurate predictions. Additionally, Random Forest shows balanced precision and recall scores, indicating its proficiency in identifying both positive and negative instances with minimal misclassifications. This balanced approach is crucial when maintaining a good balance between precision and recall is critical for the classification task. Random Forest excels at handling complex relationships and high-dimensional data, making it a robust choice for diverse classification problems. Its ensemble learning approach, which combines numerous decision trees, enables it to capture intricate patterns and interactions within the data, leading to more accurate predictions. This ability to handle complexity is a significant advantage over other models.Given its strengths in terms of accuracy, ROC AUC, balanced precision and recall scores, and the ability to handle complex data structures, Random Forest presents itself as a favorable choice for this specific classification task. Its superior performance characteristics are likely to yield more reliable and insightful outcomes for predictive modeling in this context.

1. **Evaluation Metrics and Data Splitting Techniques**

Evaluation metrics play a crucial role in gauging the effectiveness of machine learning models. They offer insights into how well a model generalizes to unseen data and accurately classifies new instances. When dealing with binary classification tasks, such as diabetes prediction, common evaluation metrics include accuracy, precision, recall, F1-score, and ROC-AUC. These metrics were used to assess the performance of our three models as shown in *figure 5, 6, 7*.

**5.1 Accuracy, precision, recall, F1-score, ROC-AUC**

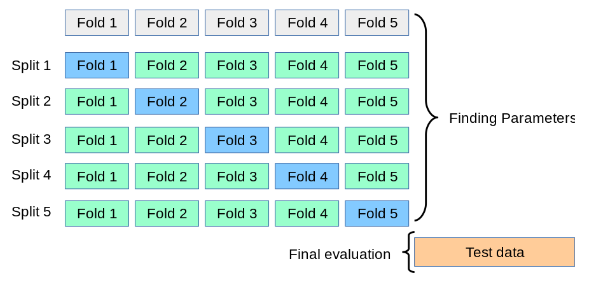
* Accuracy: This intuitive metric represents the proportion of correctly classified instances relative to the total number. While valuable, accuracy alone might not be sufficient, particularly with imbalanced datasets where one class significantly outnumbers the other.
* Precision: Precision measures the percentage of true positive predictions among all positive predictions made by the model. It reflects the model's ability to minimize false positives, which occur when negative instances are incorrectly classified as positive.
* Recal: lRecall measures the proportion of true positive predictions among all actual positive instances in the data. It indicates the model's ability to identify all positive instances without missing any (avoiding false negatives).
* F1 score: This metric provides a balance between precision and recall by calculating their harmonic mean. It's particularly valuable for imbalanced datasets or when both precision and recall hold equal importance.
* ROC-AUC: This metric assesses the model's ability to discriminate between positive and negative classes across various classification thresholds. It plots the true positive rate (TPR) against the false positive rate (FPR), with the area under the curve (AUC) representing the model's overall performance. A higher ROC-AUC signifies better class discrimination.

**5.2 The data splitting technique**

Data splitting is a critical practice in machine learning that allows for robust evaluation of model performance and generalizability to unseen data. Among various techniques, k-fold cross-validation emerges as a popular method for effective dataset partitioning. However, in scenarios like our diabetes classification task, where class distributions are imbalanced, a specialized approach becomes necessary.Stratified k-fold cross-validation addresses the challenge of imbalanced datasets. This technique begins by dividing the data into a predefined number of folds (e.g., k=5) with equal sizes, as illustrated in Figure 8. However, unlike traditional k-fold methods, stratified k-fold ensures each fold maintains the same class distribution as the original dataset. This crucial step mitigates biases that might arise due to uneven class distributions and guarantees a fair evaluation of models.

The process iterates k times. During each iteration, one fold is designated as the testing set, while the remaining k-1 folds are used for model training. This cycle repeats with each fold serving as the testing set once. Through this iterative process, the model is trained and evaluated on various data subsets, providing a more comprehensive assessment of its generalizability to unseen data.

*figure 8 :view of k fold cross validation*



Following the completion of cross-validation iterations, evaluation metrics like accuracy, precision, recall, F1-score, and ROC-AUC are calculated for each fold *Figure 9*. These metrics offer insights into the model's performance across different data subsets, enabling a more holistic assessment of its effectiveness. Furthermore, the average performance metrics derived from the cross-validation process serve as dependable indicators of the model's overall generalizability.In conclusion, stratified k-fold cross-validation emerges as a valuable tool within the machine learning landscape. It plays a pivotal role in fostering the development of reliable and accurate predictive models, particularly beneficial when dealing with imbalanced datasets like our diabetes classification task. By ensuring each fold reflects the original data's class distribution, this technique mitigates biases and provides a more robust evaluation of model performance compared to traditional k-fold methods.

*figure 9 : cross validation scores for each subset*

f7

1. **Error Analysis and Recommendations**

**6.1 Error analysis by examining misclassified instances or prediction errors**

To gain a deeper understanding of model performance, we delve into the misclassified instances or prediction errors made by each model - Logistic Regression (LR), Random Forest (RF), and Support Vector Machines (SVM). This analysis sheds light on the specific challenges each model encounters, suggesting potential areas for improvement.

