**Experiment No.8 Title:** Attribute subset selection

# Batch: A2 Roll No.: 16010421075 Experiment No.: 8 Aim: Attribute subset selection

**Resources needed:** Python

# Theory:

Attribute subset Selection is a technique which is used for data reduction in data mining process. Data reduction reduces the size of data so that it can be used for analysis purposes more efficiently.

# Need of Attribute Subset Selection-

The data set may have a large number of attributes. But some of those attributes can be irrelevant or redundant. The goal of attribute subset selection is to find a minimum set of attributes such that dropping of those irrelevant attributes does not much affect the utility of data and the cost of data analysis could be reduced. Mining on a reduced data set also makes the discovered pattern easier to understand.

# Methods of Attribute Subset Selection-

1. Stepwise Forward Selection.
2. Stepwise Backward Elimination.
3. Combination of Forward Selection and Backward Elimination.
4. Decision Tree Induction.

All the above methods are greedy approaches for attribute subset selection.

* 1. **Stepwise Forward Selection:** This procedure start with an empty set of attributes as the minimal set. The most relevant attributes are chosen(having minimum p-value) and are added to the minimal set. In each iteration, one attribute is added to a reduced set.
  2. **Stepwise Backward Elimination:** Here all the attributes are considered in the initial set of attributes. In each iteration, one attribute is eliminated from the set of attributes whose p-value is higher than significance level.
  3. **Combination of Forward Selection and Backward Elimination:** The stepwise forward selection and backward elimination are combined so as to select the relevant attributes most efficiently. This is the most common technique which is generally used for attribute selection.
  4. **Decision Tree Induction:** This approach uses decision tree for attribute selection. It constructs a flow chart like structure having nodes denoting a test on an attribute. Each

branch corresponds to the outcome of test and leaf nodes is a class prediction. The attribute that is not the part of tree is considered irrelevant and hence discarded.

# Procedure / Approach /Algorithm / Activity Diagram:

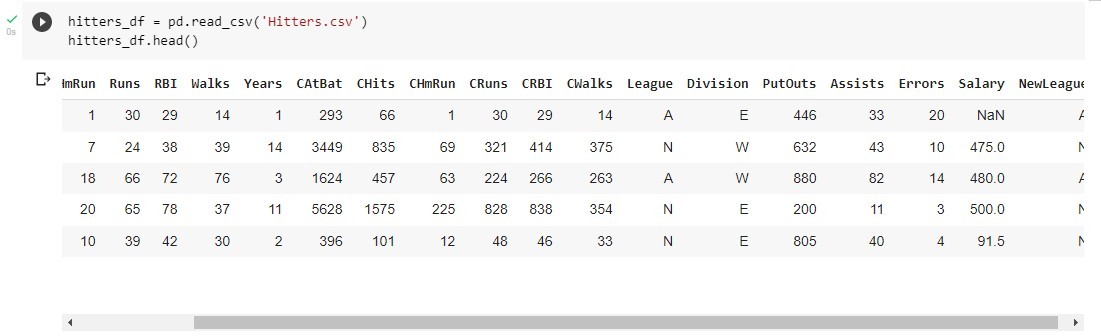
This lab on Subset Selection is a Python adaptation of p. 244-247 of "Introduction to Statistical Learning with Applications in R" by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. Adapted by R. Jordan Crouser at Smith College for SDS293: Machine Learning (Spring 2016).

