Patrick Callahan, Arjun Chimni, Andrew Schulz

CSCI 184: Machine Learning

4.15.22

Report: Feature Selection

* To begin the project, we downloaded the dataset from the website applied the header names that were necessary for the dataset.
  + Graphical user interface, text

    Description automatically generated
  + Table

    Description automatically generated
  + Table

    Description automatically generated
* After that we separated the data and cleaned it up a bit so that we could properly use it for Filter-based Feature Extraction.
* Because the input values were numerical and the output values were also numerical (0 and 1), we used the f\_regression Filter-based method.
  + Code and output:
  + A picture containing application

    Description automatically generated
  + Graphical user interface, text

    Description automatically generated with medium confidence

|  |  |  |
| --- | --- | --- |
| Method | Value of K | Feature Names |
| Filter Based | K = 3 | * + Glucose   + BMI   + Age |
|  | K = 4 | * + Glucose   + BMI   + Age   + Pregnancies |
|  | K = 5 | * + Glucose   + BMI   + Age   + Pregnancies   + DiabetesPedegreeFunction |

* After the features had been extracted, we were ready to build the Logistic Regression Models using sklearn. The following is our output for and images of the models.
  + Graphical user interface, text, application

    Description automatically generated
  + Graphical user interface, text, application, email

    Description automatically generated
  + A picture containing text

    Description automatically generated

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Case | Feature Names | Precision | Recall | F1-Score |
| 1 | BMI, Age | 0.6363636 | 0.33333 | 0.437499 |
| BMI, Glucose | 0.733333 | 0.540983 | 0.62264 |
| DiabetesPedigreeFunction, Glucose | 0.555555 | 0.390625 | 0.458715 |
| 2 | Insulin, SkinThickness | 1.0 | 0.016393 | 0.03225 |