AOS SCI C111 Project

Early Stage Diabetes Risk Prediction

Vivian Lee
UID: 005-718-435
December 6, 2024

Abstract

Machine learning methods such as Logistic Regression, Random Forest, and Support Vector Machine will be used to determine whether a person is diabetic or not based on their medical history and demographics.

1 Introduction

Diabetes is a chronic illness where blood sugar levels are higher than normal due to either an insufficient production or use of insulin. This can lead to further health problems, ranging from heart and blood disease, vision loss, kidney damage, and more. As reported by the World Health Organization (WHO) in 2022, this disease affects 38.4 million people in the United States and 830 million people worldwide. [1]

Given the prevalence and severe long-term effects of diabetes, it is crucial to understand the importance of early detection and recognize the signs and symptoms. Using the Kaggle dataset titled “Early Stage Diabetes Risk Prediction”, I trained Logistic Regression, Random Forest, and Support Vector Machine models on five data subsets: all features, male, female, young, and old subsets. The main objective of this project was to determine the indicators most closely associated with a positive diabetes diagnosis. Below I go into detail and discuss the data preprocessing steps, experimental design, and results and analysis.

2a Exploratory Data Analysis

The dataset used in this study was provided from Kaggle but originally came from the University of California, Irvine (UCI) Machine Learning Repository. [2] The information was collected using direct questionnaires from 520 patients of Sylhet Diabetic Hospital in Sylhet, Bangladesh and approved by a doctor. As outlined in Table 1, there were 14 features that described the individuals’ various symptoms, such as polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital thrush, and visual blurring. Additional attributes included itching, irritability, delayed healing, partial paresis, muscle stiffness, alopecia, and obesity. The other two features described demographic characteristics, which were age and gender. For our modeling purposes, these will be turned into categorical features, where we will explore how the diabetes diagnosis depends on age and gender, respectively.

Variable	Original Values (Unique)	Final Values
Age	16 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 72 79 85 90	15 - 47, 47 - 90
Gender	1. Male, 2. Female	1, 0
Polyuria	1. Yes, 2. No	1, 0
Polydipsia	1. Yes, 2. No	1, 0
Sudden Weight Loss	1. Yes, 2. No	1, 0
Weakness	1. Yes, 2. No	1, 0

Polyphagia	1. Yes, 2. No	1, 0
Genital Thrush	1. Yes, 2. No	1, 0
Visual Blurring	1. Yes, 2. No	1, 0
Itching	1. Yes, 2. No	1, 0
Irritability	1. Yes, 2. No	1, 0
Delayed Healing	1. Yes, 2. No	1, 0
Partial Paresis	1. Yes, 2. No	1, 0
Muscle Stiffness	1. Yes, 2. No	1, 0
Alopecia	1. Yes, 2. No	1, 0
Obesity	1. Yes, 2. No	1, 0
Class	1. Positive, 2. Negative	1, 0

Table 1: Description of Variables

To further understand the relationship between the various attributes and the target variable as well as the structure of the data, I performed Exploratory Data Analysis (EDA) by plotting the main characteristics visually.

