Alex Yeo

Capstone 2 Final Report:
Predicting Diabetes

Problem Statement and Background:

Each year it's been reported that roughly 537 million adults, aged 20 to 79 years old, and account for around 6.7 million deaths annually[1]. In addition, diabetes care is reported to cost up to $966 billion USD[2]. Diabetes is a disease that impacts a large population across the globe, with a vast majority of the population impacted by diabetes located in low and middle income countries. As a result, access and treatment to diabetes care is often difficult to obtain for the majority of the diabetes population across the globe. With more accurate prediction tools available, diabetes can further be prevented through less cost afflicting methods such as change in lifestyle choices. Our goal is to recommend an accurate tool for identifying predominant symptoms and potential patient demographic information to most accurately predict both the likelihood and early signs of contracting diabetes.

By using the "Early Stage Diabetes Risk Prediction" data from UC Irvine Machine Learning Repository, a model was created to most accurately predict and classify early stages of diabetes based upon binary responses to fifteen symptoms, or features of the data set. This dataset explores 520 patients at the Sylhet Diabetes Hospital in Sylhet, Bangladesh who responded with their age, gender, and 'yes' or 'no' to having the following symptoms: Polyuria, Polydipsia, sudden weight loss, weakness, Polyphagia, Genital thrush, visual blurring, Itching, Irritability, delayed healing, partial paresis, muscle stiffness, Alopecia, and Obesity[3].

Data Wrangling:

The dataset featured 520 rows with 16 total columns. Except for the age feature(column), all columns featured categorical binary responses. The dataset was evaluated for any missing or null values. There were no null values observed from the dataset. To improve future use of the dataset to train models, the categorical values were converted to binary numerical values, 0 and 1. '0' was used to indicate 'Male' for gender and 'No'. '1' was used to indicated 'Female' for gender and 'Yes'.

---

[1] diabetesatals.org
[2] diabetesatals.org
[3] https://archive.ics.uci.edu/dataset/529/early+stage+diabetes+risk+prediction+dataset

Alex Yeo

Exploratory Data Analysis (EDA) and Initial Findings:

Prior to converting the dataset to binary numerical values, the dataset was explored to find any initial patterns, correlations, and relationships between diagnosis of diabetes and the number of people who did or didn't show symptoms of Polyuria, Polydipsia, sudden weight loss, weakness, Polyphagia, Genital thrush, visual blurring, Itching, Irritability, delayed healing, partial paresis, muscle stiffness, Alopecia, and Obesity. In addition, the dataset was also evaluated to see if age and gender also played a role in patient diabetes diagnosis.

The dataset was initially visualized to evaluate any trends in relation to the independent features and the dependent feature, diabetes diagnoses (see Figure 1). In addition, the data further visualized the average age of patients who were diagnosed with diabetes by their genders(see Figure 2). The initial data visualization findings showed that 320 patients were diagnosed with diabetes, while the average age of patients diagnosed with diabetes were approximately 47 years old.

In order to find any initial patterns and correlation between age, gender and the symptoms in which patients provided responses, a heat map was evaluated (see Figure 3). Values most closest to '1.0' showed a strong positive correlation to the feature that's being compared. Values most closest to '0.0' show no correlation to the feature that's being compared. Values most closest to '-1.0' show a strong negative correlation to the feature that's being compared. From our initial analysis, it can be shown initially that symptoms associated with Polydipsia and Polyuria (0.648734 and 0.665922) have a correlation to Diabetes Results.

After the initial evaluation using a heatmap, a Chi-square test was used to model any significant association between categorical variables. The Chi-square test was important in comparing the observed frequencies with the expected frequencies. The Chi-square test results showed features arranged in descending order, with the features showing the highest Chi-square value suggesting a strong association between the feature and diagnosis of diabetes (see Figure 4). In addition, the p-values of the associated features to the diagnosis of diabetes were further evaluated. A p-value that is greater than 0.5 suggests that the data being explored does not have significance to reject the null hypothesis. The null hypothesis suggests that there's no correlation to the feature and the diagnosis of diabetes (see Figure 5).