Technique used: Best Subset Selection

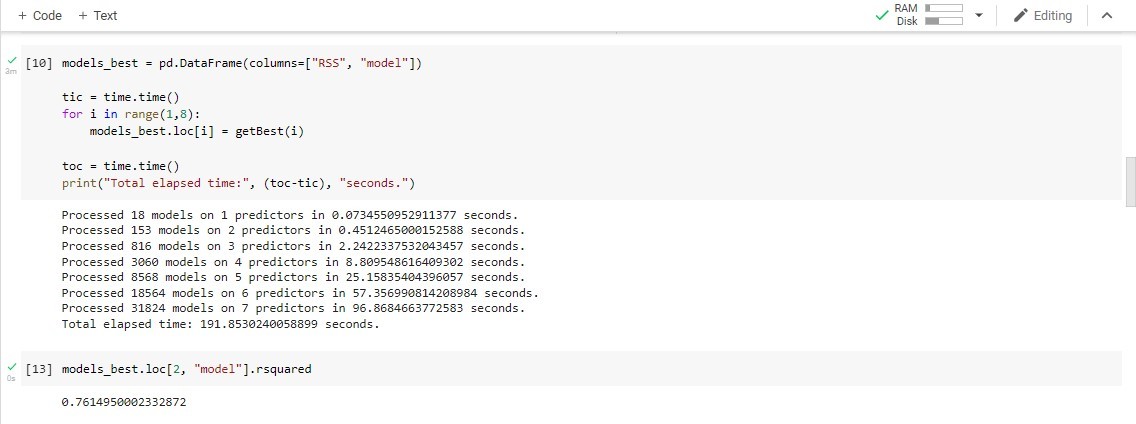
Url: <http://www.science.smith.edu/~jcrouser/SDS293/labs/lab8-py.html>