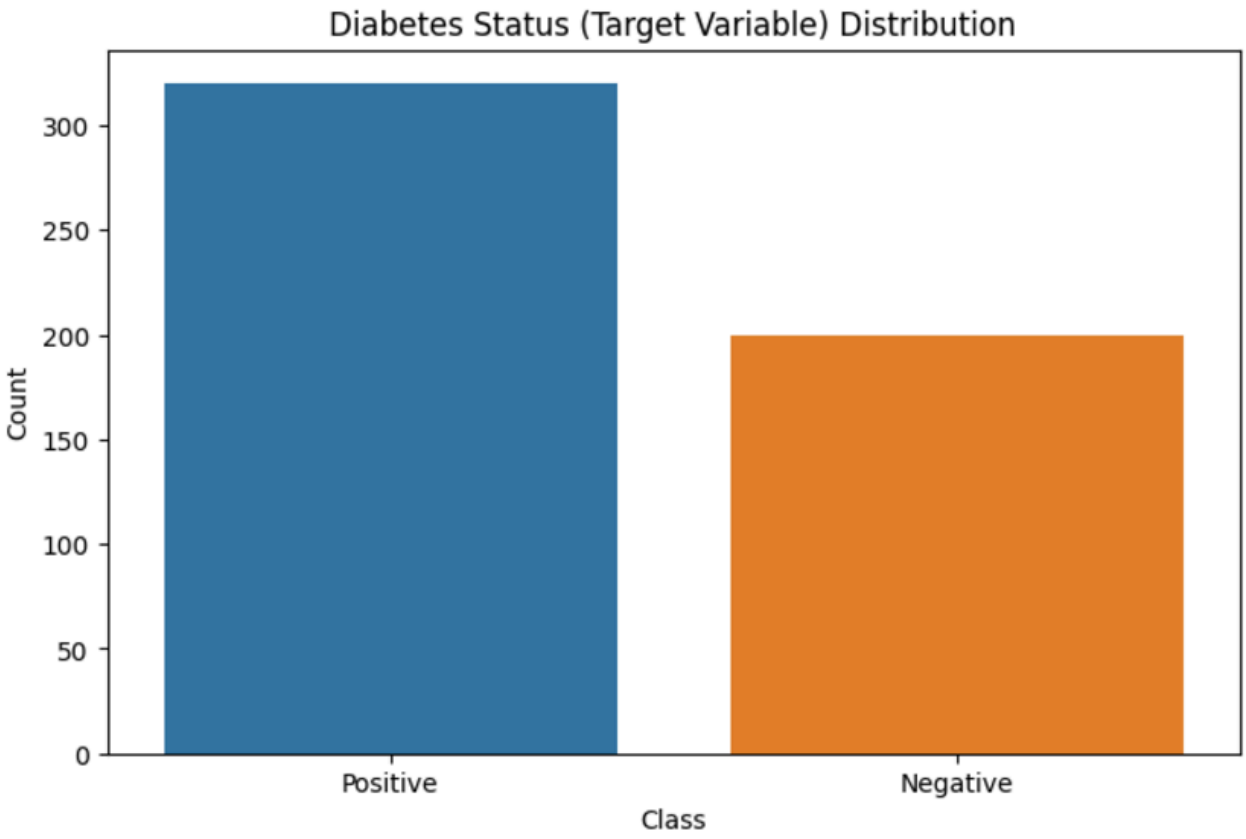

Figure 1: Bar Chart - Diabetes Status Distribution

The target variable column of diabetes status was slightly skewed and imbalanced, with the positive class having 320 instances but the negative class having 200 instances. Thus it was important to be careful in the upcoming modeling procedures to ensure that the model does not solely predict the majority class. However, because there was only a slight imbalance, I decided to train the models under the assumption that each class of the target variable should be treated equally, where the prediction errors of either class are given the same weight.

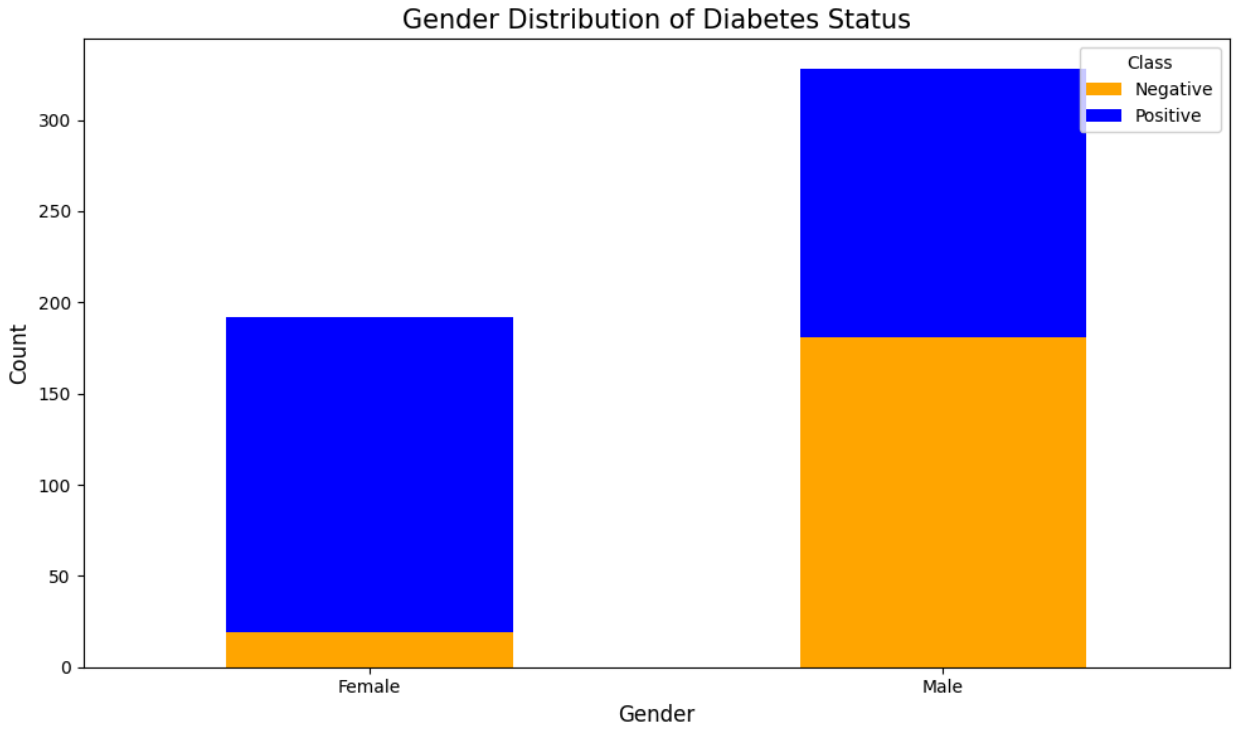

Figure 2: Stacked Bar Chart - Gender Distribution of Diabetes Status

There are much more data observations for males (328) compared to females (192). In addition, the female subset has a strong class imbalance that presents bias towards the majority class, which is positive. This may lead to skewed evaluation metrics during modeling, such as incorrectly displaying high accuracy. These will be taken into consideration in our analysis and results later.