Alex Yeo

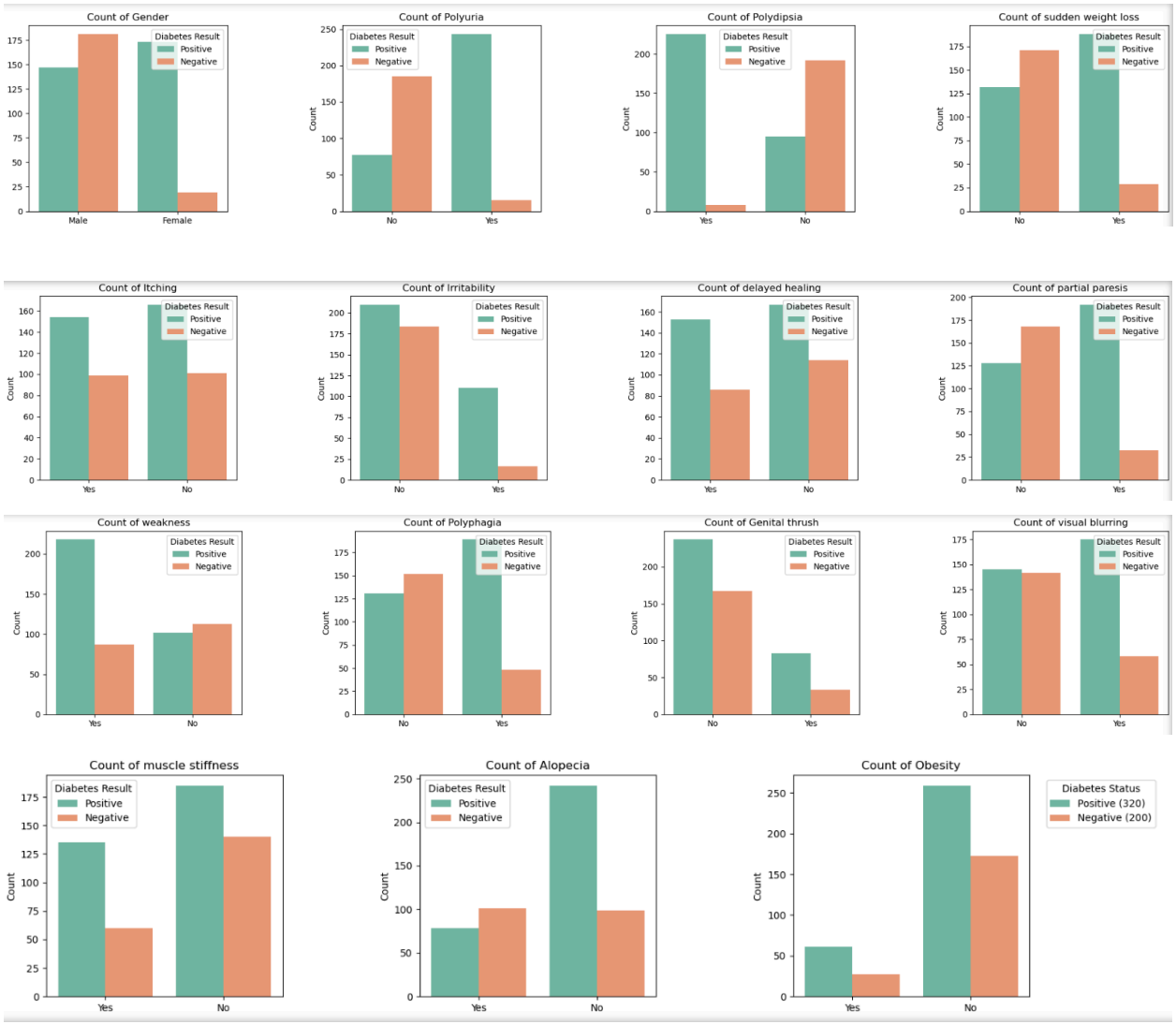


Figure 1: Initial Data Visualization

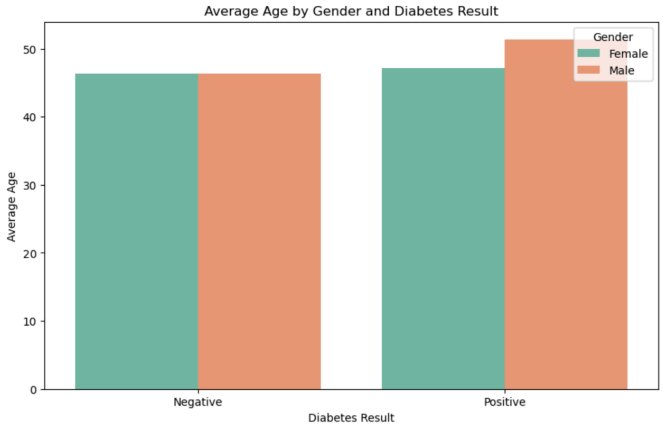


Figure 2: Average age of patients diagnosed with Diabetes by gender