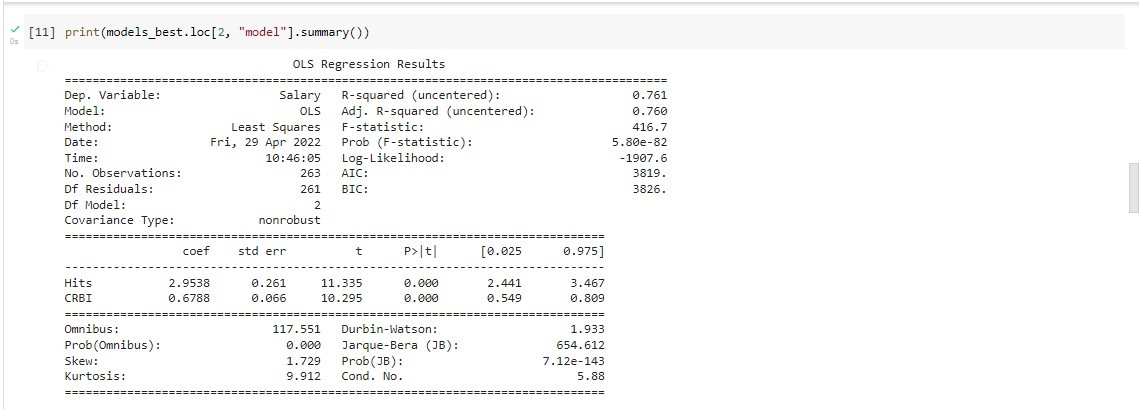
\_

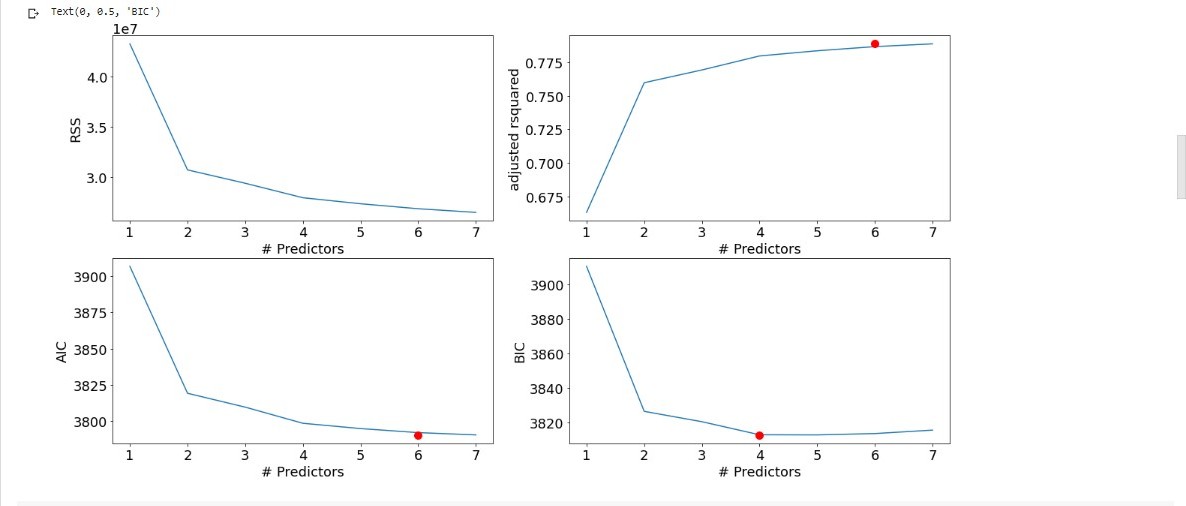
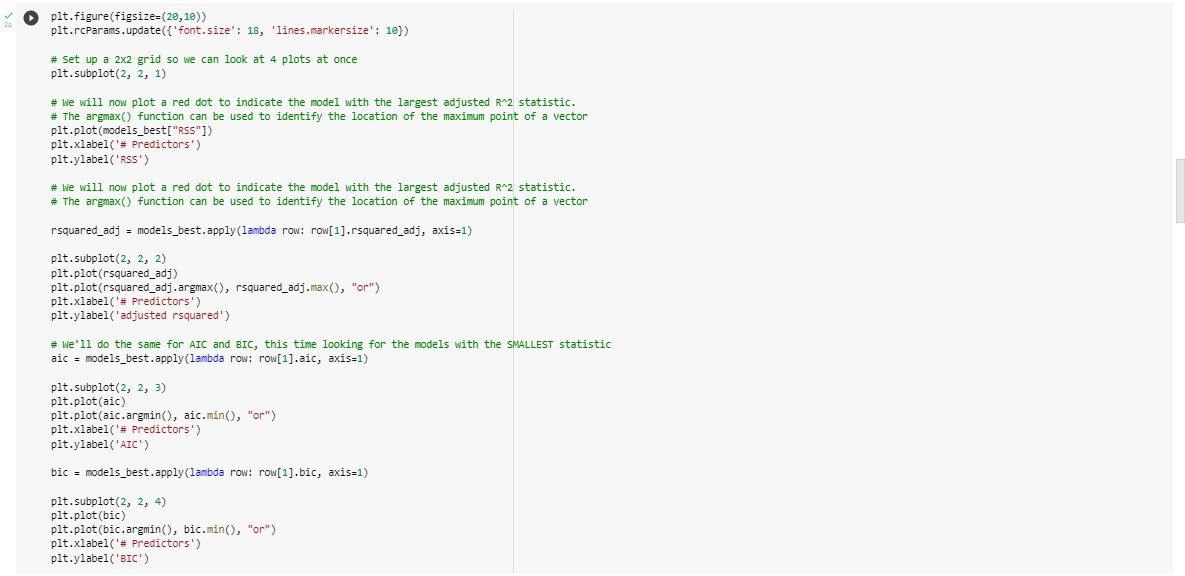
# Results: Students must submit the output of above activity.



