**Research Statement and Objectives**

This project is created to determine which symptoms are the highest predictors when diagnosing positive diabetes cases within this dataset. The dependent variable in this dataset is class or whether a patient was diagnosed positive or negative for diabetes. The independent variables include Age, Gender, Polyuria, Polydipsia, sudden weight loss, weakness, Polyphagia, Genital thrush, visual blurring, Itching, Irritability, delayed healing, partial paresis, muscle stiffness, Alopecia, and Obesity. These variables will be defined in the table below. ADD TABLE ADD HYPOTHESIS

**Background ADD**

**Method**

An opensource dataset was collected through the UCI Machine Learning Repository from a study done in 2019 using direct questionnaires on patients in Sylhet Diabetes Hospital in Sylhet, Bangladesh which was approved by a physician. This dataset was then analyzed by initially creating bar plots of the dependent variable (class) over the independent variables (various symptoms of diabetes, age, and gender) to view the distributions of the variables over the number of positive cases. After bar plots were created as well as a histogram to look at the distribution of age over the total sample.

To understand the determinants or important features that help in predicting the diagnosis of diabetes as positive or negative classes, we employed two different modeling approaches. First, we used logistic regression model to understand relationship between predictor and dependent variables and then we used decision tree model to understand the feature importance.

To fit a logistic regression model, we used stepwise regression approach to find out the model with the best set of predictors. Once we identified the final set of predictors, we fit a logistic regression model and calculated the accuracy and sensitivity of the model. The coefficients and significances of coefficients in the logistic regression helped us understand the key predictors of diabetic class.

After fitting a logistic regression model, we fit a decision tree model and calculated the accuracy and sensitivity of the model. The feature importance plot was used to identify the key predictors of diabetic class. We also compared the models based on accuracy and sensitivity and suggest used the better model to answer our hypotheses.

**Results and Analysis**

An initial analysis was done to get a better understanding of the dataset. A histogram was created to see the distribution of the participants and compare their ages in the study. It was seen that there were a wide range of participants ranging from 16 years of age to 90 years with the 45-50 years range being the most predominant. Both genders were present in the study (a third gender was not documented). From the bar plot depicting the gender distribution over the total number of participants, it can be seen that there were fewer females that participated over males; male participants made up 328/520, or 63% of the sample whereas the female participants made up 128/520, or 24.6% of the sample. In terms of class or the number of positive and negative cases for Diabetes, within the sample there were 320 positive cases for Diabetes whereas the rest of the sample was negative for the disease. When looking at the individual symptoms of Diabetes that were tested for, Polyuria (excessive urination) showed that there was not much of a difference in cases showing that there were around the same number of positive cases as negative with 258 positive cases (49.6% of the sample). When inquired about Polyphagia, or excessive eating of food, data showed that the number of participants that did not have the symptom were higher (283). The number of individuals that reported Polydipsia, or abnormally great thirst were 233 out of the 520 participants with more individuals reporting “no” than “yes”. When asked whether participants had sudden weight loss, the majority of participants had reported “no” with 217 of the sample reporting “yes”. In this particular sample, researchers found that 305 of the participants reported “yes” to feeling weakness and 116 of the population reported “yes” to Genital thrush or yeast infections. When asked about visual blurring, 233 of the participants reported “yes” along with 253 reporting “yes” to Itching. The study also asked participants about other symptoms including Irritability which 126 reported “yes” to; delayed healing which 239 reported “yes” to; 224 reported they experienced partial paresis which is a symptom where individuals feel a weakening for a muscle or group of muscles (also referred to as a mild paralysis). Data regarding prevalence of muscle stiffness showed that 195 individuals in the sample reported experiencing this symptom and 179 individuals responded with “yes” for Alopecia or hair loss. Lastly, 88 individuals reported “yes” to Obesity.