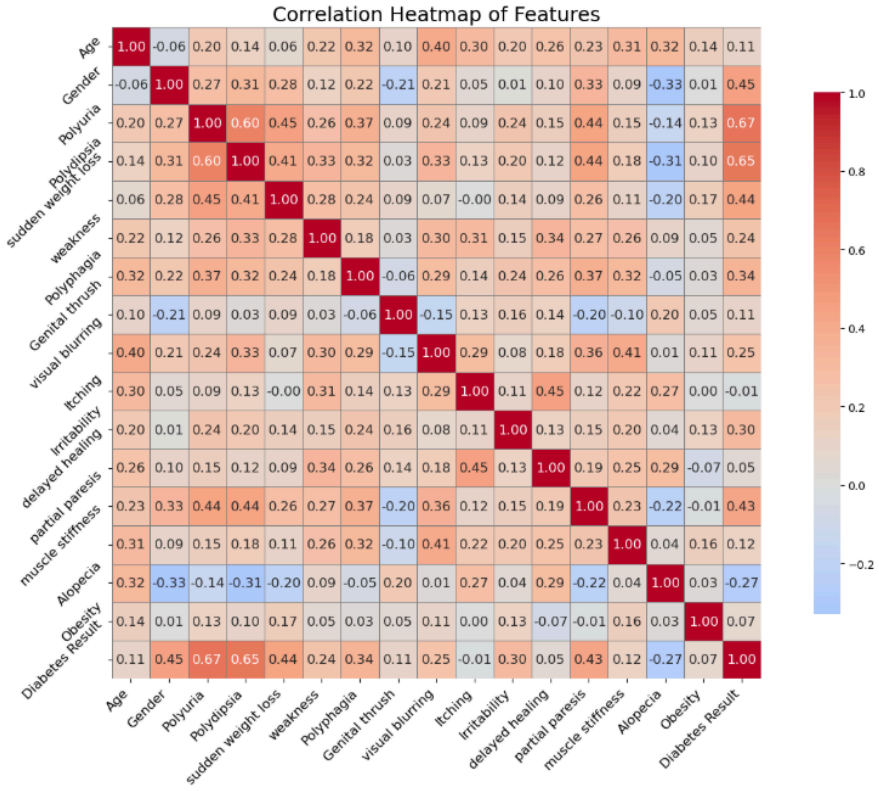


Figure 3: Heatmap of features

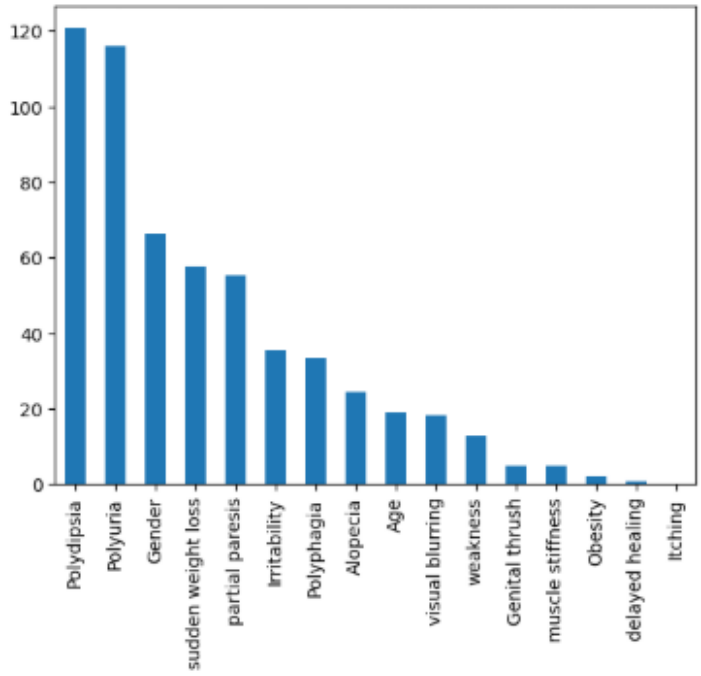


Figure 4: Chi-square results showing features with strong correlation to Diabetes diagnosis

