Prince Genel R. Umali

BSCS–3 B

I created a diabetes prediction model that uses a logistic regression model and a random forest classifier to predict diabetes. In preparing the dataset before training the model is that I see what the data looks like, and I saw that some columns, like insulin, have a value of 0, which I think can cause poor performance for my model, so I make them nan. In my dataset, there are no missing values, so I moved on to splitting the dataset into two parts: the training set and a test set. After that, I used a pipeline so that my code is clean and avoids data leaking because inside my pipeline, I used 3 functions. First is using the simpleImputer, which replaces the nan values depending on the median of the dataset. Next is scaling the dataset by using standardScaler because some columns have bigger numbers and some have smaller numbers. With the use of a standard scaler, it balances the numbers so that they a similar scale, because some models are sensitive to the feature scales. And lastly is the model, which is logistic regression, the same as the random forest classifier.

The test accuracy for the logistic regression and random forest classifier is the same, that has 0.8052. A confusion matrix helps us see how well our model is doing in predicting who has diabetes and who doesn't. We can see how many people the model correctly predicted as having diabetes, which is called True positives, or not having diabetes, which is called True negatives. And also it shows how many it got wrong predictions, which are False positives and False negatives. The logistic regression model predicted correctly 80 people without diabetes and 44 with diabetes, but had 19 false positives and 11 false negatives. The random forest classifier model identified 78 people without diabetes and 46 with diabetes, with 21 false negatives and 9 false negatives. Both models made mistakes; logistic regression made more mistakes in false negatives, while the random forest classifier made more mistakes in false positives.

Cross-validation helps us understand if our model performs consistently. We split the data into 5 parts and test the model on each part. The average accuracy across these tests tells us the performance and the standard deviation of the model. The logistic regression model has an average accuracy of 75.38% with a standard deviation of 0.0437, showing consistent performance. The random forest classifier model has a slightly higher average accuracy of 76.04% and a standard deviation of 0.0497, also showing reasonable performance.

The learning curves use a graphical representation that helps us to determine if the model is overfitting, underfitting, or a good fit. The logistic regression learning curve shows that the training and validation lines are close to each other, which means the model is a good fit. The random forest classifier model shows a gap between training and validation accuracy, with training being higher, which indicates the model a overfitting. This supports that logistic regression might be a better fit for this data.

To improve the model, we could try creating more features from the dataset, like combining age and BMI. We can also improve the model's hyperparameters to find the best combination for better performance. Also, trying different models to see

which model has a better performance on the dataset. The dataset would be better if it had more data and balanced the data that has diabetes and without diabetes, so that it does not have an imbalanced dataset, which can prevent the model from being biased. It is also trying different techniques to fill the zero values in the dataset. With this improvement, it can help to have a more accurate and reliable diabetes prediction model.