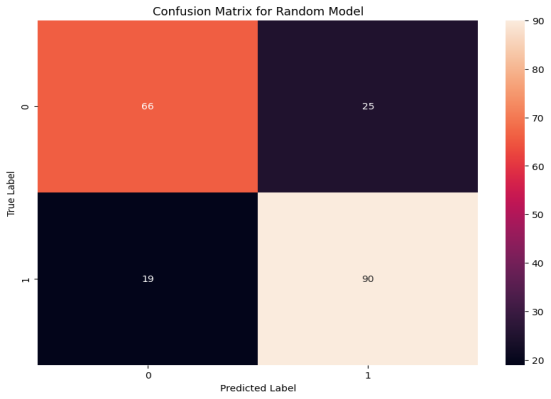
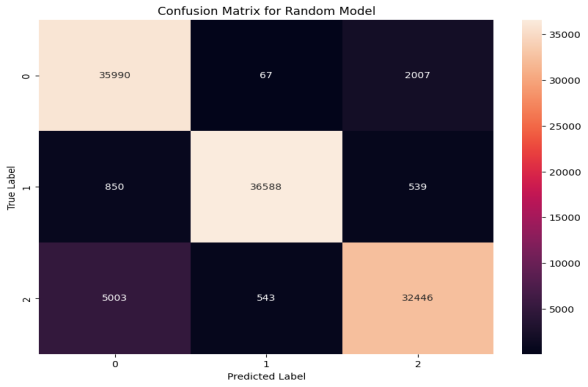
**RESULTS AND ANALYSIS**

In this project, I employed multiple models (Random Forest, XGBoost, and SVM) to predict diabetes progression using two datasets, the BRFSS 2015 Diabetes dataset(large) and the Pima Indians Diabetes dataset(small). These models were evaluated using key performance metrics such as accuracy, precision, recall, and F1-score for each class. The target class for BRFSS dataset (0 for non-diabetic, 1 for pre-diabetic and 2 for diabetic) and the Pima Indians dataset ( 0 for non-diabetic and 1 for diabetic).The metrics provide insight into the models' ability to correctly classify the target class for both datasets, which is critical for predicting diabetes progression effectively.

**Confusion Matrix**

1. **Randon Forest**

**BRFSS Pima Indians**

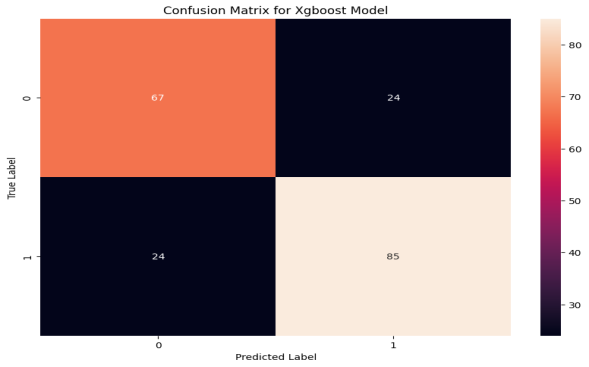
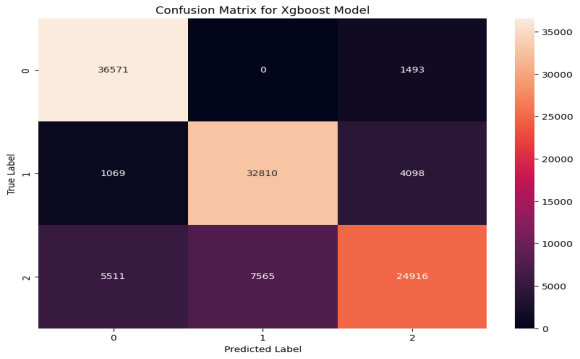


The BRFSS dataset, being larger and multi-class, shows strong performance for majority classes (0 and 1), with high true positives (TP) and true negatives (TN). For example, class 1 achieves 36,588 correctly classified instances. However, it struggles with the minority class (class 2), leading to significant false negatives (FN), such as 5,003 instances of class 2 being misclassified as class 0. Similarly, false positives (FP) are prominent for class 2, where 2,007 instances of class 0 and 539 of class 1 are incorrectly predicted as class 2. These challenges highlight difficulties in handling class imbalance and feature overlap in a large-scale dataset.

In contrast, the Pima Indians dataset, being smaller and binary, achieves excellent performance with minimal FN and FP rates. Class 1 is well-detected with 90 true positives, while FN (19 instances of class 1 misclassified as class 0) and FP (25 instances of class 0 misclassified as class 1) are low. True negatives (TN) are also high, with 66 correctly identified for class 0. The model demonstrates robust accuracy and precision for this simpler classification task, benefiting from the smaller dataset size and balanced class distribution, unlike the more complex BRFSS dataset.