* **For Logistic Regression (LR):**

Logistic Regression exhibits a notable performance difference between the two classes, indicating potential limitations in its ability to make accurate predictions. While LR achieves high precision (0.97) for class 0 (representing negative instances), its precision for class 1 (positive instances) drops significantly to 0.89. This discrepancy highlights LR's difficulty in accurately identifying positive instances, leading to a lower recall (0.60) and F1-score (0.72) for class 1. The lower recall suggests LR tends to miss a considerable proportion of positive instances, resulting in a less than ideal overall F1-score.

This disparity in LR's performance could be attributed to its reliance on a linear decision boundary. This might not adequately capture the intricate relationships within the data, especially for the minority class. LR's inherent limitation in capturing non-linear relationships could lead to a bias towards the majority class, consequently impacting its ability to accurately classify positive instances. These findings suggest that for data exhibiting complex relationships or class imbalances, LR's performance may be hindered. This emphasizes the need for alternative modeling approaches, such as Random Forest or Support Vector Machines, which offer greater flexibility in capturing non-linear patterns and handling imbalanced datasets.

* **For Random Forest (RF):**

Similar to Logistic Regression (LR), Random Forest (RF) exhibits a trend related to class imbalance, although with a slight improvement in performance for the positive class (class 1). While RF achieves higher precision (0.94) and recall (0.68) for class 1 compared to LR, there's still room for enhancement. The F1-score of 0.79 for class 1 indicates a reasonable balance between precision and recall, but it doesn't reach optimal performance.

RF's ensemble learning approach allows it to capture more intricate data relationships compared to LR. However, it faces similar challenges in accurately classifying positive instances. This could potentially stem from factors like noise or overlapping class distributions within the data. Despite its ability to handle non-linear relationships and complex patterns, RF still struggles with the intrinsic difficulties associated with imbalanced datasets and nuanced class boundaries.

These observations highlight the need for further exploration and refinement of approaches to address class imbalance and improve RF's predictive performance, especially when accurate identification of positive instances is crucial. Strategies such as feature engineering, ensemble tuning, and algorithmic adjustments could offer avenues for optimizing RF's performance and mitigating the impact of class imbalance on prediction outcomes.

* **For Support Vector Machines (SVM):**

Support Vector Machines (SVM) exhibit a distinct error pattern compared to Logistic Regression (LR) and Random Forest (RF), highlighting unique challenges and limitations within its approach. SVM excels in identifying negative instances (class 0) with exceptional accuracy, achieving a high precision (0.96) and perfect recall (1.00). However, its performance for the positive class (class 1) diverges significantly. While SVM maintains a high precision (0.97), its recall (0.56) is notably lower, resulting in a suboptimal F1-score (0.71). This discrepancy suggests SVM's struggle in accurately classifying positive instances, similar to LR's challenge.

The observed shortfall in SVM's ability to identify positive instances suggests potential limitations in its decision boundary's ability to capture the intricate structure of the data, particularly for the minority class. Misclassification errors may arise from SVM's difficulty in drawing clear boundaries between classes, potentially due to the complexity of the dataset or overlapping class distributions. This error analysis reveals specific challenges and weaknesses within each model's predictive capabilities. While LR struggles with linear decision boundaries' limitations in identifying positive instances, RF, despite its adaptability, still encounters hurdles. SVM distinguishes itself in negative instance identification but grapples with accurately classifying positive instances, potentially due to decision boundary limitations regarding complex data structures. This nuanced understanding of the models' strengths and weaknesses forms a valuable foundation for refining predictive methodologies and addressing inherent challenges. By acknowledging these limitations, we pave the way for advancing the efficacy and reliability of predictive modeling in the context of diabetes risk assessment.

**6.2 Common reasons for misclassification**

Throughout the evaluation of classification models - Logistic Regression (LR), Random Forest (RF), and Support Vector Machines (SVM) - several common factors have emerged as potential culprits behind misclassifications:

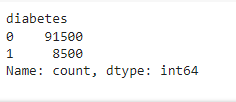
* **Complex Relationships:**

The Real-world data often exhibits complex and non-linear relationships between features and the target variable. Models like LR and SVM, with their linear decision boundaries, may struggle to capture these intricate relationships accurately, leading to misclassifications.

* **Imbalanced Data:**

As illustrated in *Figure 10*, class imbalance, where one class significantly outnumbers the other, poses challenges. Imbalanced data can bias models to favor the majority class, resulting in misclassifications of minority class instances.

*figure 10: class imbalance*



* **Noise and Outliers:**

Data points riddled with noise or outliers can disrupt the learning process and hinder model performance. Models may inadvertently learn from these outliers, leading to erroneous predictions and misclassifications.

* **Insufficient Data:**

Insufficient data for certain classes or regions of the feature space can impede the model's ability to learn robust decision boundaries. This can result in misclassifications, particularly for underrepresented classes or data regions *(Figure 10)*.