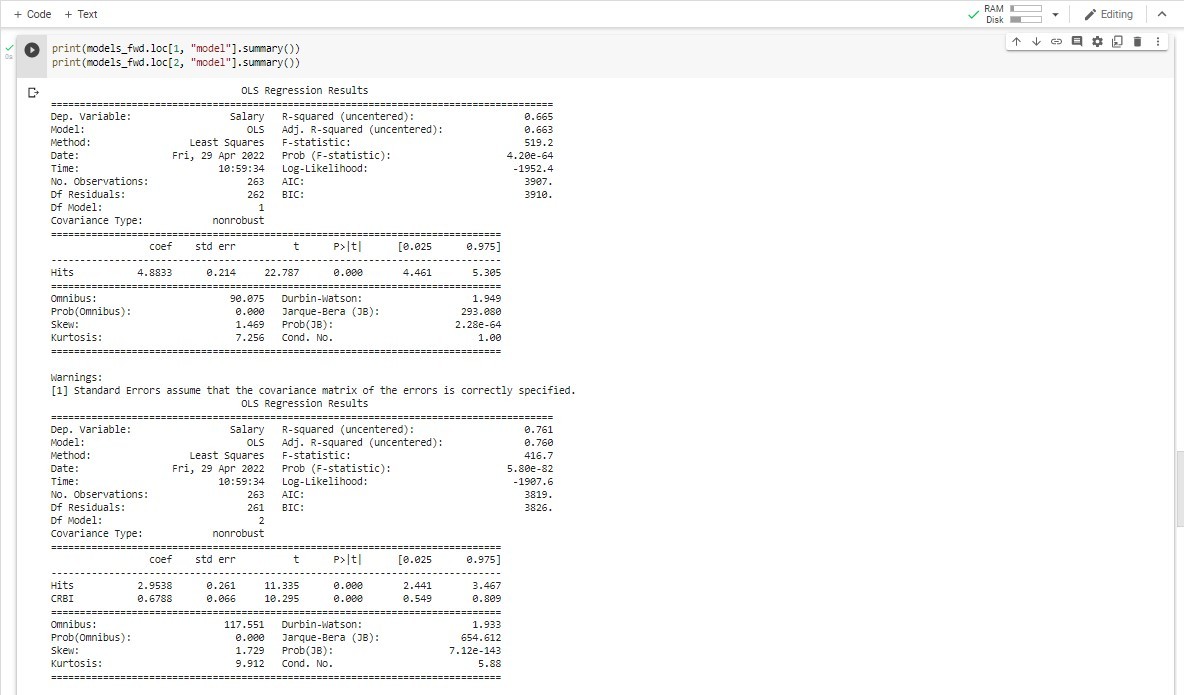
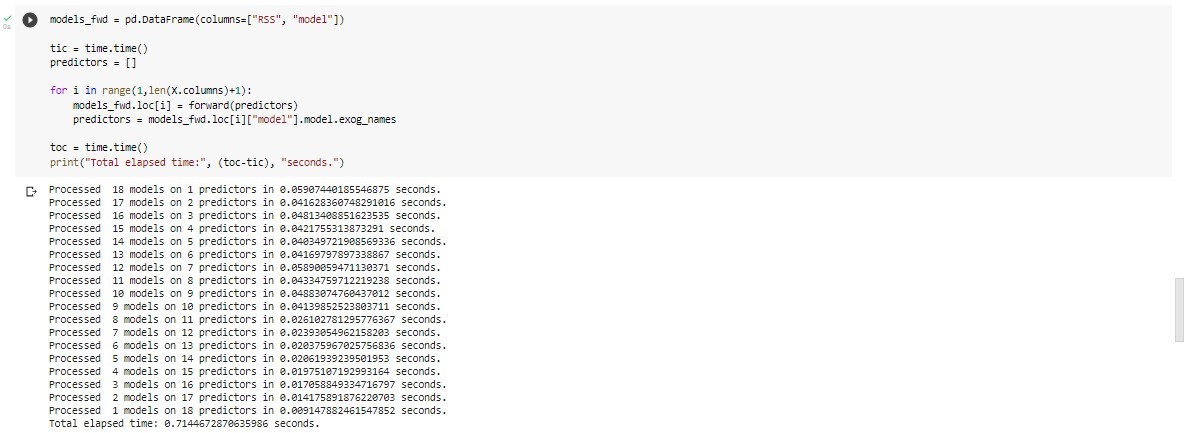


