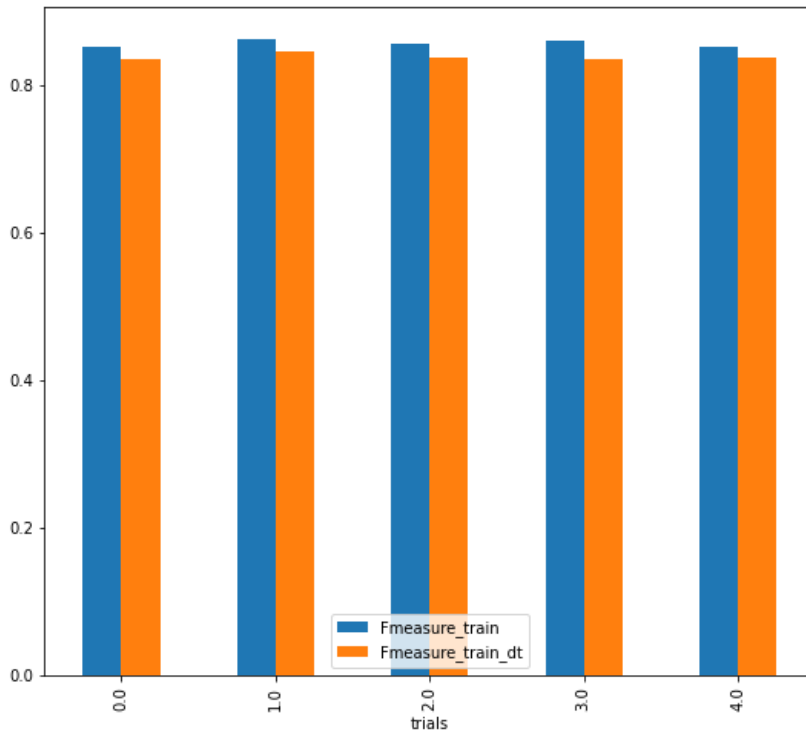


Part 1

Dataset: [https://archive.ics.uci.edu/ml/datasets/statlog+\(german+credit+data\)](https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)) (Links to an external site.)

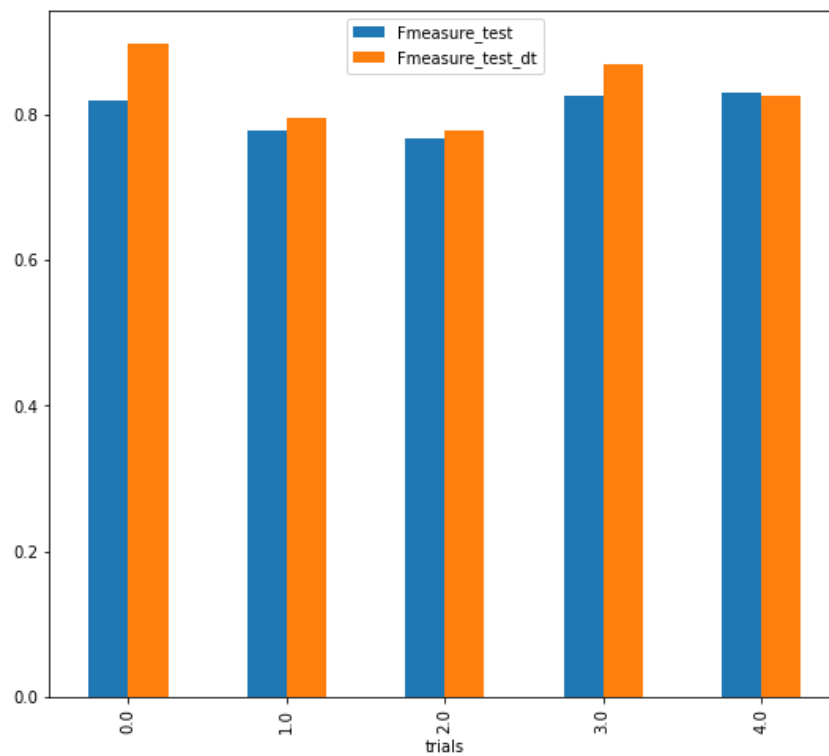
```
fullreport.plot(x='trials', y=['Fmeasure_train', 'Fmeasure_train_dt'], kind="bar", figsize=(9, 8))  
]: <AxesSubplot:xlabel='trials'>
```



Interpretations: The Fmeasure of the KK classifier is higher than Fmeasure of the decision tree on the training data. The KK classifier is made on the training data. That's why the F-measure for both the classifier is quite high.