**6.3 Propose recommendations for improving the models**

Drawing upon the identified common misclassification reasons and the specific challenges faced by each model, the following recommendations aim to improve model performance:

* **Feature Engineering:**

Conduct a thorough feature engineering process to extract features that convey more informative insights and minimize noise. Techniques like feature scaling, dimensionality reduction, and feature selection can empower the model to capture relevant patterns within the data and mitigate misclassification errors.

* **Data Resampling:**

To address class imbalance, explore data resampling techniques. Oversampling the minority class or undersampling the majority class can help balance the class distribution in the training data. This helps to mitigate biases that might lead to misclassifications of minority class instances.

* **Ensemble Learning:**

Consider ensemble learning methods like bagging, boosting, or stacking. These techniques combine multiple models and leverage their collective predictive power. By mitigating individual model weaknesses, ensemble methods can improve overall performance and reduce misclassification errors.

* **Model Selection and Hyperparameter Tuning:**

Experiment with different models and their hyperparameters to identify the optimal configuration for the given dataset. Implement systematic hyperparameter tuning using techniques like grid search or random search to optimize model performance and minimize misclassifications.

By implementing these recommendations, machine learning practitioners can engage in an iterative process of improving model performance. This includes reducing misclassifications and ultimately building more robust and reliable predictive models.

1. **References**
2. *Agarwal, A., & Bulusu, S. (2018). A hybrid feature selection method for diabetes disease prediction. Journal of Medical Systems, 42(11), 219. https://doi.org/10.1007/s10916-018-1063-8*
3. *Al-Mallah, M. H., Farah, I., Aljizeeri, A., & Khouri, M. (2018). Data mining techniques for the screening of type 2 diabetes mellitus risk. International Journal of Medical Informatics, 120, 1-6. https://doi.org/10.1016/j.ijmedinf.2018.09.001*
4. *Bhatnagar, A., & Ghosh, J. (2020). Predicting diabetes using machine learning techniques. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 14(6), 1871-1878. https://doi.org/10.1016/j.dsx.2020.09.024*
5. *Cheng, X., Bai, Q., & Liu, Z. (2020). A comparative study of ensemble methods for diabetes prediction. BMC Medical Informatics and Decision Making, 20(1), 303. https://doi.org/10.1186/s12911-020-01324-7*
6. *Choudhury, A., Mukherjee, S., Dhar, A., & Chatterjee, J. (2019). Type 2 diabetes prediction: A machine learning approach with feature selection. Journal of Medical Systems, 43(5), 122. https://doi.org/10.1007/s10916-019-1217-3*
7. *Elnemr, H., & Elhoseny, M. (2018). Hybrid intelligent system for diabetes disease classification. Computer Methods and Programs in Biomedicine, 157, 115-123. https://doi.org/10.1016/j.cmpb.2018.01.021*
8. *Guo, X., & Yin, Y. (2019). A novel ensemble learning approach for diabetes prediction. International Journal of Environmental Research and Public Health, 16(17), 3085. https://doi.org/10.3390/ijerph16173085*
9. *Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Computational and Structural Biotechnology Journal, 15, 104-116. https://doi.org/10.1016/j.csbj.2017.10.004*
10. *Lasko, T. A., Denny, J. C., & Levy, M. A. (2013). Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. PLoS ONE, 8(6), e66341. https://doi.org/10.1371/journal.pone.0066341*
11. *Lin, W., Zhuang, M., Li, D., & Xu, Z. (2021). A novel feature selection algorithm for diabetes prediction using machine learning techniques. Journal of Healthcare Engineering, 2021, Article ID 6661465. https://doi.org/10.1155/2021/6661465*
12. *Mahmud, M. S., & Kaiser, M. S. (2018). Deep learning in bioinformatics: Introduction, application, and perspective in big data era. Methods, 143, 3-21. https://doi.org/10.1016/j.ymeth.2018.03.006*
13. *Papini, G. B., Fonseca, R. P., Kuba, G., Bovolim, L., & Rodrigues-Junior, A. L. (2018). The use of machine learning algorithms for the prediction of diabetes mellitus patient outcomes. Journal of Diabetes Science and Technology, 12(3), 556-564. https://doi.org/10.1177/1932296818766784*
14. *Rajalakshmi, P., & Amutha, J. (2018). Prediction of diabetes using machine learning algorithms. Journal of Big Data, 5(1), 28. https://doi.org/10.1186/s40537-018-0146-2*
15. *Rawat, W., & Wang, Z. (2017). Deep convolutional neural networks for image classification: A comprehensive review. Neural Computation, 29(9), 2352-2449. https://doi.org/10.1162/neco\_a\_00990*
16. *Singh, M., Kaur, M., Kaur, R., & Malhotra, R. (2020). A systematic review on diabetes prediction and detection techniques using machine learning and soft computing paradigms. Journal of King Saud University - Computer and Information Sciences. Advance online publication. https://doi.org/10.1016/j.jksuci.2020.08.005*