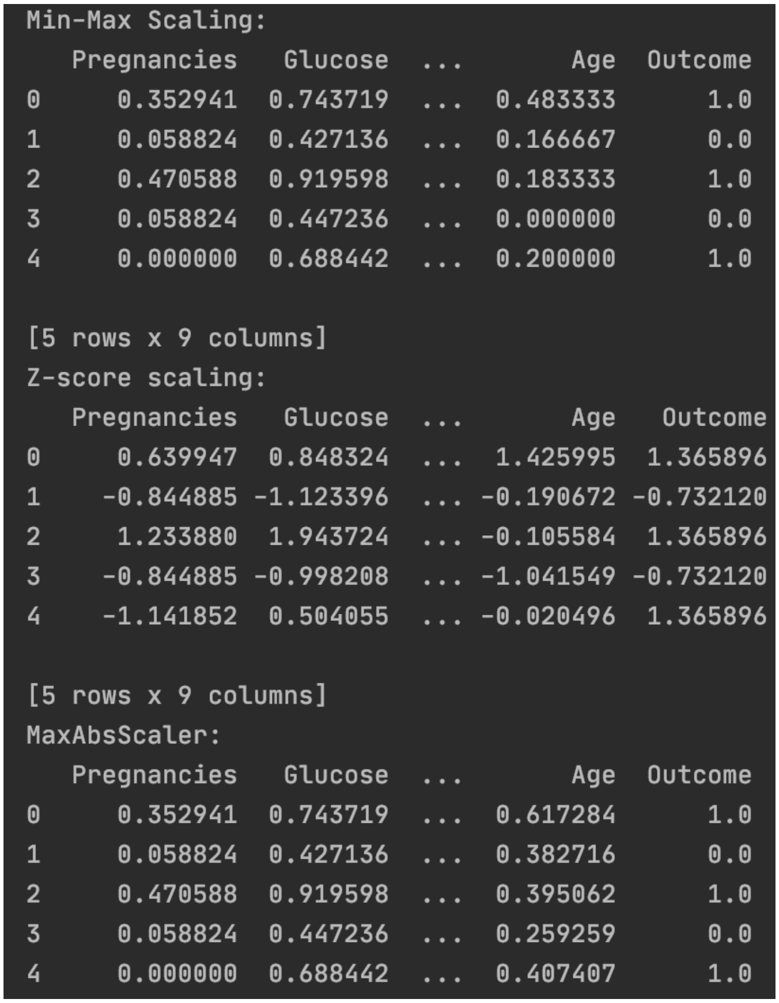
Keilani Bailey

CIS 335

Sorio Boit

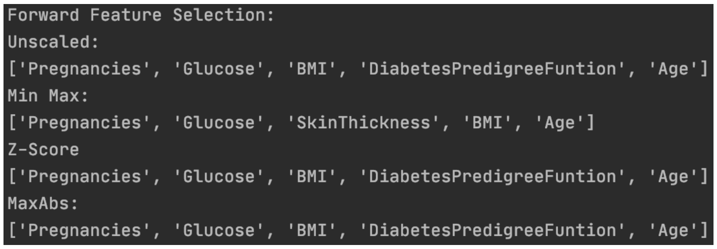
March 1, 2022

1. I used Z-score normalization, MaxAbs scaling and Min-Max scaling. Both Min-Max and MaxAbs scaling resulted in the minimum value becoming 0 and the maximum value becoming 1. This is different from z-score normalization, which resulted in values that were negative and larger than 1.



