**DATA ANALYSIS ON**

**JOB RECRUITMENT OF PEOPLE**

Shwethambari V1, Sruthi R2, Yazhini Karthik3

Dr. M. Srividya, Ms. K.H. Vani, Ms J. Rathika

“2nd year M. Sc, Department of Computing - Data Science,

Coimbatore Institute of Technology, Coimbatore, Tamil Nadu, India”

**ABSTRACT:**

The main objective of this paper is to analyse a dataset on job recruitment of people using statistical concepts. The data frame consists of 12 attributes and a total of 200 rows. Sorting concept is used where graphs using sorted values are used for better visual representation than unsorted graphs. Next decision trees concept is used consisting of entropy method where accuracy is predicted by training the data frame. Naive bayes theorem is also implemented similar to decision tree using gaussian method. The accuracies from decision tree and naïve bayes theorem are compared and inferred to be more or less similar having accuracy rate of 80% approximately. Therefore, the advantage of sorted graphs is analysed and comparison of accuracies from naïve bayes theorem and decision trees are inferred in this paper.

DATASET USED: JOB RECRUITMENT OF PEOPLE DATASET

DATASET SOURCE: KAGGLE

**DATA STRUCTURES AND ALGORITHMS CONCEPTS USED:**

* SORTING
* DECISION TREE

These concepts are also implemented, refered and compared with a few probability concepts.

LIBRARIES USED:

**Pandas**-To import the dataset

**NumPy**- Application of arrays

**Matplotlib.pyplot, %matplotlib.inline** -For visual representation of data using graphs

**Import sklearn**:

**Sklearn.model\_selection importing train\_test\_split**-To select models for train test in Naïve bayes method

**Sklearn.naive\_bayes import gaussianNB**-For application of naïve bayes method

**Sklearn.metrics import accuracy\_score** Inbuilt function to calculate accuracy in a particular method

**Sklearn.tree import DecisionTreeClassifier** For decision tree application

**Scipy.stats**- to work with statistical functions

**SORTING:**

**Sorting refers to ordering data in an increasing or decreasing fashion according to some linear relationship among the data items. This is done so that we can analyze data more effectively.**

Comparing the unsorted data members with the sorted members by graphical representation:

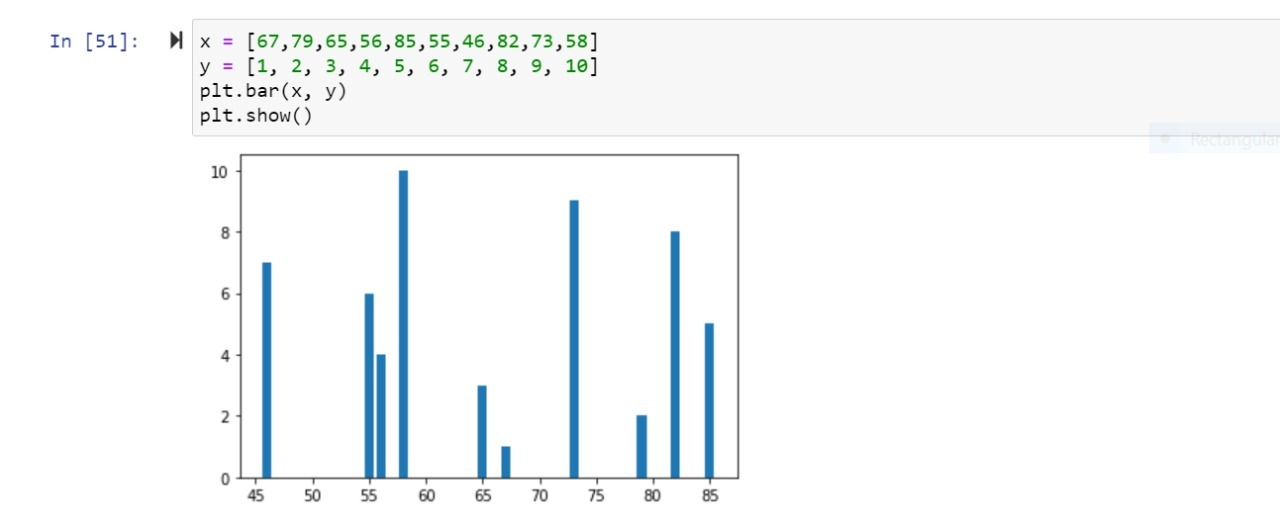
Few unsorted values of secondary score percentage column from the data set are taken and the graph is plotted where x-axis contains the dataset values and y-axis contains the index numbers.