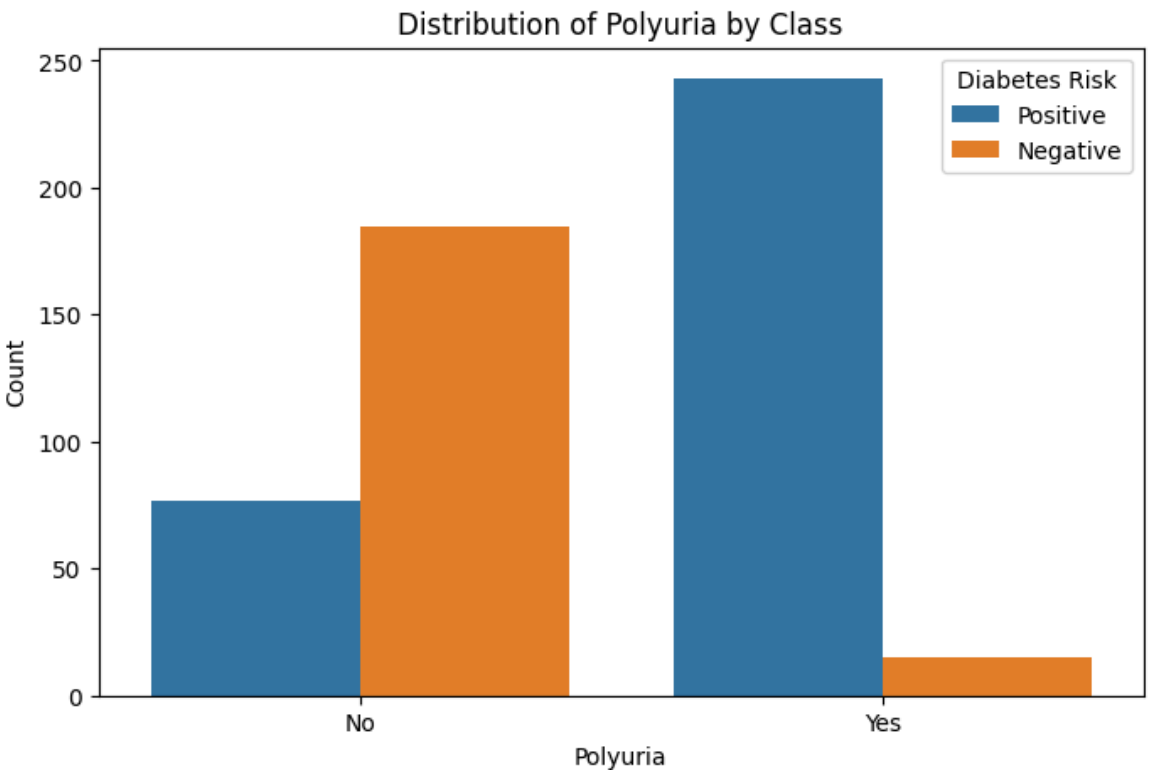

Figure 3: Grouped Bar Chart - Polyuria Distribution of Diabetes Status

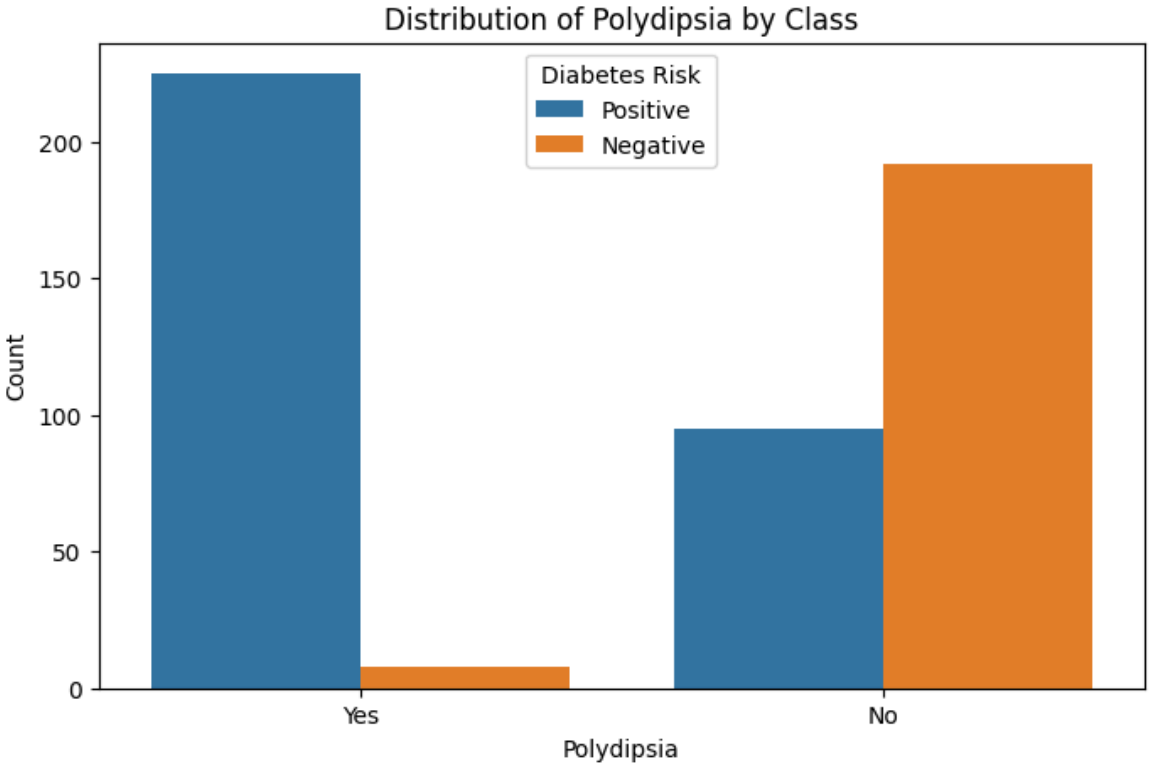

Figure 4: Grouped Bar Chart - Polydipsia Distribution of Diabetes Status

From all of the grouped stacked bar charts that display the distribution between each of the 14 symptom variables and diabetes status, I found that the features polyuria and polydipsia appeared to have the most significant impact on the target variable. More specifically, patients who were diagnosed with polyuria and polydipsia were highly likely to also have diabetes. For background context, polyuria means excessive urine production, while polydipsia means excessive thirst or drinking. [3, 4] Clearly, these conditions are interdependently connected as they both affect the urinary system.