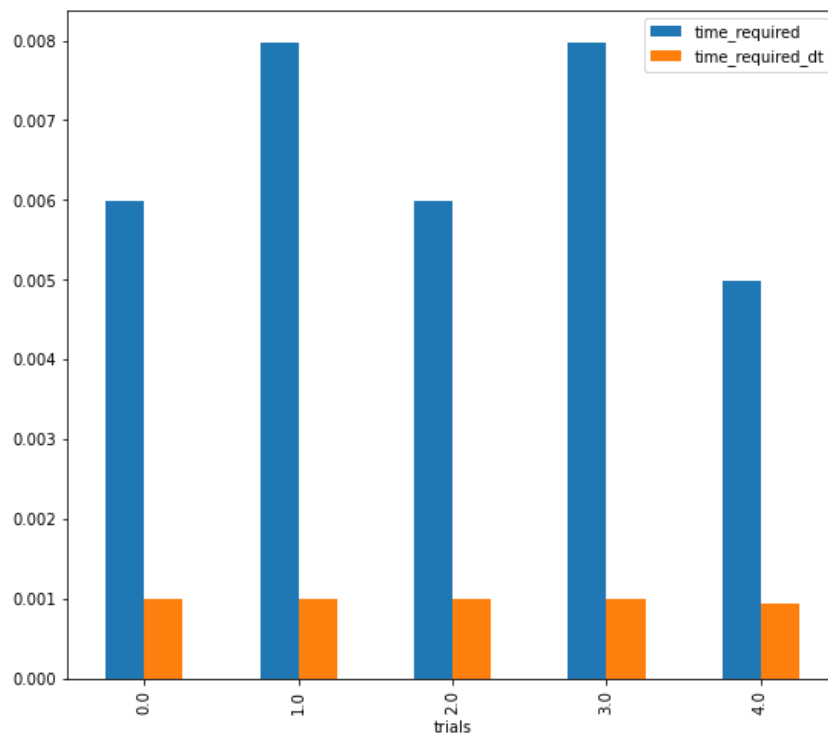
```
fullreport.plot(x='trials', y=['Fmeasure_test', 'Fmeasure_test_dt'], kind="bar", figsize=(9, 8))
```

```
43]: <AxesSubplot:xlabel='trials'>
```



Here we can see that the Fmeasure for the decision tree is higher than the kk classifier for most of the cases. These trials are done on the test data set.

```
4]: fullreport.plot(x='trials', y=['time_required', 'time_required_dt'], kind="bar", figsize=(9, 8))
it[44]: <AxesSubplot:xlabel='trials'>
```



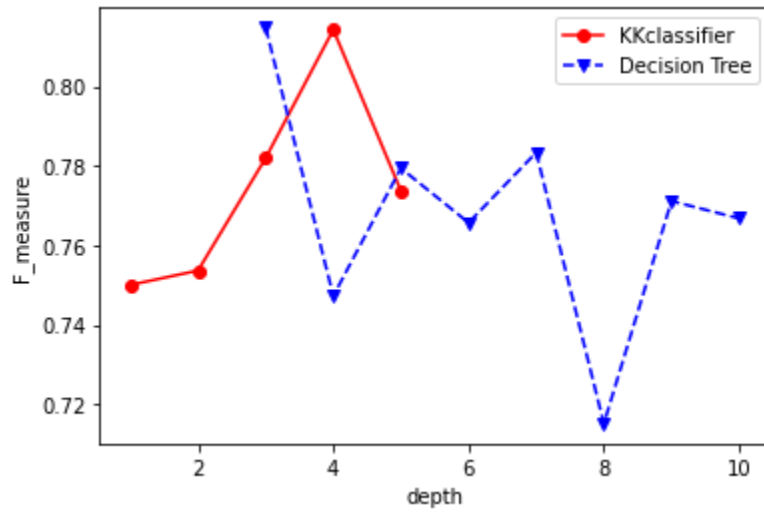
The time required for KK classifier is significantly higher than the time required for decision tree.

```
import statistics
mean_train_f=statistics.mean(fullreport["Fmeasure_train"])
mean_test_f=statistics.mean(fullreport["Fmeasure_test"])
mean_train_f_dt=statistics.mean(fullreport["Fmeasure_train_dt"])
mean_test_f_dt=statistics.mean(fullreport["Fmeasure_test_dt"])
mean_time=statistics.mean(fullreport["time_required"])
mean_time_dt=statistics.mean(fullreport["time_required_dt"])
mean_test_f,mean_train_f,mean_test_f_dt,mean_train_f_dt,mean_time,mean_time_dt
```

```
!9]: (0.804086770954206,
      0.8572113347295547,
      0.8335291792307666,
      0.838583653762321,
      0.0065821647644042965,
      0.0009845733642578126)
```

The mean test f measure is lower than the mean f measure for train data. The mean f measure for test for decisiontree is more than mean test f measure for kk classifier. The mean time required for kk classifier is more than mean time required for decision tree as we saw in the bar chart.

