CA05A – Logistic Regression

Elisabeth Webb

Steps:

First, I checked for multicollinearity of the variables using a correlation matrix. I removed the variable 'hip' as it had a high correlation (above 0.7) to waist. I experimented with removing other variables such as those with a correlation score above 0.6 (waist circumference, av_weight_kg, and neck20), but I found that the accuracy score didn't improve. The model performed best when only removing hip. Next, I converted the categorical variables to dummies. Once I identified the X and y variables for the model, I split the dataset into train and test, a 70-30 ratio split. The dependent variable for this model is the risk of Cardiovascular Disease, using the independent variables age, race, education, marital status, waist circumference, neck circumference, average weight, smoking in packs per year, cups of tea containing caffeine per day, self-reported hypertension, diabetes status, health limits to bending over, kneeling, or stooping, happiness, tiredness, and how much your health limited social activities. Next, I built the logistic regression model and fit it to the training dataset. Then, I identified the features that were most important in determining if the person is at risk for cardiovascular disease and found that waist circumference has the most influence followed by health limiting social activities level 4/5, and marital status 1 (being married).

When using the model for prediction, I found that the model predicted the risk of cardiovascular disease on the test dataset with about 69% accuracy. Lastly, I evaluated the model's performance using accuracy score, precision, recall, f1 score, AUC value, and the ROC graph. The results show that precision and recall are balanced as the f1 score is high at 0.78. The model has a high recall, which means that of all the people that are actually at risk for CVD, the model correctly predicted 87% of them. For precision, out of the people that we predicted as at risk for CVD, 69% of those predictions are right. Accuracy is also high at 69%. The AUC value from the predicted scores is 0.64, which is not particularly high as 0.5 means the classification is almost random 50-50. When plotting the ROC from X_test and y_test, it shows that the AUC is 0.71. These scores indicate that the model is good and acceptable, as it does a better job than random, but it is not an excellent model in its ability to classify patient's CVD risk based on these independent variables.

```
              precision    recall  f1-score   support

           0       0.70      0.39      0.50       388
           1       0.69      0.89      0.78       585

    accuracy                           0.69       973
   macro avg       0.69      0.64      0.64       973
weighted avg       0.69      0.69      0.66       973

Accuracy: 0.69
Precision: 0.69
Recall: 0.89
F1 Score: 0.78
AUC Value: 0.64
<sklearn.metrics._plot.roc_curve.RocCurveDisplay at 0x7f1a4836d080>
```