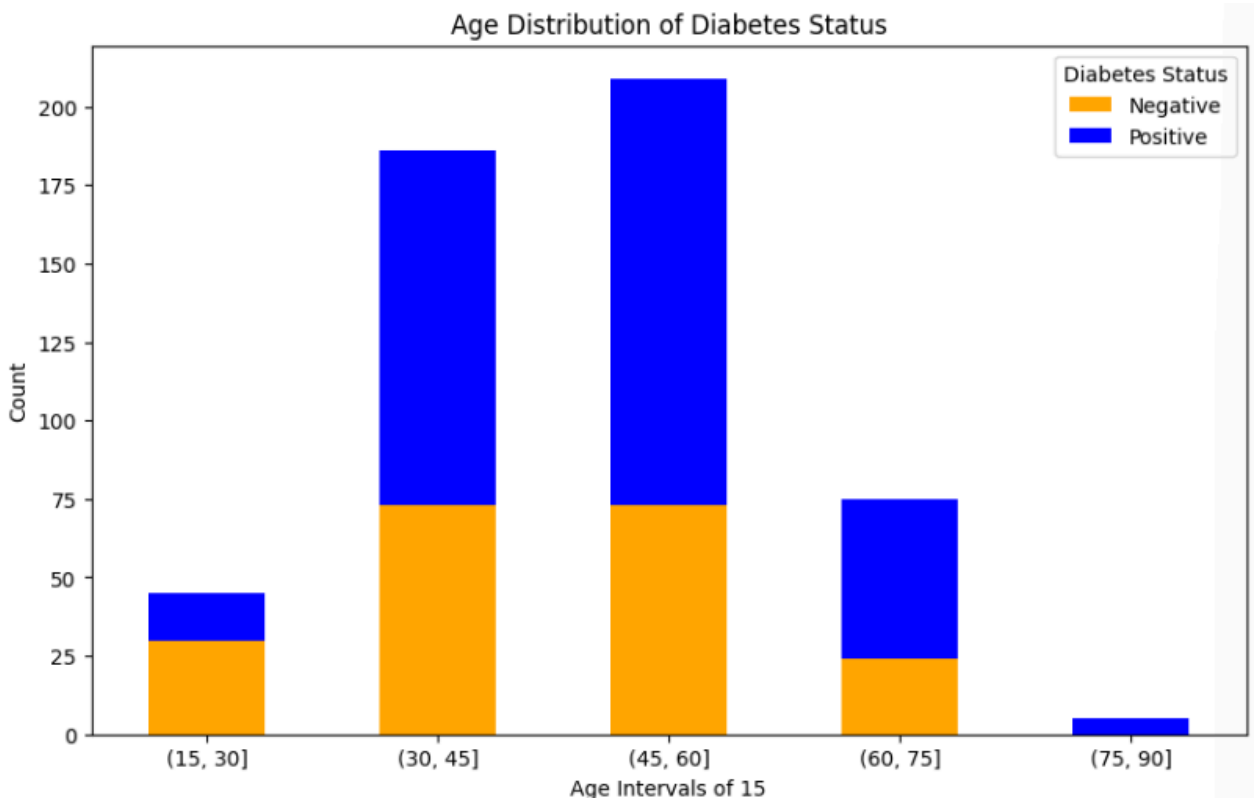

Figure 5: Stacked Bar Chart - Age Distribution of Diabetes Status (5 Intervals)

Age Groups	Positive Class Count	Negative Class Count	Total
15-30	15	30	45
31-45	113	73	186
46-60	136	73	209
61-75	51	24	75
76-90	5	0	5

Table 2: Class Count of Age Group (5 Intervals)

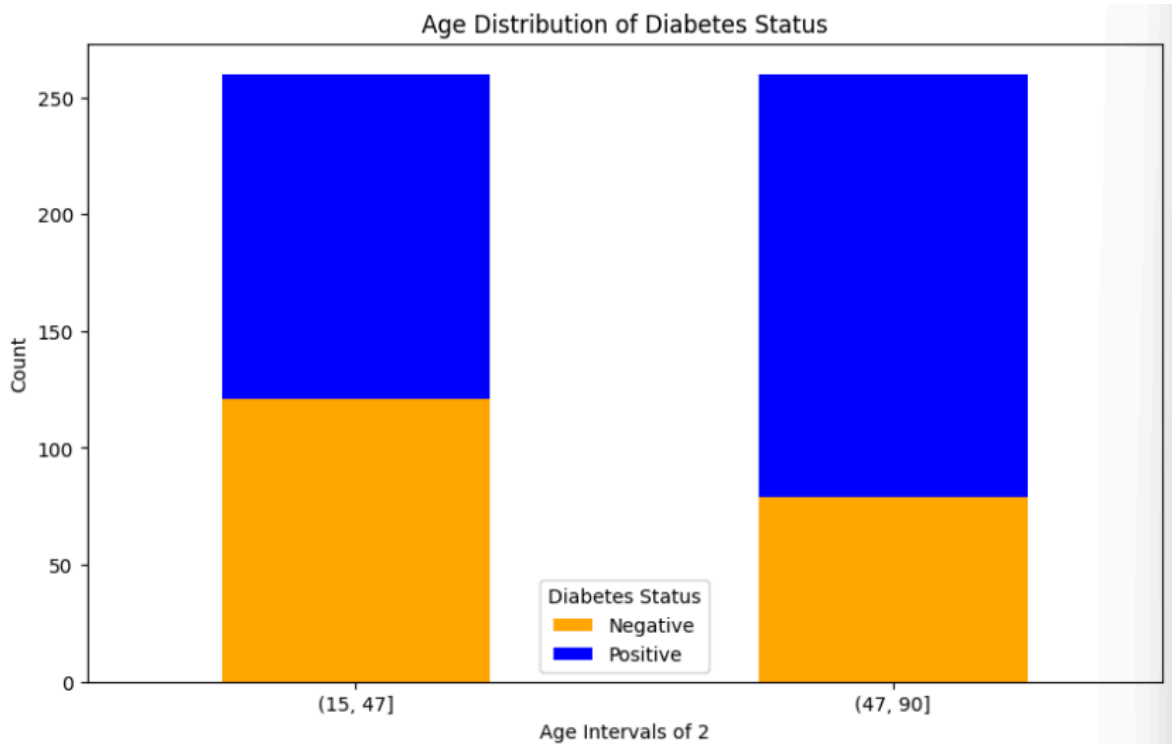

Figure 6: Stacked Bar Chart - Age Distribution of Diabetes Status (2 Intervals)

Age Groups	Positive Class Count	Negative Class Count	Total
15-47	139	121	260
48-90	181	79	260

Table 3: Class Count of Age Group (2 Intervals)

As the only numeric column, age provided a wider range of information due to its granularity. Delving into the summary statistics, the ages ranged from 16 to 90 with a standard deviation of approximately 12.1514 and an average age of 48.0288.