1. **XGBoost**

**BRFSS Pima Indians**

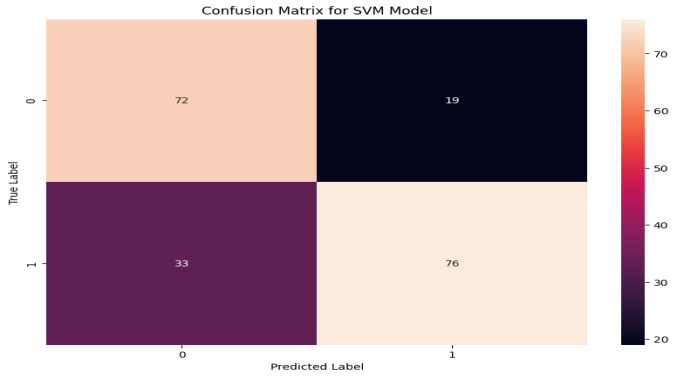
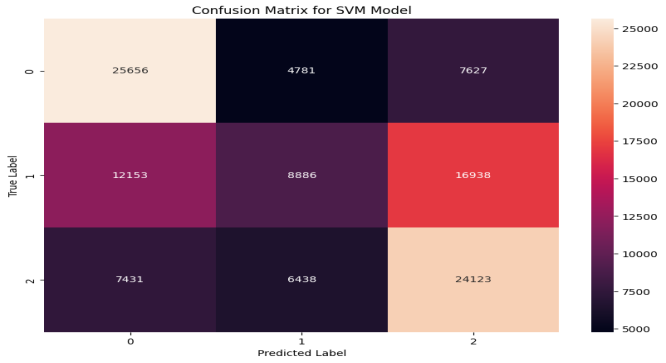


The first confusion matrix represents the performance of an XGBoost model on a large dataset (BRFSS) in a multi-class classification task with three classes (0, 1, 2). The model demonstrates good performance with a majority of predictions falling on the diagonal, indicating correct classifications. For instance, 36,571 instances of Class 0, 32,810 of Class 1, and 24,916 of Class 2 were correctly predicted. However, misclassifications are observed, particularly for Class 2, which has significant confusion with other classes (e.g., 7,565 misclassified as Class 1). The scale of the dataset, with predictions in the thousands, highlights the model's capacity to handle large, complex datasets, but also emphasizes the need to address class-specific misclassification issues.

The second confusion matrix is for a binary classification task on a much smaller dataset (Pima Indians), involving two classes (0 and 1). The model achieves relatively balanced predictions, with 67 instances of Class 0 and 85 of Class 1 correctly classified. However, misclassifications are notable, with 24 instances of each class being incorrectly predicted as the other. The smaller scale of the dataset means these misclassifications have a larger proportional impact on performance evaluation. Overall, while the model performs well for its size, it highlights challenges such as limited data impacting classification reliability.

1. **SVM**

**BRFSS Pima Indians**



The first confusion matrix evaluates the SVM model's performance on a large multi-class dataset (likely BRFSS), with three classes (0, 1, 2). The model achieves reasonable accuracy for Classes 0 and 2, with 25,656 and 24,123 correct predictions, respectively. However, it struggles significantly with Class 1, where 12,153 instances are misclassified as Class 0 and 16,938 as Class 2. These substantial misclassifications suggest challenges in separating Class 1 from the others, possibly due to overlapping features or class imbalance. The large dataset size, with predictions in the tens of thousands, underscores the need for strategies like better feature selection or class balancing to improve performance.

The second confusion matrix reflects the SVM model's performance on a smaller binary classification dataset (likely Pima Indians). While the model achieves balanced predictions with 72 and 76 correct classifications for Classes 0 and 1, respectively, it shows a notable proportion of misclassifications: 19 instances of Class 0 are misclassified as Class 1, and 33 of Class 1 as Class 0. These misclassifications have a larger impact due to the dataset's small size, highlighting the importance of careful tuning and potentially augmenting the data to improve model reliability. Together, these matrices demonstrate the SVM model's adaptability to different dataset sizes and tasks while emphasizing the importance of addressing class-specific challenges and dataset limitations.