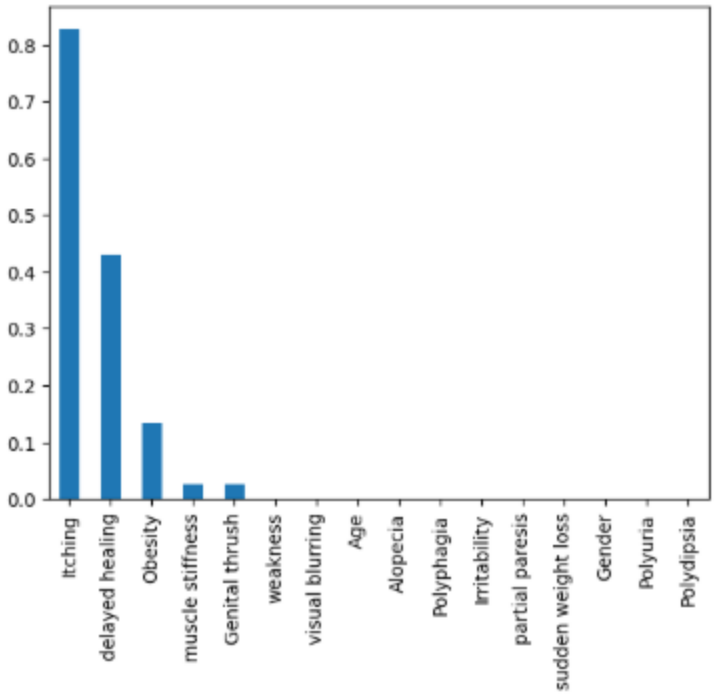


Figure 5: P-value results

In-Depth Analysis and Initial Model Training and Development:

After the observations were captured to evaluate trends and correlation of features to Diabetes diagnosis, the data set was further analyzed to rank features by highest to lowest importance through feature engineering (see Figure 6), and initially modeled to evaluate the accuracy only utilizing the most importantly selected features. Within the initial model training and development, the dataset was evaluated using logistical regression and random forest modeling. The performance of these models were first evaluated for accuracy and then analyzed using the following methods: Receiver Operating Curve (ROC), confusion matrix, and precision and recall score. From the initial modeling and training, it was observed that utilizing only the most important features yielded less optimal results as compared to including all features of the dataset. In addition, the random forest model had showed the poorest performance during initial model training and development.

After the conclusion of the in-depth analysis and initial model training and development, it was found that logistic regression utilizing all the features of the dataset had shown most optimal performance in comparison to only utilizing a model with the most importantly selected features. Figures 7-9 detail the results of the Receiver Operating Curve, precision and recall scores, and confusion matrix. “Logistic Regression Model 1” denotes the model that contain all features of the dataset, while

“Logistic Regression Model 2” denotes the model that only uses 5 selected features ('Polydipsia', 'Polyuria', 'sudden weight loss', 'partial paresis', 'Gender').