**USE IN CODE:**

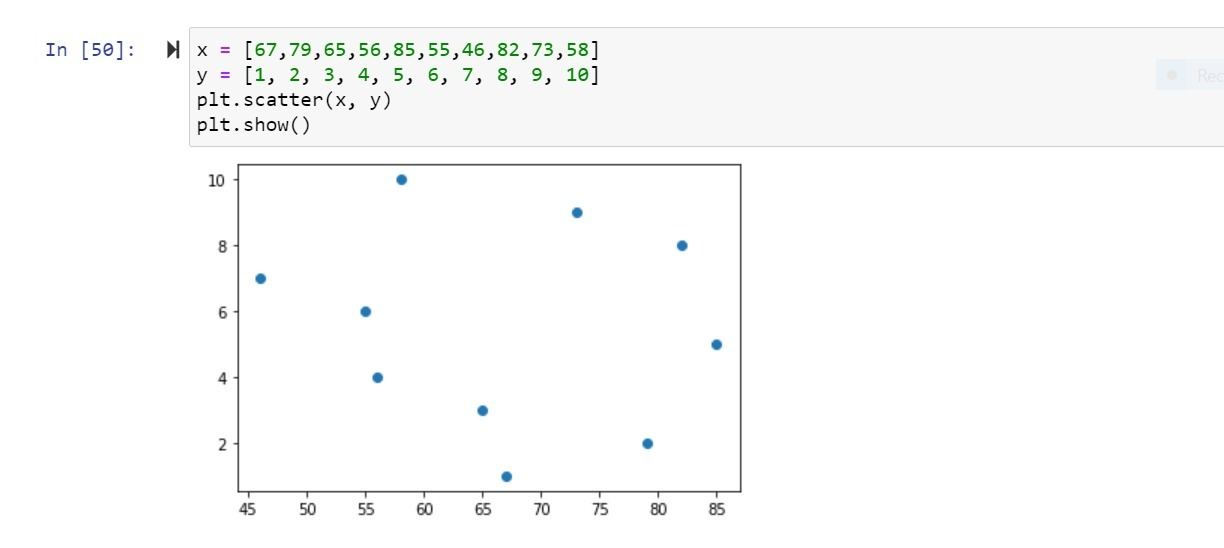
BAR GRAPH:

x-axis: unsorted values of ssc\_% column

y-axis: index numbers



SCATTER PLOT:



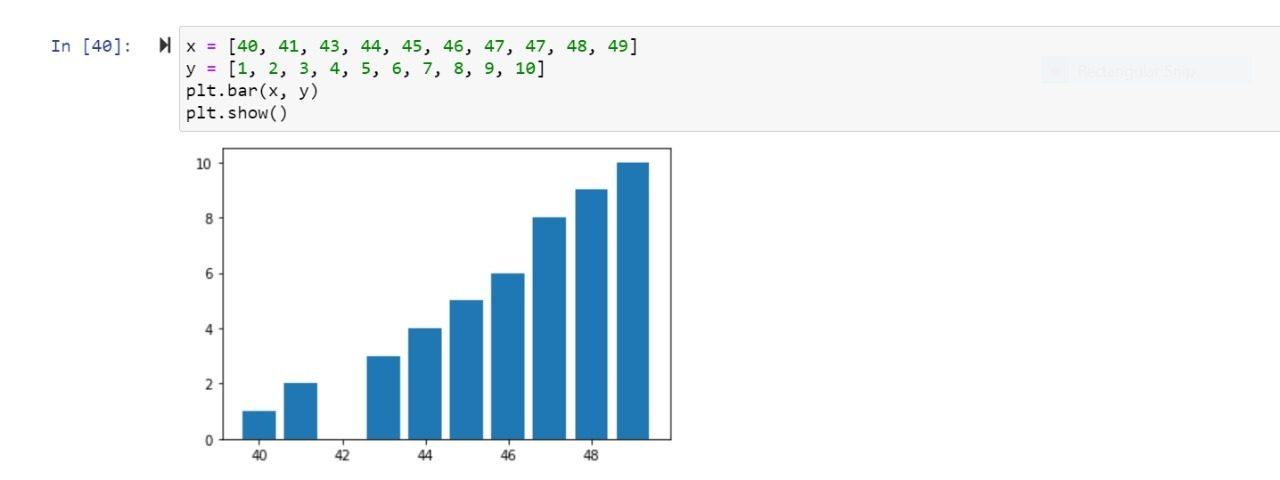
* The previous picture symbolizes the bar and scatter graphs of ssc\_p column without getting sorted. Take a look on the varying proportion of the values in the graph.
* That is because the values are not sorted.

Now, we’ll have a look on the graphs from the sorted values of secondary score percentage column.