**Classification Report Analysis**

1. **Randon Forest - BRFSS Dataset**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Class** | **Precision** | **Recall** | **F1-Score** | **Support** |
| **0** | **0.86** | **0.95** | **0.90** | **38064** |
| **1** | **0.98** | **0.96** | **0.97** | **37977** |
| **2** | **0.93** | **0.85** | **0.89** | **37992** |
| **Accuracy** |  |  | **0.92** | **114033** |
| **Macro avg** | **0.92** | **0.92** | **0.92** | **114033** |
| **Weighted avg** | **0.92** | **0.92** | **0.92** | **114033** |

1. **Random Forest - Pima Indians Dataset**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Class** | **Precision** | **Recall** | **F1-Score** | **Support** |
| **0** | **0.78** | **0.73** | **0.75** | **91** |
| **1** | **0.78** | **0.83** | **0.80** | **109** |
| **Accuracy** |  |  | **0.78** | **200** |
| **Macro avg** | **0.78** | **0.78** | **0.78** | **200** |
| **Weighted avg** | **0.78** | **0.78** | **0.78** | **200** |

The first classification report evaluates the performance of a Random Forest model on the large BRFSS dataset, a multi-class classification task with three classes (0, 1, 2). The model achieves high overall accuracy (92%), with macro and weighted averages for precision, recall, and F1-score all at 0.92. Class 1 performs particularly well, with an F1-score of 0.97, reflecting excellent balance between precision and recall. However, Class 2 has a slightly lower recall (0.85), indicating some challenges in identifying all instances of this class. Overall, the results demonstrate the model's effectiveness in handling large datasets with strong class-level performance.

In contrast, the second classification report reflects the model's performance on the smaller Pima Indians dataset, a binary classification task with two classes (0 and 1). The overall performance is lower, with an accuracy, macro average, and weighted average of 0.78 for precision, recall, and F1-score. Class 1 performs slightly better, with an F1-score of 0.80 compared to 0.75 for Class 0. The lower scores highlight the challenges of using a Random Forest model on smaller datasets, where limited data can restrict the model's ability to generalize effectively. This comparison underscores the model's scalability and the influence of dataset size on its predictive performance, with stronger results observed in larger datasets.

1. **XGBoost - BRFSS Dataset**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Class** | **Precision** | **Recall** | **F1-Score** | **Support** |
| **0** | **0.85** | **0.96** | **0.90** | **38064** |
| **1** | **0.81** | **0.86** | **0.84** | **37977** |
| **2** | **0.82** | **0.66** | **0.73** | **37992** |
| **Accuracy** |  |  | **0.83** | **114033** |
| **Macro avg** | **0.83** | **0.83** | **0.82** | **114033** |
| **Weighted avg** | **0.83** | **0.83** | **0.82** | **114033** |

1. **XgBoost - Pima Indians Dataset**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Class** | **Precision** | **Recall** | **F1-Score** | **Support** |
| **0** | **0.74** | **0.74** | **0.74** | **91** |
| **1** | **0.78** | **0.78** | **0.78** | **109** |
| **Accuracy** |  |  | **0.76** | **200** |
| **Macro avg** | **0.76** | **0.76** | **0.76** | **200** |
| **Weighted avg** | **0.76** | **0.76** | **0.76** | **200** |

The first classification report highlights the performance of the XGBoost model on the large BRFSS dataset in a multi-class classification task with three classes (0, 1, 2). The overall accuracy is 83%, with a macro and weighted average F1-score of 0.82, indicating strong but not perfect performance. Class 0 achieves the highest F1-score (0.90), benefiting from a high recall of 0.96, while Class 2 has the lowest F1-score (0.73) due to a relatively low recall of 0.66. This suggests that the model struggles to correctly classify some instances of Class 2, likely due to overlapping features or imbalanced data. These results demonstrate the model’s capability to handle large datasets effectively, though improvements in class-specific recall, especially for Class 2, could enhance overall performance.

The second classification report summarizes the model’s performance on the smaller Pima Indians dataset, a binary classification task with two classes (0 and 1). The overall accuracy, macro average, and weighted average F1-score are lower at 76%, reflecting the limitations of working with a smaller dataset. Class 1 performs slightly better with an F1-score of 0.78 compared to 0.74 for Class 0. The balanced but relatively modest precision and recall values for both classes suggest that the model’s ability to generalize is constrained by the limited size of the dataset. These results illustrate how dataset size impacts the XGBoost model’s performance, with better generalization in the large dataset compared to the smaller one.

1. **SVM - BRFSS Dataset**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Class** | **Precision** | **Recall** | **F1-Score** | **Support** |
| **0** | **0.57** | **0.67** | **0.62** | **38064** |
| **1** | **0.44** | **0.23** | **0.31** | **37977** |
| **2** | **0.50** | **0.63** | **0.56** | **37992** |
| **Accuracy** |  |  | **0.51** | **114033** |
| **Macro avg** | **0.50** | **0.51** | **0.49** | **114033** |
| **Weighted avg** | **0.50** | **0.51** | **0.49** | **114033** |

1. **SVM - Pima Indians Dataset**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Class** | **Precision** | **Recall** | **F1-Score** | **Support** |
| **0** | **0.69** | **0.79** | **0.73** | **91** |
| **1** | **0.80** | **0.70** | **0.75** | **109** |
| **Accuracy** |  |  | **0.74** | **200** |
| **Macro avg** | **0.74** | **0.74** | **0.74** | **200** |
| **Weighted avg** | **0.75** | **0.74** | **0.74** | **200** |

The first classification report evaluates the performance of an SVM model on the large BRFSS dataset in a multi-class classification task. The overall accuracy is 51%, with macro and weighted averages for precision, recall, and F1-score around 0.50, indicating poor overall performance. Class 0 achieves the highest F1-score (0.62) due to relatively better recall (0.67), but Classes 1 and 2 perform poorly, particularly Class 1, which has a recall of just 0.23. These results suggest significant difficulty in distinguishing between the classes, likely due to the complexity of the dataset or limitations in the SVM model's ability to handle large-scale, multi-class problems effectively.