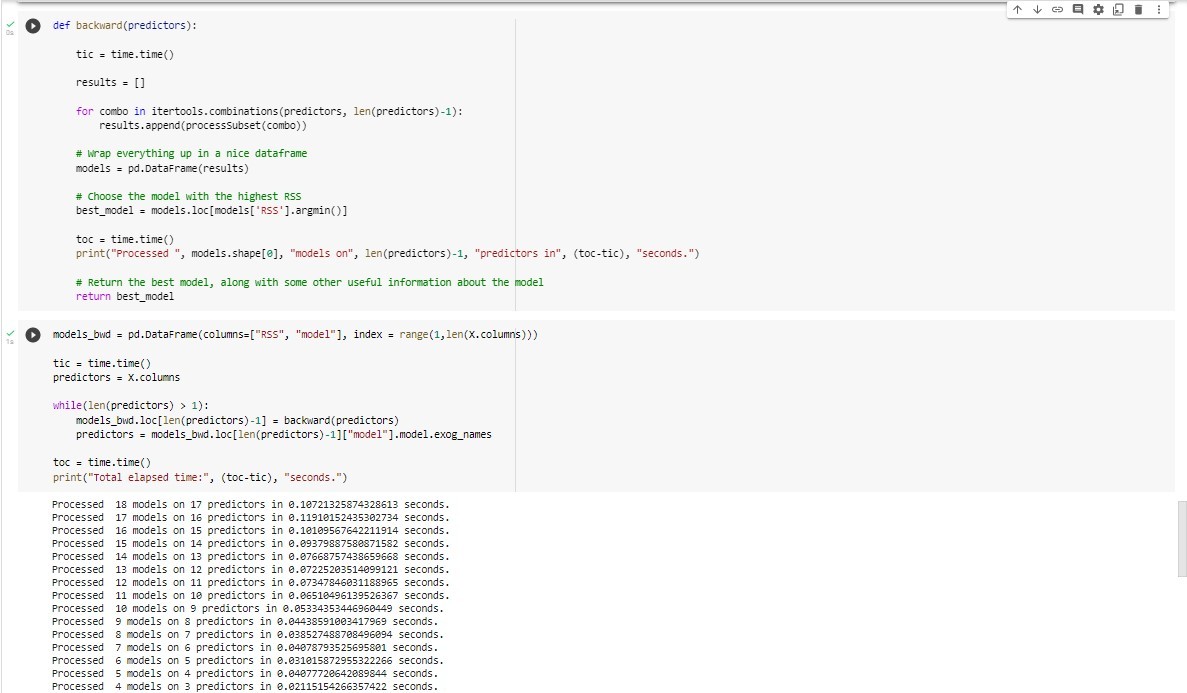


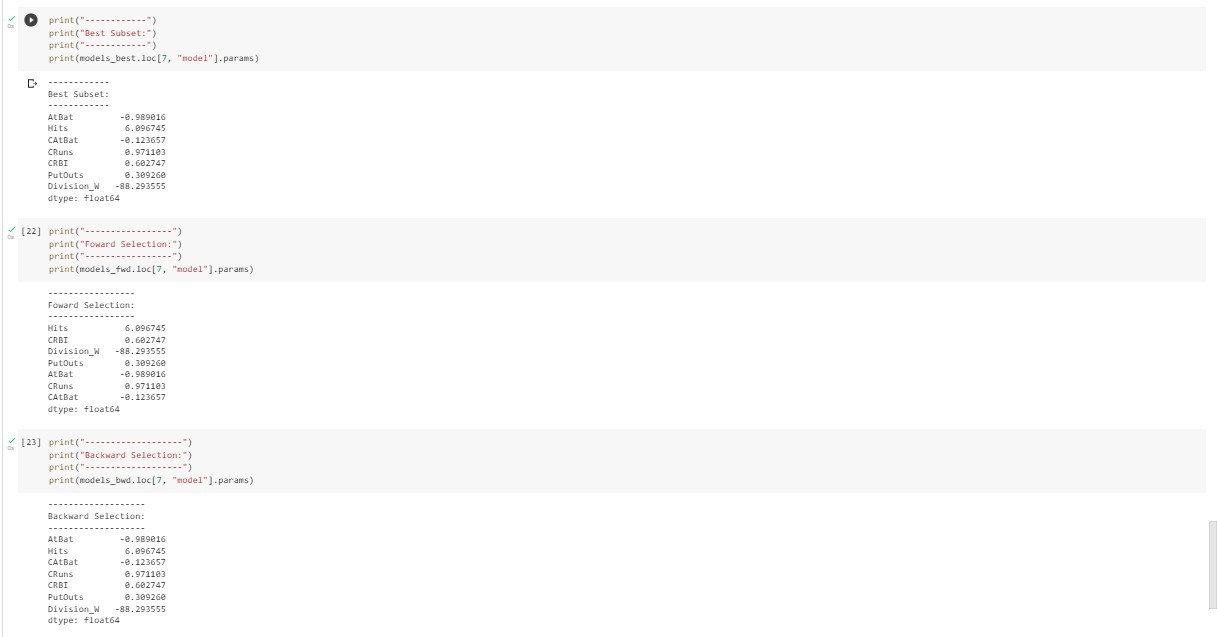












**Questions:**

* + 1. Explain other data reduction techniques in brief.

# Ans.

* + - 1. **Dimensionality Reduction:**

Whenever we encounter weakly important data, we use the attribute required for our analysis. Dimensionality reduction eliminates the attributes from the data set under consideration, thereby reducing the volume of original data. It reduces data size as it eliminates outdated or redundant features. Here are three methods of dimensionality reduction.

* + - 1. **Numerosity Reduction:**

The numerosity reduction reduces the original data volume and represents it in a much smaller form. This technique includes two types parametric and non parametric numerosity reduction.

# Non-Parametric:

A non-parametric numerosity reduction technique does not assume any model. The non-Parametric technique results in a more uniform reduction, irrespective of data size, but it may not achieve a high volume of data reduction like the parametric. There are at least four types of Non-Parametric data reduction techniques, Histogram, Clustering, Sampling, Data Cube Aggregation, and Data Compression.

* **Histogram:** A histogram is a graph that represents frequency distribution which describes how often a value appears in the data. Histogram uses the binning method to represent an attribute's data distribution. It uses a disjoint subset which we call bin or buckets.

A histogram can represent a dense, sparse, uniform, or skewed data. Instead of only one attribute, the histogram can be implemented for multiple attributes. It can effectively represent up to five