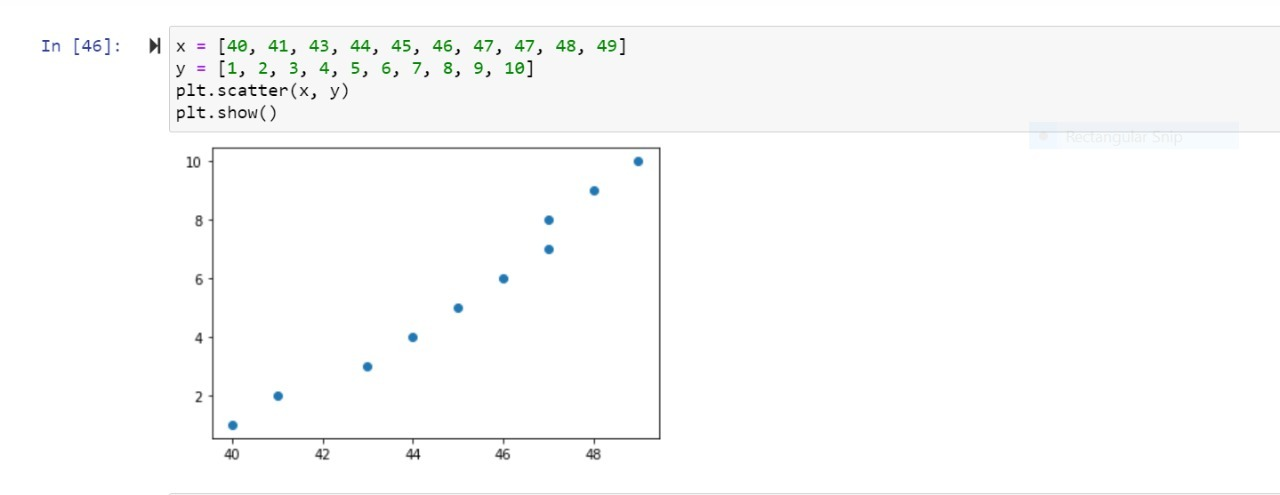
BAR GRAPH:

x-axis: sorted values of ssc\_% column

y-axis: index numbers



SCATTER PLOT:



**INFERENCE:**

We Can also use unsorted graph for interpretation, but the more organized way is approached in a sorted graph. So, when sorting is used, the graph becomes clearer and more understandable. Hence, both the graphs are proper ones to be used.

**DECISION TREE:**

A decision tree is a very specific type of probability tree that enables you to make a decision about some kind of process. For example, you might want to choose between manufacturing item A or item B.

Here we are going to compare the accuracy obtained from normal train test method and the accuracy obtained from decision tree.

**Method 1: NAÏVE BAYES METHOD**

**USE IN CODE:**

In this method we use naïve bayes theorem to get the accuracy.

**It includes five steps:**

1. DATA PRE-PROCESSING:
   1. We’ll create the x and y variables from dataset using train\_split\_test function then split the data into train and test sets (subsets).
   2. There is no need to divide the dataset manually.
   3. For the output of train\_test\_split, we get x\_train,x\_test and y\_train,y\_test
2. TRAIN THE MODEL:
   1. Turning the models to match their requirements. We are training the model using GAUSSIAN method.
3. PREDICTION:
   1. Once the model is trained it is ready to make predictions. We can use the predict method on the model and pass x\_test as a parameter to get the output as y\_pred.
   2. Prediction output is an array of real numbers corresponding to the input array.
4. MODEL EVALUATION:
   1. We evaluate our model by finding the accuracy.
   2. score produced by the model.
5. PERCENTAGE:
   1. We’ll find the percentage value of the accuracy
   2. produced by the model.

**METHOD:2 DECISION TREE CLASSIFIER METHOD:**

**USE IN CODE:**

**Step 1:** Similar to train and test sets, we use train and test features in decision tree classifier using decision tree classifier function using ENTROPY concept.

**STEP 2:** Prediction and finding accuracy are done with the code related to decision tree (I.e.) tree. Predict and tree.score.

**INFERENCE:**

Both the methods, decision tress and Naïve Bayes theorem are found to have an accuracy rate of 80% approximately.

Naïve bayes classifier is the Normal method we can use for prediction. If we are in need of data structure concept, we can use decision tree classifier method for accuracy calculation.

**CONCLUSION:**

1. Sorting method is used to plot graphs of sorted column values which are visually clear and more understandable than graphs plotted from unsorted column values.
2. Comparison of accuracies are proceeded by two methods- Naïve bayes and Decision tree where, it is inferred that both the concepts are equally similar in providing the accuracy values.

Hence, The end of this paper.

**REFERENCES:**

<https://machinelearningmastery.com/implement-decision-tree-algorithm-scratch-python/>

<https://www.geeksforgeeks.org/decision-tree-implementation-python/>

<https://towardsdatascience.com/simple-way-to-find-a-suitable-algorithm-for-your-data-in-scikit-learn-python-9a9710c7c0fe>

<https://realpython.com/pandas-sort-python/>

<https://www.kaggle.com/benroshan/factors-affecting-campus-placement>