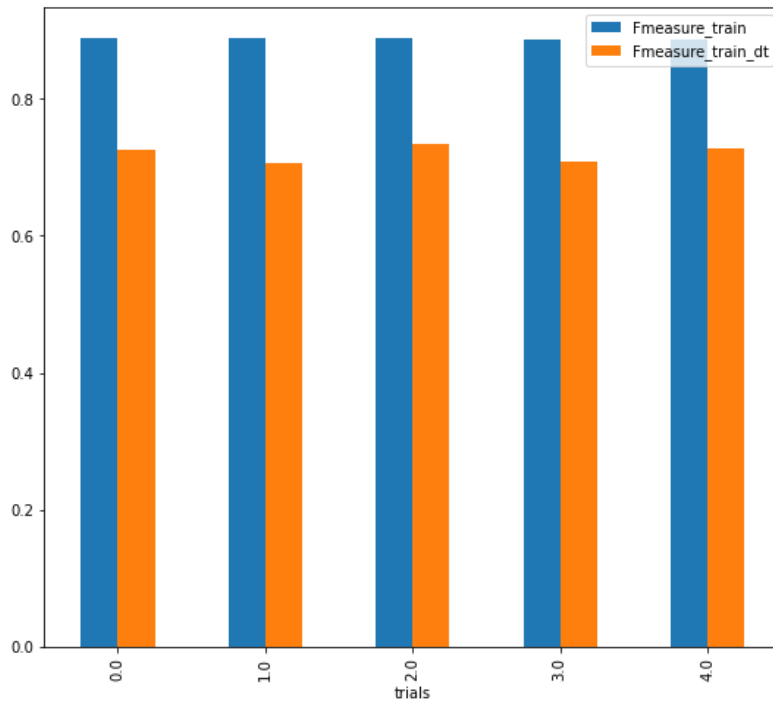
Part 2:



We can see that the KK classifier has the highest f_measure when the k value is 4. That's the suitable value for KK classifier. And for the decision tree the highest f_measure is when the depth is 3. The f_measure is lowest for decision tree when the depth is 8. When the k=1 the F_measure is lowest for kk classifier. This calculation is being done in the validation set.

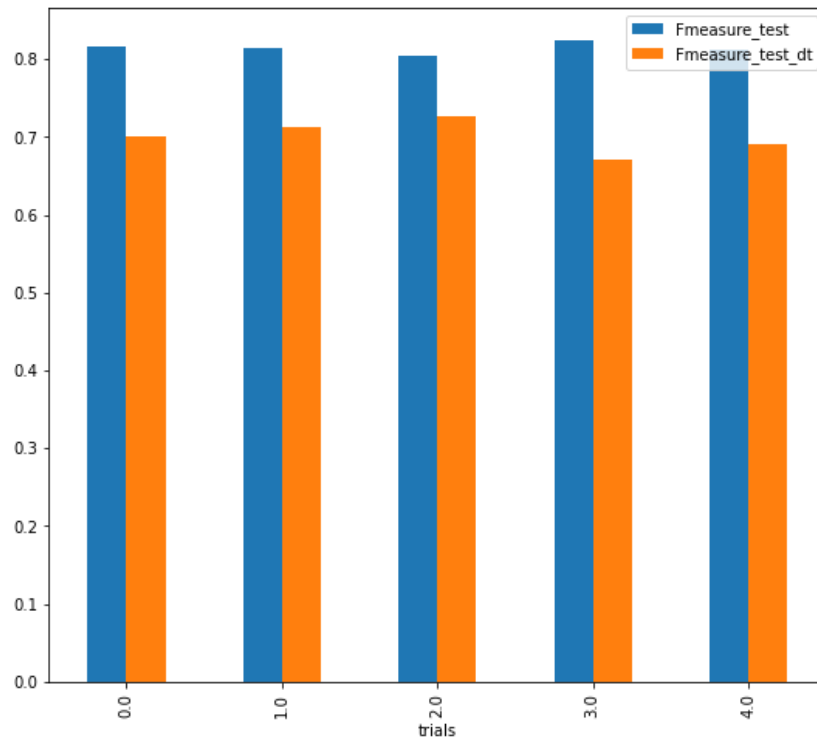
Dataset: [https://archive.ics.uci.edu/ml/datasets/Waveform+Database+Generator+\(Version+1\)](https://archive.ics.uci.edu/ml/datasets/Waveform+Database+Generator+(Version+1))

```
fullreport2.plot(x='trials', y=['Fmeasure_train', 'Fmeasure_train_dt'], kind="bar", figsize=(9, 8))  
]: <AxesSubplot:xlabel='trials'>
```



The F_measure for KK classifier is quite larger than that of decision tree for this dataset.