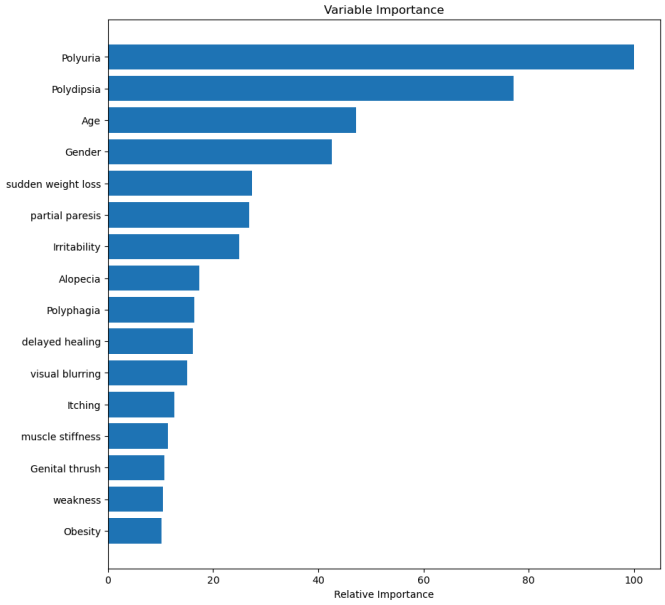


Figure 6: Features ranked by importance

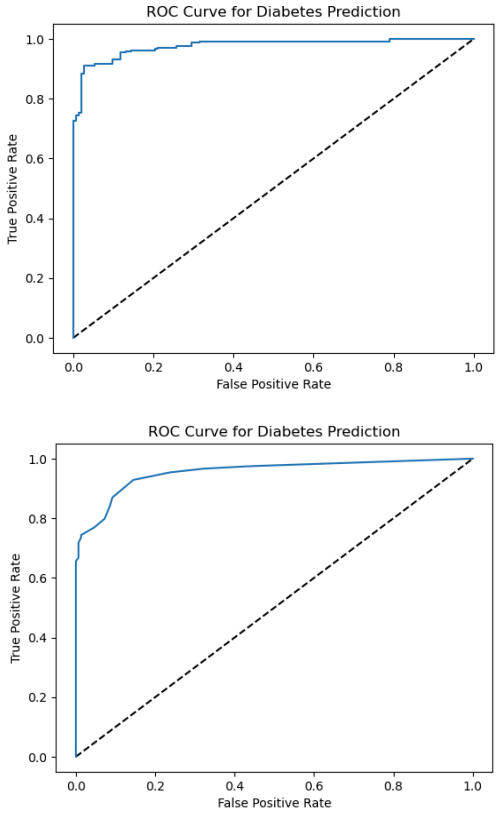


Figure 7: ROC results - “Logistic Regression Model 1” (Top) and “Logistic Regression Model 2” (Bottom)

	precision	recall	f1-score	support
0	0.88	0.92	0.90	152
1	0.95	0.92	0.93	238
accuracy			0.92	390
macro avg	0.91	0.92	0.91	390
weighted avg	0.92	0.92	0.92	390

	precision	recall	f1-score	support
0	0.85	0.88	0.86	152
1	0.92	0.90	0.91	238
accuracy			0.89	390
macro avg	0.89	0.89	0.89	390
weighted avg	0.89	0.89	0.89	390

Figure 8: Precision and Recall results - “Logistic Regression Model 1” (Top) and “Logistic Regression Model 2” (Bottom)

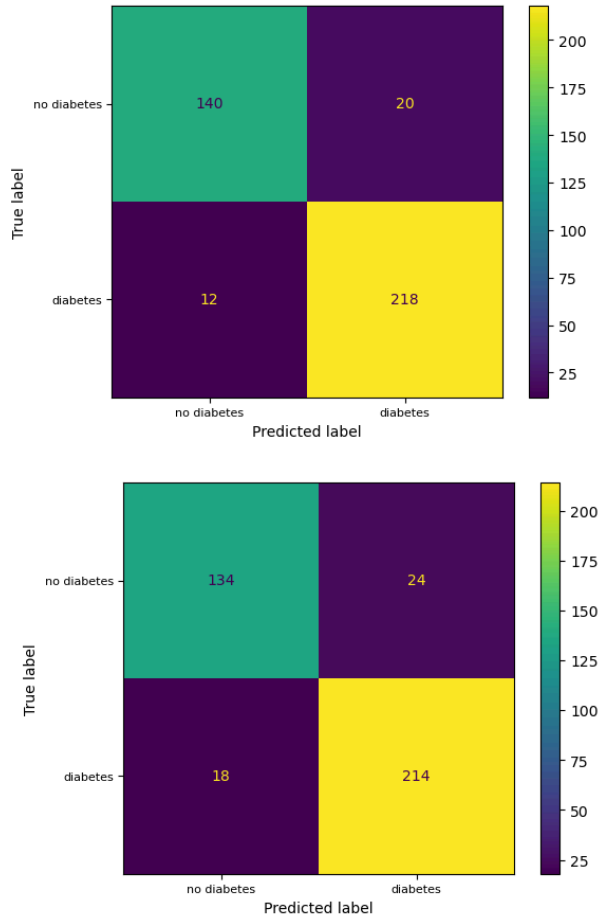


Figure 9: Confusion Matrix results - “Logistic Regression Model 1” (Top) and “Logistic Regression Model 2” (Bottom)

Figure 7 visualizes the ROC curve, where a true positive rate that's at 1 or 100% shows that the model is a perfect classifier and is able to provide the frequency for a true positive classification for diabetes diagnosis, while the false positive rate would be 0 or 0% and would show the frequency for a false positive classification for diabetes diagnosis . "Logistic Regression Model 1" (Top) had an ROC score of 0.97, while "Logistic Regression Model 2" (Bottom) had an ROC score of 0.95.