attributes.

* **Clustering:** Clustering techniques groups similar objects from the data so that the objects in a cluster are similar to each other, but they are dissimilar to objects in another cluster.The quality of the cluster depends on the diameter of the cluster, i.e., the max distance between any two objects in the cluster.

The cluster representation replaces the original data. This technique is more effective if the present data can be classified into a distinct clustered.

* **Sampling:** One of the methods used for data reduction is sampling, as it can reduce the large data set into a much smaller data sample. Below we will discuss the different methods in which we can sample a large data set D containing N tuples.

# Outcomes:

CO3. Apply the transformations required on data to make it suitable for mining.

# Conclusion: (Conclusion to be based on the objectives and outcomes achieved)

We successfully implemented data reduction of subsets.

# Grade: AA / AB / BB / BC / CC / CD /DD

Signature of faculty in-charge with date

# References:

Books/ Journals/ Websites:

1. Han, Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann 3nd Edition
2. Subset Selection is a Python adaptation of p. 244-247 of "Introduction to Statistical Learning with Applications in R" by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. Adapted by R. Jordan Crouser at Smith College for SDS293: Machine Learning (Spring 2016).
3. Dataset: https://[www.kaggle.com/code/omeryasirkucuk/salary-prediction-models-on-](http://www.kaggle.com/code/omeryasirkucuk/salary-prediction-models-on-) hitters-dataset/data?select=Hitters.csv