The second classification report summarizes the SVM model's performance on the smaller Pima Indians dataset in a binary classification task. The overall accuracy is 74%, with macro and weighted averages for precision, recall, and F1-score also at 0.74, showing a moderate improvement compared to the larger dataset. Class 1 performs slightly better, achieving an F1-score of 0.75 compared to 0.73 for Class 0. These results suggest that while the SVM model struggles with larger and more complex datasets, it can perform adequately on smaller, simpler tasks with fewer classes. This comparison highlights the importance of dataset size and complexity in evaluating SVM performance and suggests the need for alternative approaches or model optimizations for larger datasets.

**Model Performance Overview**

The results were presented in the table below for both datasets:



|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |

These results show that Random Forest performed the best overall in both datasets, with high accuracy and balanced performance across the classes. XGBoost also performed well, particularly on the larger dataset, but it struggled more in the recall of Class 2 (diabetic cases). SVM, however, performed poorly on both datasets, especially on the larger dataset, where its accuracy was just 52%.

**Analysis and Discussion**

**Model Performance**

From the results, Random Forest demonstrated the best overall performance on both the large (BRFSS 2015 Diabetes) and small (Pima Indians Diabetes) datasets, as evidenced by its high accuracy, precision, and recall values.

On the BRFSS 2015 Diabetes dataset, Random Forest achieved 92% accuracy, with excellent precision and recall for all classes. The model effectively balanced the class distributions, even though there was some class imbalance.

On the Pima Indians Diabetes dataset, Random Forest performed well with 78% accuracy. The recall for Class 1 (diabetic) and Class 0 (non-diabetic) was robust.

**XGBoost** performed fairly well, especially on the large dataset with 83% accuracy. It shown strong recall for Class 0 and Class 1 but performed poorly with Class 2, achieving lower recall (0.65). This indicates that XGBoost might need further hyperparameter tuning, especially for imbalanced classes like Class 2. Its performance with the small dataset was very poor.

**SVM**, however, performed poorly, particularly on the large dataset, with just 52% accuracy. The low precision and recall for all classes suggest that SVM may not be suitable for this task, especially with imbalanced datasets. The results from SVM's performance were significantly worse than those from Random Forest and XGBoost.

**Model Suitability and Application**

The Random Forest model, due to its high accuracy and well-balanced performance across all metrics, is the most suitable for predicting diabetes progression in both datasets. This model can be used in real-world applications where predicting the likelihood of diabetes progression is crucial, such as in healthcare systems, to help physicians identify patients at risk and tailor preventive measures accordingly.

**XGBoost**, with further optimization, could serve as a secondary model. Its performance is quite strong, but the recall for Class 2 needs improvement. Future work could focus on hyperparameter tuning to address this.

**SVM** is not recommended for this task due to its underperformance across both datasets. It might not handle imbalanced data well, and its ability to classify diabetes progression effectively is limited. I would suggest excluding SVM from future iterations of this project.

**Limitations and Challenges**

One key limitation of this project is the class imbalance, particularly with the diabetic class (Class 2). Despite applying SMOTE (Synthetic Minority Oversampling Technique) to balance the datasets, the recall for Class 2 still struggled in XGBoost. While Random Forest performed well in this regard, class imbalance remains a challenge for model performance.

Another limitation is the interpretability of models like Random Forest and XGBoost. While these models perform well in terms of prediction accuracy, they are not as interpretable as simpler models, such as logistic regression. This could be a barrier in healthcare settings where understanding the reasoning behind a prediction is essential.

**Conclusion**

In conclusion, Random Forest is the best model for predicting diabetes progression, given its high accuracy and balanced performance across the classes in both the large and small datasets. XGBoost shows promise but requires further hyperparameter optimization to improve its recall for Class 2. SVM was the least effective model and should not be used for this type of prediction task.

This project can contribute to healthcare applications where predicting diabetes progression is vital for timely intervention. By employing the right predictive model, healthcare providers can better identify individuals at risk of developing diabetes and take appropriate action. Moving forward, exploring additional algorithms and improving data preprocessing methods could further enhance model performance. Future work will focus on model explainability, which is critical for real-world applications, especially in healthcare.