CS746 Fall 2024 Big HW3 Feature Selection and PCA

Using the SAT dataset ('CASE1201.ASC.txt')

Create a .ipynb notebook to submit your assignment. The file name should start with “BHW3\_yourwsuid”.

1. Display the correlation and covariance tables for the raw features.
2. Build a: (a) simple (intercept + takers) and (b) full linear regression model to predict the **SAT scores**. (that is, simply copy the first two models from the SAT\_example.ipynb presented in class.)
3. Choose at least one method from each of the three feature selection techniques (filter, wrapper, and embedded). Select and report the “best” features using these three methods.
4. Did you expect to see any differences between the “best features” in the different models you chose? Do you see any significant differences? (answer this using a markdown cell in the .ipynb file).
5. Calculate the Principal Components for the same data set: (a) Normalize the data, (b) display correlation and covariance tables fort the normalized data, build the PCA following the in-class model, and finally display the explained variance for all the principal components.
6. In your opinion, what is the “best” number of Principal Components to use? (again, answer using a markdown cell).

Submit the .ipynb for your assignment.