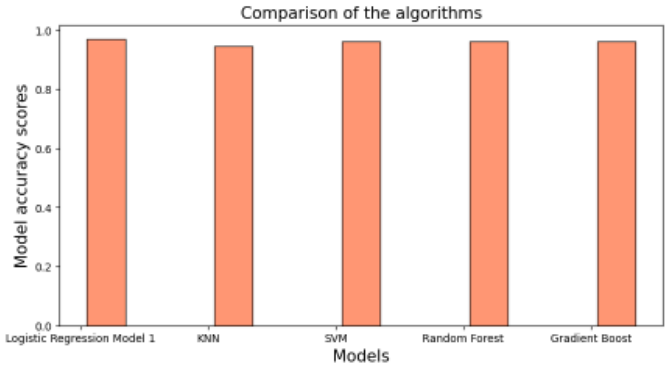
Figure 8 shows the summary of results around the evaluation of the two models pertaining to precision, recall, and F-1 score. Precision measures how many of the predicted positive instances of diabetes were actually positive diagnosis. Recall (or Sensitivity) measures how many of the actual positive instances for diabetes were correctly identified. The F1 score is the average or mean of the precision and recall results. When evaluating the performance of the models, the accuracy, macro average and the weighted average of precision, recall, and F1 score were compared. "Logistic Regression Model 1" (Top) had an accuracy result of 92%, while "Logistic Regression Model 2" (Bottom) had an accuracy result of 89%. "Logistic Regression Model 1" (Top) had a macro average result of 0.91, 0.92, and 0.91 respectively for Precision, Recall, and F1-Score**,** while "Logistic Regression Model 2" (Bottom) had a macro average result of 0.89 respectively across the Precision, Recall, and F1-Score.

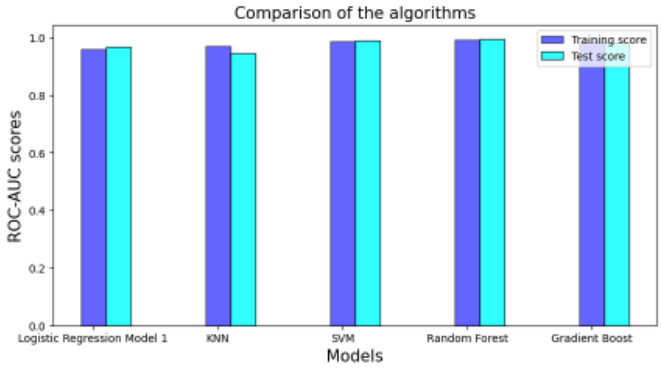
Figure 9 compares the result of the confusion matrix result between "Logistic Regression Model 1" (Top) and "Logistic Regression Model 2" (Bottom). For "Logistic Regression Model 1" (Top), the model correctly predicted 140 instances as class 0 (negative class), incorrectly predicted 12 instances as class 1 when they were actually class 0, incorrectly predicted 20 instances as class 0 when they were actually class 1, and correctly predicted 218 instances as class 1 (positive class). The model achieves high accuracy (**92%**) and performs well in both precision (**94.7%**) and recall (**91.6%**) for class 1. The F1-score for class 1 (**93.1%**) indicates a strong balance between precision and recall. There are relatively few false positives (**12**) and false negatives (**20**), indicating the model is reliable for both classes. The false negatives (**20**) are slightly higher than the false positives (**12**), suggesting that the model is slightly more conservative in predicting class 1 (positive class).