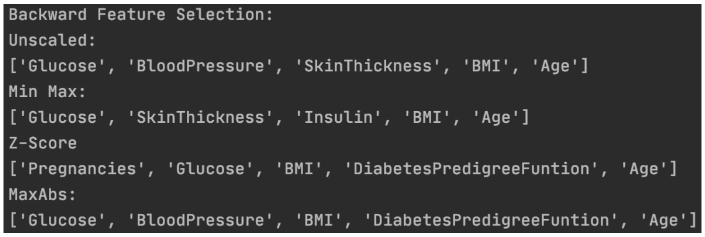
```
Min-Max Scaling:
   Pregnancies   Glucose  ...       Age  Outcome
0     0.352941  0.743719  ...  0.483333      1.0
1     0.058824  0.427136  ...  0.166667      0.0
2     0.470588  0.919598  ...  0.183333      1.0
3     0.058824  0.447236  ...  0.000000      0.0
4     0.000000  0.688442  ...  0.200000      1.0

[5 rows x 9 columns]
Z-score scaling:
   Pregnancies   Glucose  ...       Age   Outcome
0     0.639947  0.848324  ...  1.425995  1.365896
1    -0.844885 -1.123396  ... -0.190672 -0.732120
2     1.233880  1.943724  ... -0.105584  1.365896
3    -0.844885 -0.998208  ... -1.041549 -0.732120
4    -1.141852  0.504055  ... -0.020496  1.365896

[5 rows x 9 columns]
MaxAbsScaler:
   Pregnancies   Glucose  ...       Age  Outcome
0     0.352941  0.743719  ...  0.617284      1.0
1     0.058824  0.427136  ...  0.382716      0.0
2     0.470588  0.919598  ...  0.395062      1.0
3     0.058824  0.447236  ...  0.259259      0.0
4     0.000000  0.688442  ...  0.407407      1.0
```

2. I used forward feature selection, backward feature selection, and RFE. PCA is also a feature reduction method that I used as well. Forward feature selection starts with an empty set and adds variables:



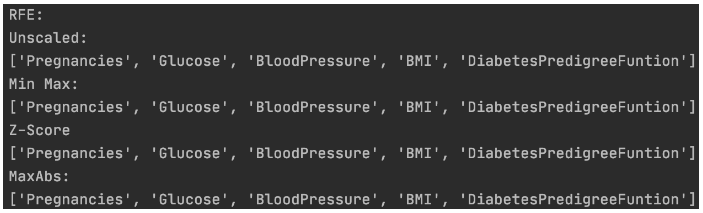
```
Forward Feature Selection:
Unscaled:
['Pregnancies', 'Glucose', 'BMI', 'DiabetesPredigreeFuntion', 'Age']
Min Max:
['Pregnancies', 'Glucose', 'SkinThickness', 'BMI', 'Age']
Z-Score
['Pregnancies', 'Glucose', 'BMI', 'DiabetesPredigreeFuntion', 'Age']
MaxAbs:
['Pregnancies', 'Glucose', 'BMI', 'DiabetesPredigreeFuntion', 'Age']
```

Backward feature selection starts with a complete set and removes variables:



```
Backward Feature Selection:
Unscaled:
['Glucose', 'BloodPressure', 'SkinThickness', 'BMI', 'Age']
Min Max:
['Glucose', 'SkinThickness', 'Insulin', 'BMI', 'Age']
Z-Score
['Pregnancies', 'Glucose', 'BMI', 'DiabetesPredigreeFuntion', 'Age']
MaxAbs:
['Glucose', 'BloodPressure', 'BMI', 'DiabetesPredigreeFuntion', 'Age']
```

RFE feature selection fits a model and removes the weakest features:



```
RFE:
Unscaled:
['Pregnancies', 'Glucose', 'BloodPressure', 'BMI', 'DiabetesPredigreeFuntion']
Min Max:
['Pregnancies', 'Glucose', 'BloodPressure', 'BMI', 'DiabetesPredigreeFuntion']
Z-Score
['Pregnancies', 'Glucose', 'BloodPressure', 'BMI', 'DiabetesPredigreeFuntion']
MaxAbs:
['Pregnancies', 'Glucose', 'BloodPressure', 'BMI', 'DiabetesPredigreeFuntion']
```

The results are only consistent using RFE feature selection. Forward feature selection and backward feature selection are both inconsistent.

3. Without the scaling methods the results for step 2 change. For every method except for RFE, the selected features differ. For forward feature selection, the unscaled data matches the features selected for the z-scored and MaxAbs normalized values. For backward feature selection, the features selected for the unscaled data are similar to the min-max normalized data. There is only one feature that is different.