Now knowing the summary statistics of this demographic variable, I first tried to bin the data into five different age groups with an interval of 15 (1. 15-30, 2. 31-45, 3. 46-60, 4. 61-75, 5. 76-90). I chose this specific interval width because it was close to the calculated standard deviation of 12. However, Figure 5 and Table 2 clearly demonstrated that not all age groups have sufficient data, especially ages 15-30 and 76-90. Specifically, the oldest age group had no instances of the negative class at all, most likely due to increased mortality rates and decreased life expectancy for individuals over 75 years old. The lack of balanced representation would cause overfitting to the positive class as the model would have no basis to learn the distinction between the two classes. Thus I decided to instead split the age group in half according to the median age of 47.5, separated as the young (15-47 years) and old (48-90 years) group as displayed in Figure 6 and Table 3. Now, age can be utilized as an effective categorical feature.

2b Cleaning & Preprocessing

The dataset file was initially loaded as a dataframe object into a Google Colab notebook. Tuples with incomplete data were already discarded by previous research on the subject, so there were no missing values that had to be addressed through the imputation or removal process. In addition, the majority of the features and the target variable were binary, characterized by "Yes/No" or denoted as "Positive/Negative." Thus I converted these values to 1 and 0, respectively. Lastly, I created four different subsets that separated the data by gender (male and female) and age (young and old) as outlined above during the Exploratory Data Analysis.

3 Modeling

All of the data provided were labeled, where both the input features and the corresponding output feature were given. Thus three main techniques under the supervised learning were selected: Logistic Regression, Random Forest, and Support Vector Machine. The approaches for fitting each model as well as its advantages and disadvantages are outlined below.

1. **Logistic Regression.** Logistic Regression is a discriminative model which estimates the conditional probability that a sample x_n belongs to the positive class 1, denoted by $P(Y = 1 \mid X = x_n)$. The coefficients represent the log odds, providing insight into the relationship between predictors and the target and making the algorithm simple to implement and understand. The model performs poorly with the existence of outliers in the data, or when its assumptions of a linear relationship and independent features are violated. [5, 6]

2. **Random Forest.** Random Forest is an ensemble learning method that aggregates predictions from a set of decision trees given random subsets of the data and classifies based upon the majority vote of the trees. The result is thus significantly better than any one classifier on its own, improving prediction accuracy and reducing overfitting with its robustness to noise. However, the algorithm can be computationally expensive, time consuming, as well as being difficult to understand how specific features contribute to individual predictions. [5, 7]

3. **Support Vector Machine (SVM).** Support Vector Machine is a large margin classifier that finds the hyperplane, which is the optimal dividing line that separates data points into different classes by maximizing the margin between classes. The technique performs well with limited data, which aligns with our problem since we lack samples for the negative class. Some limitations include its computational complexity, difficulty in parameter tuning (i.e. kernel selection), and sensitivity to feature scaling. [5, 8]

Using the default parameters, each algorithm evaluated five data sets: all features, male, female, young, and old subsets. With the general modeling framework identified, I proceeded to randomly split each dataset into 80% training data and 20% testing data. After training each of the three models, I assessed each on the testing data, and the accuracy of each model was given in Table 4. Other evaluation metrics included classification reports (Tables 5, 6, 7), confusion matrices (Figure 7), ROC curves (Figure 11), and feature importance (Figures 8, 9, 10).

4 Results and Discussion

In general, the best result was achieved using the Random Forest algorithm, where 97.11% instances were classified correctly using the dataset with all features. Logistic Regression had the lowest accuracy at 90.38%, while Support Vector Machine performed in the middle with a 94.23% accuracy rate. The testing accuracy values from Table 4 were supported by the Receiver Operating Characteristic (ROC) curve shown in Figure 11. This graph plotted the model's sensitivity, or True Positive Rate (TPR), against its specificity, or False Positive Rate (FPR). The Area Under the Curve (AUC) represented the model's overall accuracy in distinguishing between classes. Again, Random Forest had the highest AUC, while Logistic Regression had the lowest. All of these outcomes were consistent with our prior understanding of each method's strengths and weaknesses. Random Forest integrated results from multiple different classifiers to accurately capture non-linear patterns and interactions in the data. On the other hand, Logistic Regression struggled with complex, non-linear relationships.

From all the datasets characterized by demographics, the Young subset had the lowest performance across all models, with the lowest being 84.61% testing accuracy for logistic regression. This can be attributed to the subset containing more variability or noise, making it difficult for the models to generalize. Another important point of observation was when the model achieved 100% testing accuracy. There were three instances where this took place: when the male subset was tested by Random Forest and Support Vector Machine, and when the old subset was tested by Random Forest. This most likely occurred due to overfitting and low complexity of the limited training set.

Dataset	Logistic Regression Testing Accuracy	Random Forest Testing Accuracy	Support Vector Machine Testing Accuracy
All Features	0.9038	0.9711	0.9423
Male Subset	0.9393	1.0000	1.0000
Female Subset	0.9487	0.9743	0.9743
Young Subset	0.8461	0.9807	0.9038
Old Subset	0.9423	1.0000	0.9807

Table 4: Testing Performance across all Models and Datasets

	Precision	Recall	F1 Score	Support
0	0.85	0.89	0.87	38
1	0.94	0.91	0.92	66
Accuracy			0.90	104

Macro Avg	0.89	0.90	0.90	104
Weighted Avg	0.91	0.90	0.90	104

Table 5: Classification Report of Logistic Regression for Testing Data (All Features)

	Precision	Recall	F1 Score	Support
0	0.97	0.95	0.96	38
1	0.97	0.98	0.98	66
Accuracy			0.97	104
Macro Avg	0.97	0.97	0.97	104
Weighted Avg	0.97	0.97	0.97	104

Table 6: Classification Report of Random Forest for Testing Data (All Features)

	Precision	Recall	F1 Score	Support
0	0.92	0.92	0.92	38
1	0.95	0.95	0.95	66
Accuracy			0.94	104
Macro Avg	0.94	0.94	0.94	104
Weighted Avg	0.94	0.94	0.94	104

Table 7: Classification Report of Support Vector Machine for Testing Data (All Features)

The confusion matrices from Figure 7 were used to create the classification reports for each algorithm, provided above in Tables 5, 6, and 7. From the precision and recall results, the model appeared to be slightly better at predicting labels of 1 (positive diabetes status) than 0 (negative diabetes status). In a future revision, it would be beneficial to explore balancing classes to ensure that the models are not biased toward predicting the majority class.