Model Selection and Takeaways:

The results in comparing "Logistic Regression Model 1" (Top) and "Logistic Regression Model 2" (Bottom) were used as a basis for training and testing a variety of different classification models using all of the features that were part of the original dataset. The goal in training and testing various classification models was to see which classification model should be selected based on evaluation of performance using the most optimal parameters. Each model was evaluated for accuracy score, and

'ROC-AUC' score for both the training and test data, and visualized. The two best performing models were the Random forest and the Support Vector Machine (SVM). A randomized grid search cross validation (CV) method was utilized for hyperparameter tuning for both the models separately. Randomized grid search cross validation was chosen in order to remedy the longer time gradient search CV takes. The result of fitting the models with the optimal hyperparamaters resulted in better accuracy for the SVM model, while fitting the models with the optimal hyperparamaters resulted in worse accuracy for the random forest model. **Figure 10** summarizes the comparison of the accuracy and the ROC-AUC scores of all the classification models that were evaluated prior to hyperparamater tuning.





	Algorithm	Model accuracy score
0	Logistic Regression Model 1	0.969231
1	KNN	0.946154
2	SVM	0.961538
3	Random Forest	0.961538
4	Gradient Boost	0.961538

	Algorithm	ROC-AUC train score	ROC-AUC test score
0	Logistic Regression Model 1	0.960486	0.968282
1	KNN	0.970383	0.945712
2	SVM	0.985362	0.990134
3	Random Forest	0.990693	0.994392
4	Gradient Boost	0.982628	0.980985

Figure 10: Summary of Classification Model Evaluation Results (pre hyperparamater tuning)

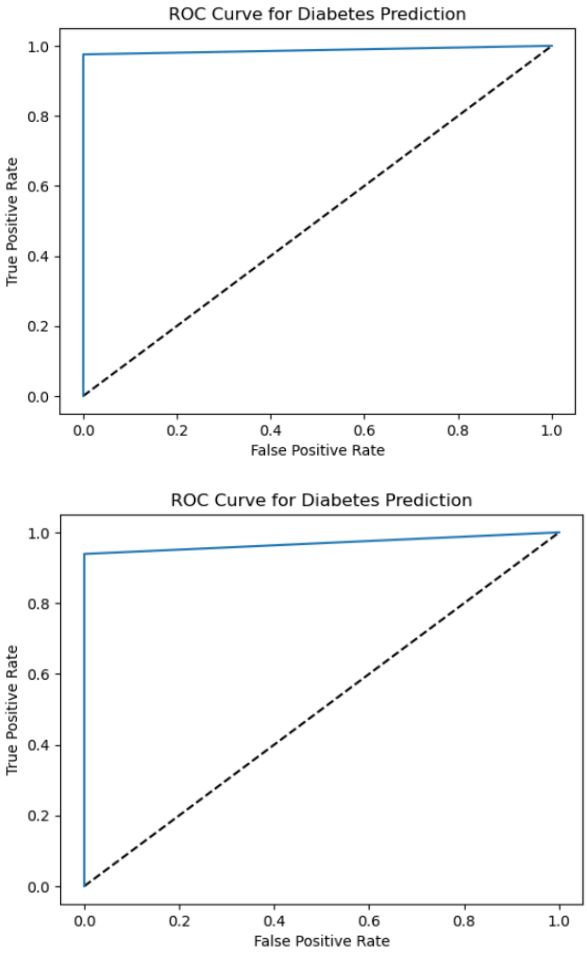


Figure 11: ROC results - “SVM Model 2.0” (Top) and “Random Forest Model 1.0” (Bottom)

	precision	recall	f1-score	support
0	0.96	1.00	0.98	48
1	1.00	0.98	0.99	82
accuracy			0.98	130
macro avg	0.98	0.99	0.98	130
weighted avg	0.99	0.98	0.98	130

	precision	recall	f1-score	support
0	0.91	1.00	0.95	48
1	1.00	0.94	0.97	82
accuracy			0.96	130
macro avg	0.95	0.97	0.96	130
weighted avg	0.97	0.96	0.96	130

Figure 12: Precision and Recall results - “SVM Model 2.0” (Top) and “Random Forest Model 1.0” (Bottom)