Chart, histogram

Description automatically generatedChart, bar chart

Description automatically generatedChart

Description automatically generated

Chart

Description automatically generatedChart, shape, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generatedChart, shape, bar chart

Description automatically generated

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

After creating bar plots to evaluate distributions of age, gender, positive/negative cases, as well as prevalence of symptoms in the sample, the sample was converted into a binary format in order to work with data more easily making all “yes” responses as 1 and “no” as 0 and dividing gender into “0” and “1” as well with “1” representing males and “0” representing females. A logistic regression linear model was then used to examine the relationships between the dependent and independent variables. A stepwise procedure was taken to select which independent variables among the various symptoms, age, and gender should be included in the final model for analysis. A reduced version of the stepwise logistic regression was first taken with a boundary of 1, containing no independent variables. The output of this iteration is shown in *Table 1.* Text

Description automatically generated

In this the AIC value (Alkaline Information Criteria) was 694.93. In the full stepwise regression iteration, all the independent variables were included (Age, Polyuria, Polydipsia, sudden weight loss, weakness, Polyphagia, Genital Thrush, visual blurring, itching, irritability, delayed healing, partial paresis, muscle stiffness, Alopecia, and Gender). The output of this iteration is depicted in *Table 2*. A screenshot of a computer

Description automatically generated with low confidence

Text

Description automatically generated

The AIC value of this iteration is computed to be 205.65 with this being more optimal than the previous reduced iteration. In this iteration, Polyuria, Polydipsia, Itching, Irritability, and Gender had the highest number of significant codes (3) and partial paresis, polyphagia, and age showing significance of 0.01. These two iterations were useful in determining the boundaries for the stepwise function. When the full stepwise regression was run, adding independent variables one by one to find the most accurate model with the highest predictors, the last model with the independent variables: Polyuria, Polydipsia, Gender, Itching, Irritability, Genital thrush, partial paresis, Polyphagia, Age, and weakness was the most accurate with 198.36 as the AIC value (the lowest out of the previous iterations). This is depicted in *Table* 3.

Table

Description automatically generated

From this it is clear that the highest significant predictors for this dataset include Polyuria, Polydipsia, Itching, Irritability, partial paresis, Age, Genital thrush, Polyphagia, weakness and Gender.

After completing the stepwise regression and figuring out which were the highest predictors, the average marginal effects were calculated over the final (optimal) stepwise iteration. These marginal effects helped to see the average marginal effect on probability of diagnosing as positive class for a one unit change in the independent variable or if the independent variable is categorical then it refers to the change in average marginal effect on probability of diagnosing as positive class for category 1 compared to reference category of the variable. Through this it is seen that Polydipsia has the highest average marginal effect on the probability of testing positive for diabetes (depicted in *Table 4)*.

Text

Description automatically generated

The next step of the analysis included calculating the predicted classes using logistic regression model to create a confusion matrix. The confusion matrix is depicted in *Table 5*. From these calculations it is seen that the accuracy is 93% ((183+301)/520) and the sensitivity is (301/(301+19)) giving 94.06%. This accuracy and sensitivity will then be compared to the tree model done on the same predictors and dependent variable.

A picture containing diagram

Description automatically generated

A tree model was used to compare which model was most accurate in predicting variables associated with class(whether an individual had diabetes or not. Decision trees allow for a visual representation of the decision-making process. These models are helpful when determining which strategy should be used to reach a particular goal. In this case, the tree model was used to depict the decision-making process when predicting diabetes in an individual. This tree model is depicted in *Table 6*. Diagram

Description automatically generated

From this it can be seen that one must go through the symptoms starting with Polyuria, then Polydipsia, then partial paresis, then visual blurring, and lastly delayed healing. After the decision tree was created, feature importance was established and plotted in order to see which symptoms were calculated to have the most importance when being diagnosed with diabetes. Chart

Description automatically generated

From the figure it is seen that Polyuria, Polydipsia, sudden weight loss, visual blurring, partial paresis, and Polyphagia had the highest importance. Another confusion matrix was calculated for this model in order to calculate the accuracy and sensitivity like what was done in the previous model. The accuracy of the model was calculated to be 89.8% ((177+290)/520) and the sensitivity was calculated to be 90.62% (290/(290+30)). Comparing these calculations to the linear model created previously it can be seen that the logistic regression model had higher accuracy and sensitivity compared to tree-based model. So, we use the logistic regression model to the importance of predictors and also make predictions of diabetes class.

A picture containing text

Description automatically generated

**Conclusion ADD**