Feature importance for each model was calculated and ranked in Figures 7, 8, and 9 as well. The two most relevant features were consistently polydipsia and polyuria, which correctly aligned with the insights identified earlier from the Exploratory Data Analysis. Another insight was that Random Forest and Support Vector Machine both ranked Age and Gender as significant features.

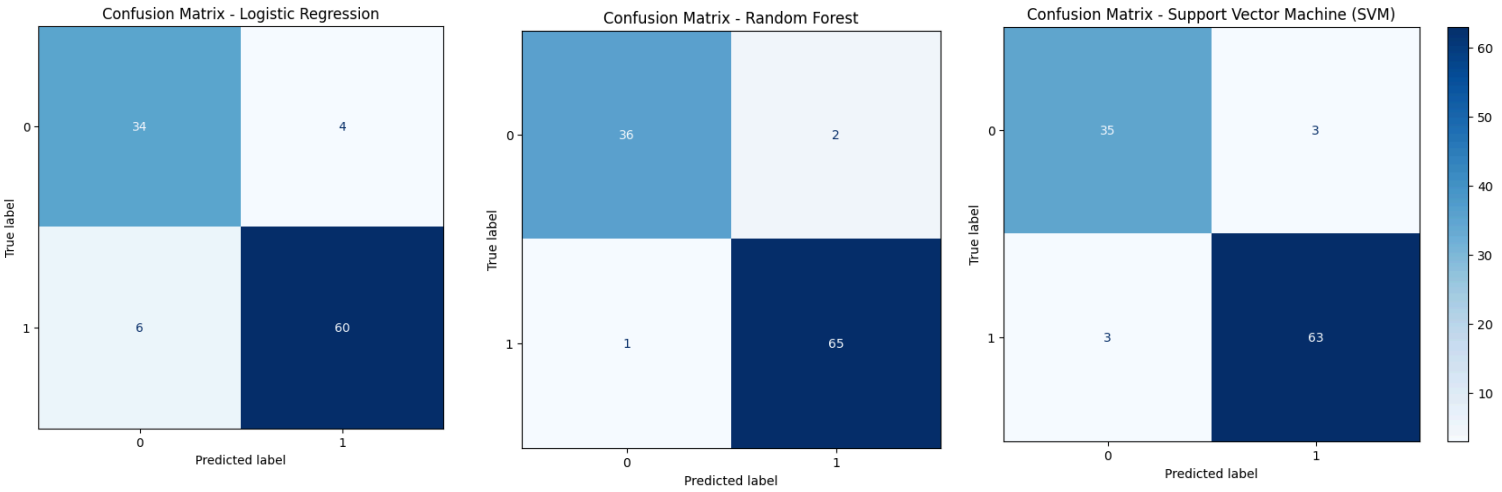

Figure 7: Confusion Matrices (All Features)

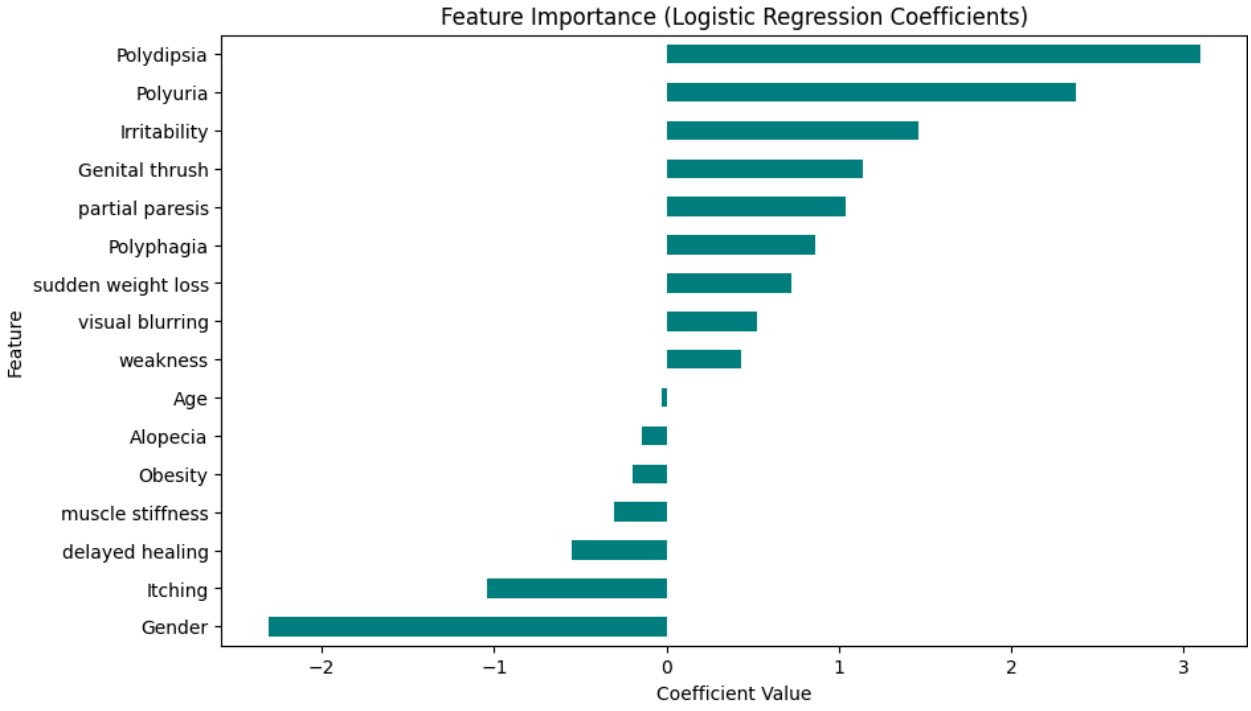

Figure 8: Feature Importance of Logistic Regression (All Features)

