Problem 2 – Logistic regression (35%)

Using the dataset 'diabetes.csv' included in the assignment, develop a logistic regression model to predict whether a patient has developed diabetes as a function of the remaining variables in the dataset.

References:

- Logistic Regression using Statsmodels - GeeksforGeeks
- Logistic Regression Using StatsModels
- How to Interpret Pr(>|z|) in Logistic Regression Output in R - Statology

15% credit. Estimate the classification rate on validation data using k-fold cross validation. Generate a plot of the average classification rate across folds for different values of k (2,3,..10).

Before proceeding with the KFold Cross validation, I had assumed that increase of k folds would result in increased classification rate, as increasing k folds both increases the variance, but lowers the bias of the model. This is because as K goes up, the size of the partitions goes down, and thus, we end up training with larger partitions and testing with smaller partitions.

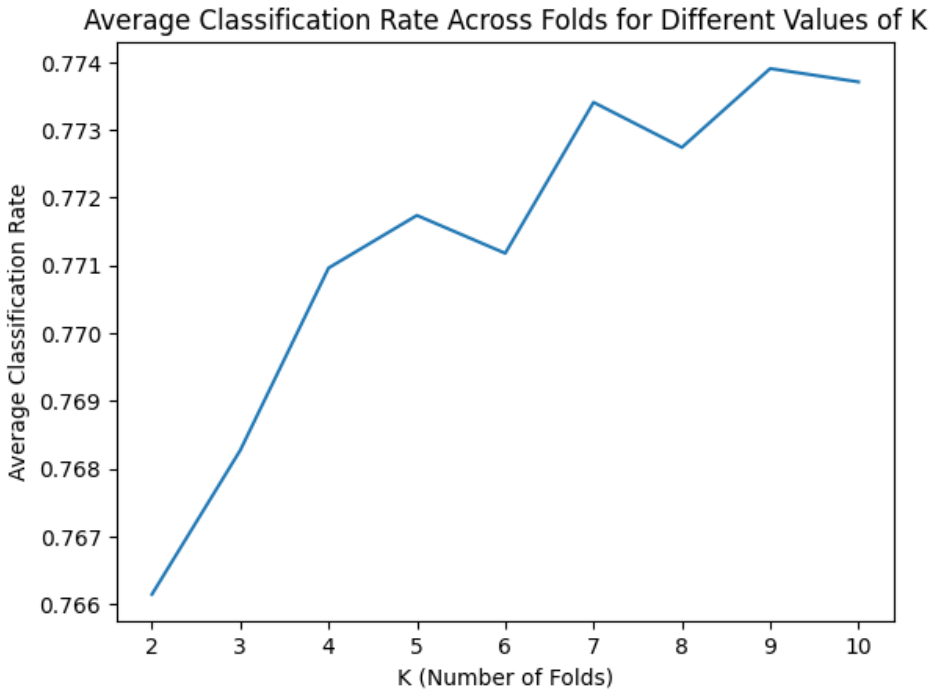

Figure 1. Average Classification Rate Across Folds for Different Values of K

Discussion of findings:

Based on Figure 1, it appears that as the number of folds increase, the average classification rate of the logistic regression model initially goes up and experiences areas of local maximums in terms of average classification rates, while also experiencing areas of local minimums where surrounding ACR (average classification rates) are higher. For example, k = 5, k=7, and k = 9 all experience local maximums in terms of average classification rate.
However, despite the growing trend of ACR increasing as the number of folds increases, the model appears to slow around k = 9, where the ACR gain following k = 9 slows. It is also important to note the y-axis of Figure 1, as while the graph suggests that the ACR of the model using 4 folds results in a ACR gain compared to that of 2 folds, upon further inspection, the average model performance only increases 0.5%, which is not ideal. Even when comparing the average model performance gain between that of 2 folds (the lowest ACR) and 9 folds (the highest ACR), the model only improves in classification rate by 0.8%.

Thus, this tells us that while at first glance, the figure suggests that cross validation using increased folds improves classification accuracy, because the increase in classification accuracy is so small between that of the best k and the worst k chosen for accuracy, cross validation does not contribute much to increase the accuracy of the logistic regression model. Typically, cross validation is used to train the model and test it on unseen data, with higher k values increasing variance in the data; however, it appears that the resulting logistic regression models generated from differing folds in cross validation are generally stable or that the data is relatively

homogenous (if the unseen data in the testing set is very similar to that of the training data, cross validation doesn't give us the advantage of allowing the model to see new patterns or gain any insight). The low performance rate may be attributed to other factors, such as overfitting or underfitting, complexity, etc.