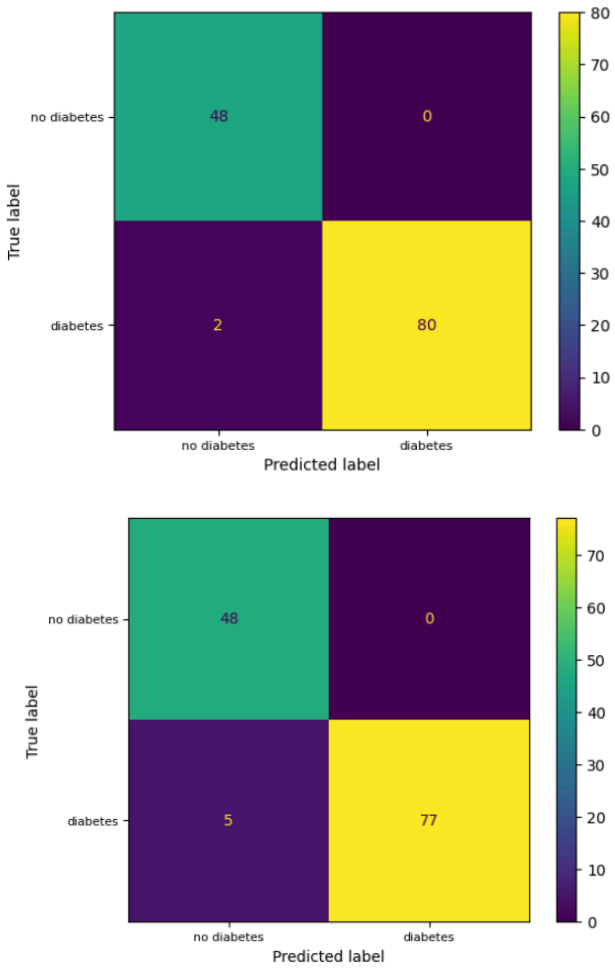


Figure 13: Confusion Matrix results - “SVM Model 2.0” (Top) and “Random Forest Model 1.0” (Bottom)

Figure 11 visualizes the ROC curve, where a true positive rate that’s at 1 or 100% shows that the model is a perfect classifier and is able to provide the frequency for a true positive classification for diabetes diagnosis, while the false positive rate would be 0 or 0% and would show the frequency for a false positive classification for diabetes diagnosis . “SVM Model 2.0” (Top) had an ROC score of 0.98, while “Random Forest Model 1.0” (Bottom) had an ROC score of 0.96.

Figure 12 shows the summary of results around the evaluation of the two models pertaining to precision, recall, and F-1 score for any unseen data (test data). Precision measures how many of the predicted positive instances of diabetes were actually positive diagnosis. Recall (or Sensitivity) measures how many of the actual positive instances for diabetes were correctly identified. The F1 score is the average or mean of the precision and recall results. When evaluating the performance of the models, the accuracy, macro average and the weighted average of precision, recall, and F1 score were compared. “SVM Model 2.0” (Top) had an accuracy result of 98%, while “Random

Forest Model 1.0” had an accuracy result of 96%. “SVM Model 2.0” (Top) had a macro average result of 0.98-0.99 respectively for Precision, Recall, and F1-Score**,** while “Random Forest Model 1.0” (Bottom) had a macro average result of 0.95-0.97 respectively across the Precision, Recall, and F1-Score.

Figure 13 compares the result of the confusion matrix result between “SVM Model 2.0” (Top) and “Random Forest Model 1.0” (Bottom). For “SVM Model 2.0” (Top), The model correctly predicted 48 instances as class 0 (negative class), no instances were incorrectly predicted as class 1 when they were actually class 0, two instances were incorrectly predicted as class 0 when they were actually class 1, and correctly predicted 80 instances as class 1 (positive class). The model demonstrates **exceptional performance**, achieving a high accuracy of **98%** and nearly perfect precision, recall, and F1-scores. The “SVM Model 2.0” (Top) had the most optimal performance, showing improvements in every metrics when it was evaluated using unseen data (test data).

Future Considerations:

As new technologies evolve around prevention and treatment of diabetes, the final developed model will serve as a simplified tool that aims to enable the predictions for the diagnosis of diabetes. For future development and deployment of the model, it would be most optimal to evaluate the model's performance on a greater population of patients who may be potentially experiencing any of the symptoms. In addition, expanding the model to a more broad population of patients would also open the opportunity for incorporating ensemble methods, if the model’s performance is impacted by expanding the number of patients and locations. However, this model achieves savings in the number and types of responses required for patients to predict the likelihood of developing diabetes.