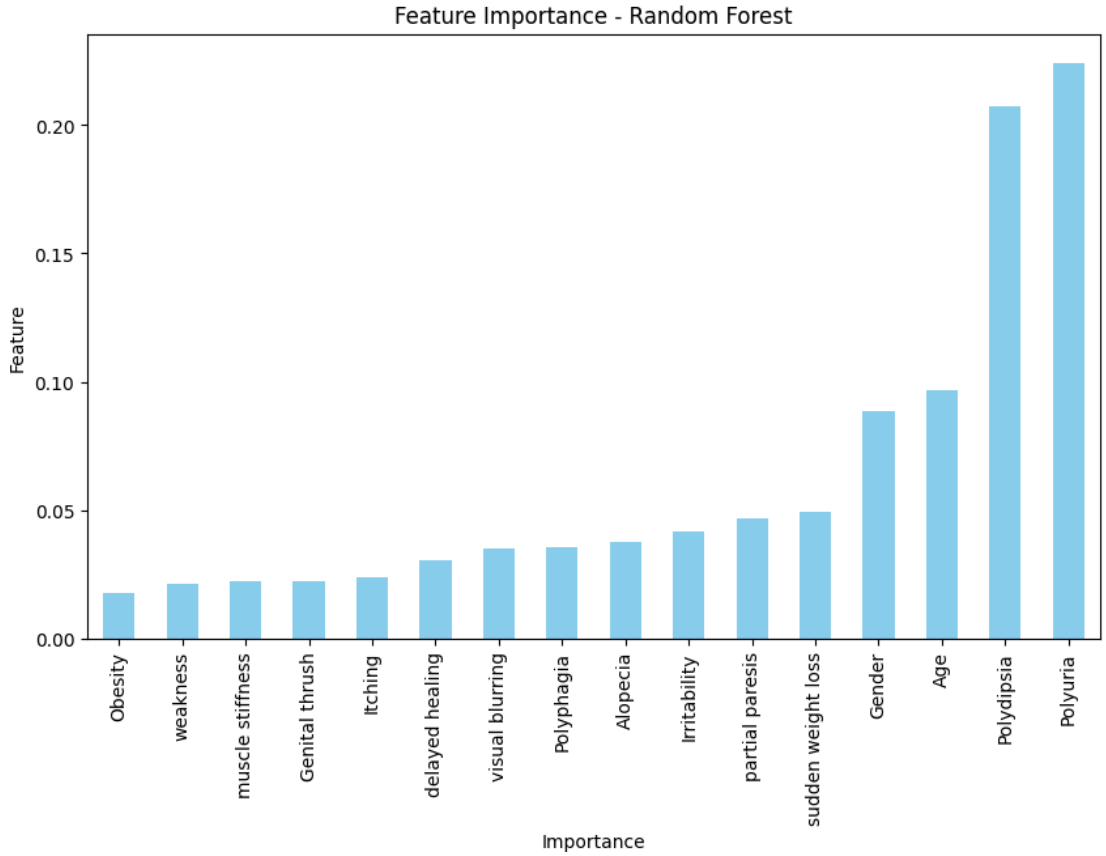

Figure 9: Feature Importance of Random Forest (All Features)

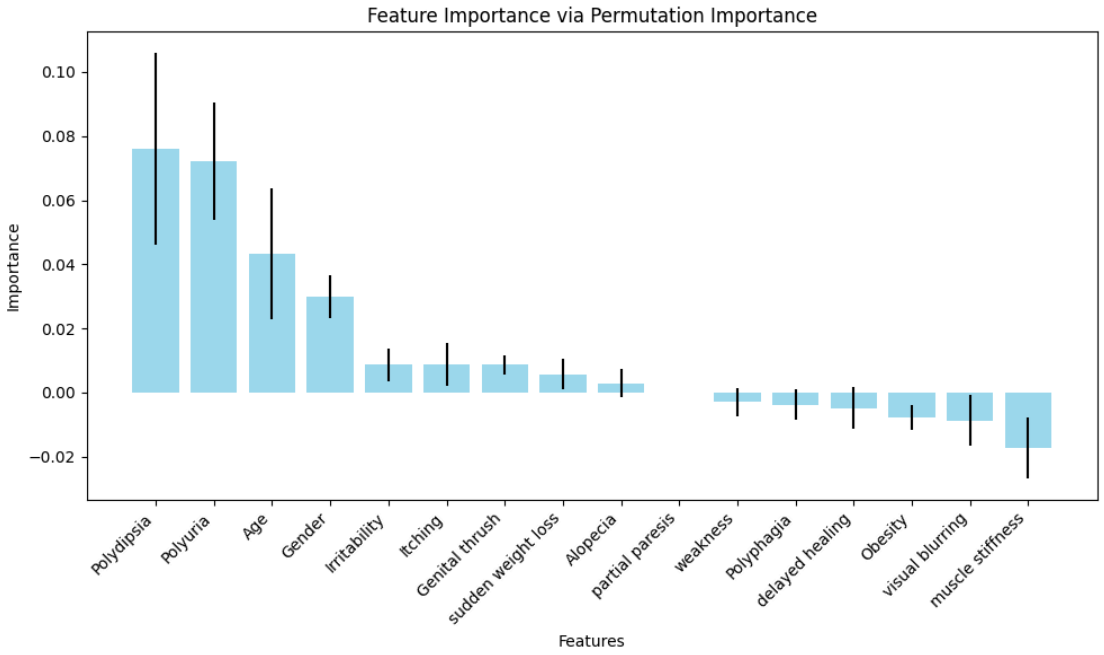

Figure 10: Feature Importance of Support Vector Machine (All Features)