20% credit. Interpret the regression coefficients of the model. These coefficients will vary somewhat from fold to fold, but they should be somewhat consistent. This will require you to develop some familiarity with the individual predictors and how they are related to diabetes.

	Coefficient	P-Value
Intercept	-8.1903	< 0.0001
Pregnancies	0.1392	< 0.0001
Glucose	0.0338	< 0.0001
Blood Pressure	-0.0125	0.032
Skin Thickness	0.0014	0.857
Insulin	-0.0020	0.068
BMI	0.0899	< 0.0001
Diabetes Pedigree Function	0.8519	0.011
Age	0.0121	0.233

Figure 2. Regression Coefficients of the Best Performing Model

Discussion of findings:

Before proceeding with the main objective of this discussion, we first need to analyze the individual predictors for Diabetes and their relation to a diagnosis.

- Pregnancy: During pregnancy, some may develop gestational diabetes (GD) due to increased high blood sugar. This form of diabetes does not require a patient to have diabetes prior to pregnancy, but rather appears because of the pregnancy.
- Glucose: Glucose levels play a large role in diabetes as blood sugar levels following an overnight fast and glucose tolerance tests can indicate whether or not a patient is normal, has prediabetes, or has diabetes. Because diabetes is a disease that occurs when blood glucose is too high, it is reasonable to think that glucose levels could be a good predictor of diabetes.
- Blood Pressure: John Hopkins Medicine states that “[h]igh blood pressure is twice as likely to strike a person with diabetes than a person without diabetes.” For this reason, is

it reasonable to think that BP could be a good indicator of diabetes due to this connection studied in medicine.

- Skin Thickness: If a patient is afflicted with diabetes and their blood sugar levels remain high, CDC states that "digital sclerosis can cause your skin to become hard, thick, and swollen and can spread throughout your body." Because patients who are afflicted with diabetes are more likely to experience digital sclerosis, it is viable to think that skin thickness could be a good predictor of diabetes.
- Insulin: This hormone plays a vital role in regulation of blood sugar in the body, and plays a big role in determining whether or not a patient develops Type 2 Diabetes. NHS inform states that type 2 diabetes typically occurs when "the pancreas doesn't produce enough insulin or the body's cells don't react to insulin".
- BMI: The paper The Relationship between BMI and Onset of Diabetes Mellitus and its Complications concludes that "Having even moderately elevated BMI is associated with increased risk of developing [Type 2 Diabetes] complications."
- Diabetes Pedigree Function: Diabetes Pedigree Functions calculate the likelihood of developing diabetes based on a patient's age and family history. Because some types of diabetes are genetic, it is safe to assume DPF is a good predictor of diabetes.
- Age: Typically for type 2 diabetes, it is most often developed in people over age 45 (source: Type 2 Diabetes | CDC). Thus, since there is an increased risk of diabetes as one grows older, we can assume that age could also be a good predictor of diabetes.

Based on Figure 2 above, we see that the three best predictors of diabetes involve the Diabetes Pedigree Function, pregnancies, and BMI. Thus, the model tells us that those with higher Diabetes Pedigree Functions, with more pregnancies, and higher BMIs tend to have higher $p(diabetes)$ then those with lower instances of those features. However, this is the initial interpretation, and to better understand/interpret the coefficients, we need to draw attention to the p-values of each coefficient.

For the sake of this analysis, we compare with $\alpha = 0.05$. Interestingly enough, we get the following interpretations:

- p-values for pregnancies, glucose, Blood Pressure, BMI and the Diabetes Pedigree Function are all significant. This tells us that these predictors are associated with $p(diabetes)$.
- p-values for Skin Thickness, Insulin, and Age are large. Thus, these variables appear to not be associated with $p(diabetes)$.

This is notable, as looking at Figure 3, we see that there exist a range of low to moderate correlations between certain features, including those that are considered to not be significant and those who are considered to be significant.