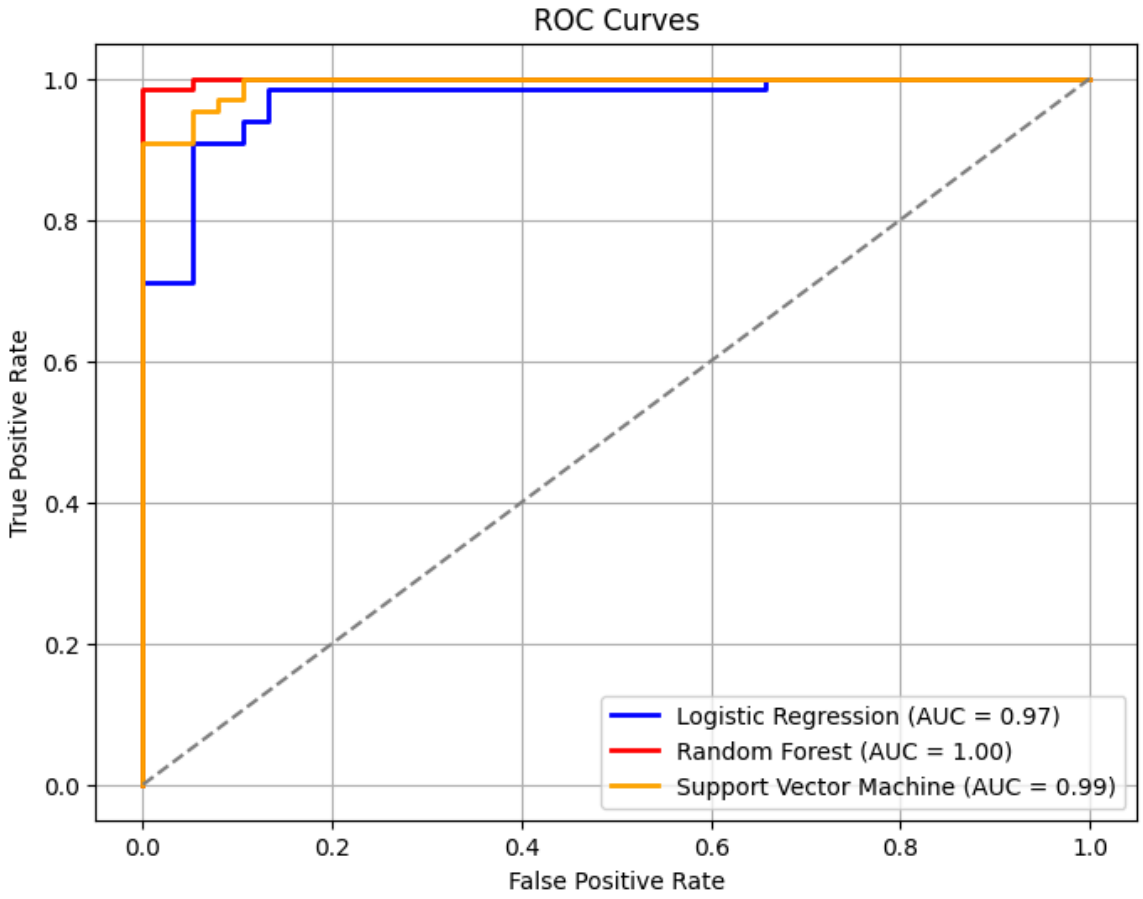

Figure 11: ROC Curves (All Features)

6 Conclusion

Some areas of improvement include having access to more data to increase model performance in terms of testing accuracy. This includes having more class balance for the target variable of the diabetes status column, and simply having more samples for each demographic group, especially females and individuals over 75 years old. Having this information would help see which age/gender group the model is best at predicting.

To expand this project further, I could explore hyperparameter tuning of the chosen three models. For Logistic Regression, this would include regularization strength, learning rate, and number of iterations; for Random Forest, it would be bagging, splitting, and maximum leaf nodes; and lastly for Support Vector Machine, it would comprise of the regularization parameter, kernel coefficient, and kernel function. Another avenue could be using other classification algorithms in supervised machine learning, namely K-Nearest Neighbors using Hamming distance for binary features.

This project provided a practical, hands-on application of a machine learning problem applied to a real-life setting. These kinds of projects utilizing patient data have the ability to potentially improve diagnostic accuracy and patient outcomes, allowing for timely interventions and encouraging at-risk individuals to make lifestyle changes.

7 References

[1] *Diabetes*. (2024). World Health Organization. https://www.who.int/news-room/fact-sheets/detail/diabetes

[2] Early Stage Diabetes Risk Prediction [Dataset]. (2020). UCI Machine Learning Repository. https://doi.org/10.24432/C5VG8H.

[3] Painter, K. Fowler, P. *Polyuria (Excessive Urine Production)*. (2024). WebMD. https://www.webmd.com/diabetes/polyuria-too-much-urine

[4] *Polydipsia.* (2022). Cleveland Clinic. https://my.clevelandclinic.org/health/symptoms/24050-polydipsia

[5] Bortnik, J. (2020). *Introduction to Machine Learning for the Physical Sciences* [PowerPoint presentation]. Canvas. https://bruinlearn.ucla.edu/courses/195891/modules

[6] Lawton, G. Burns, E. Rosencrance, L. (2022). *What is logistic regression?* Business Analytics. https://www.techtarget.com/searchbusinessanalytics/definition/logistic-regression

[7] *What is Random Forest?* IBM. https://www.ibm.com/topics/random-forest

[8] *What is Support Vector Machine (SVM)?* (2017). TechTarget. https://www.techtarget.com/whatis/definition/support-vector-machine-SVM