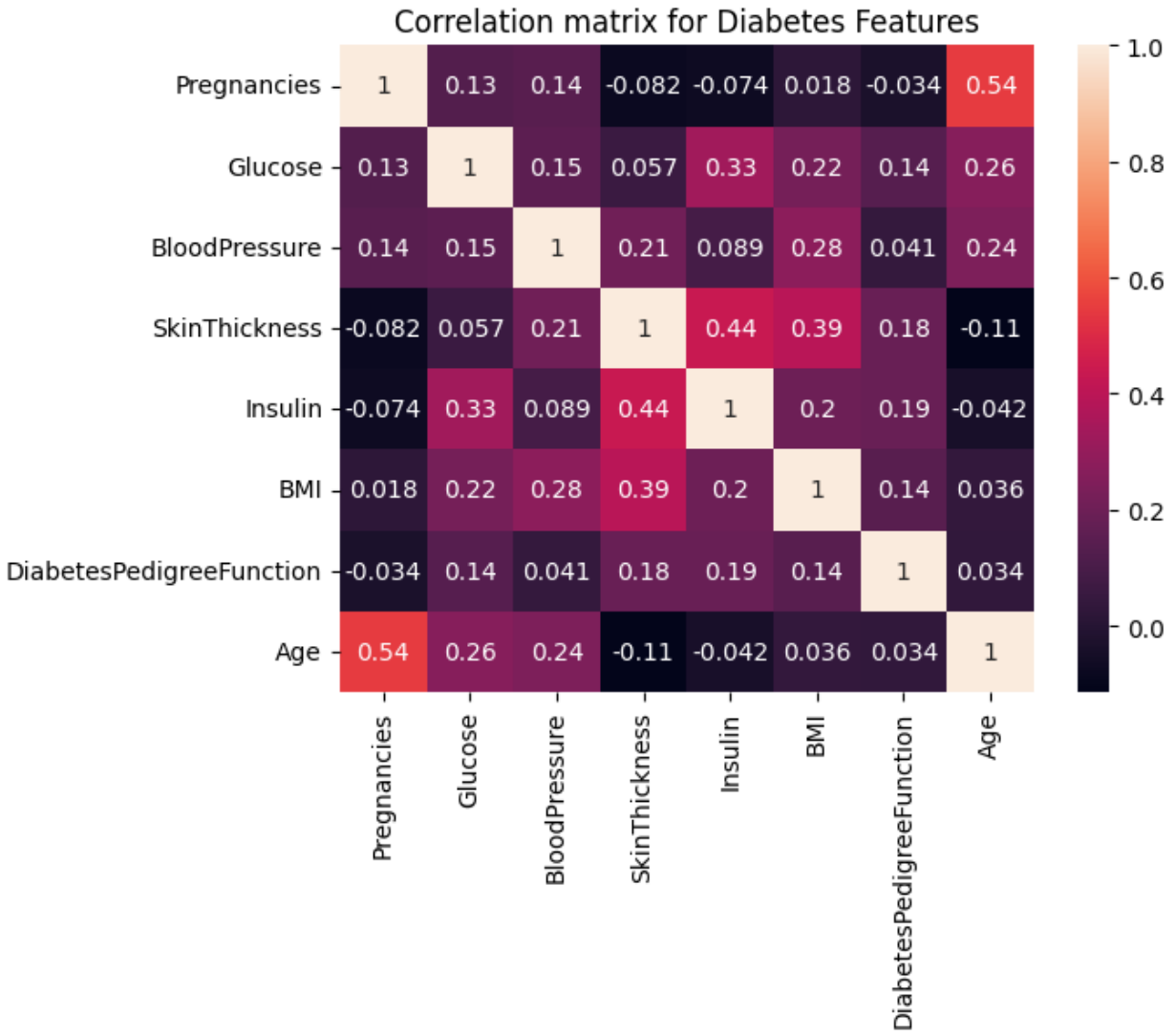

Figure 3: Correlation Matrix for Diabetes Features

For example, the following pairs show a range of low to moderate correlation:

- age and pregnancy
- skin thickness and insulin
- glucose and insulin
- blood pressure and BMI
- skin thickness and BMI

In general, these correlations are in line with current understanding of diabetes and nature, as those who are older are more likely to experience pregnancies, high blood sugar levels often result in digital sclerosis, and those who have higher BMIs are more likely to experience hypertension. What draws attention the most is how pregnancy is considered to be a variable associated with diabetes, but age not considered to be a variable despite the moderate correlation between the variables.

Thinking over this again reinforces our understanding of the results of the graph, as while age and pregnancy may be correlated, as one cannot become pregnant unless they experience puberty first, this is reliant on the fact that the individual is capable of pregnancy first, and thus, correlation between these two variables does not imply that both are significant predictors of $p(diabetes)$.

In short, the coefficients in Figure 2 tell us that:

- those with Higher Diabetes Pedigree Function scores tend to have higher $p(diabetes)$ then those with low DPF scores (with DPF scores being the strongest predictor of whether or not an individual develops diabetes).
- Those with higher Glucose Levels tend to have higher $p(diabetes)$.
- Those with higher Blood Pressure are less likely to have $p(diabetes)$ then those with low BP.
- Those with higher Glucose Levels tend to have higher $p(diabetes)$.
- Those who experience more pregnancies tend to have higher $p(diabetes)$.