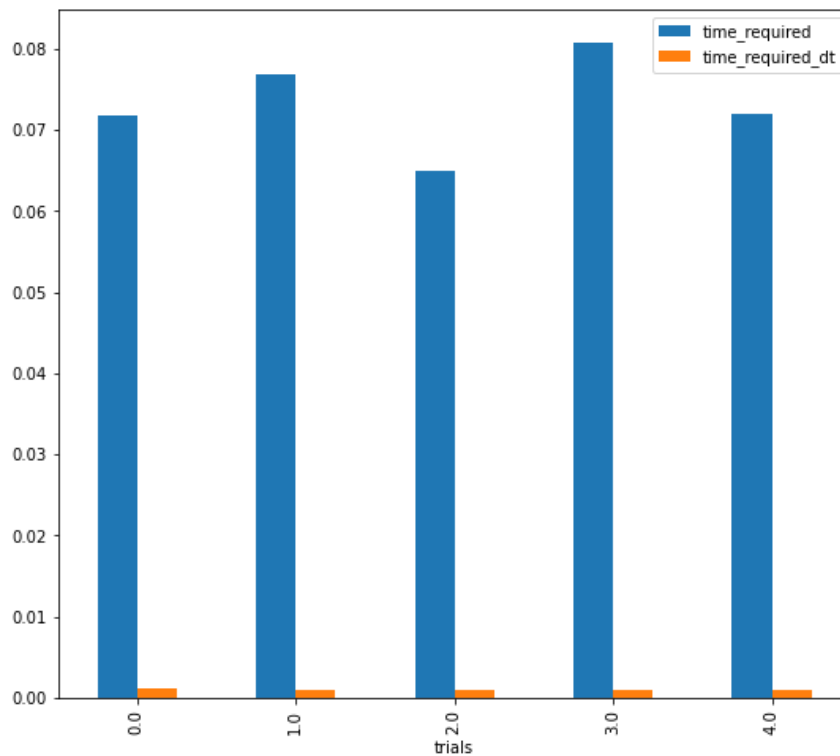
```
9]: fullreport2.plot(x='trials', y=['Fmeasure_test', 'Fmeasure_test_dt'], kind="bar", figsize=(9, 8))
it[29]: <AxesSubplot:xlabel='trials'>
```



Same kind of representation is evident for the f_measure of test data.

```
fullreport2.plot(x='trials', y=['time_required','time_required_dt'], kind="bar", figsize=(9, 8))
```

```
l]: <AxesSubplot:xlabel='trials'>
```

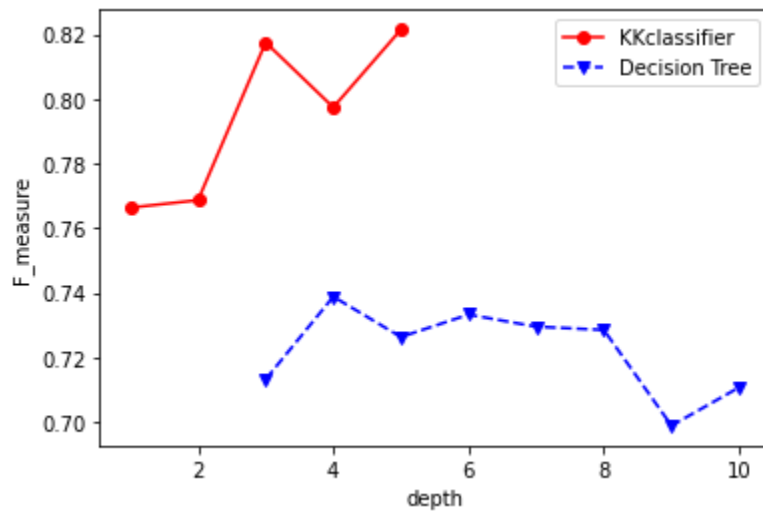


Here we see that time required for kk classifier is quite higher than the time required for decision tree. We observed the same thing for the other dataset.

```
import statistics
mean_train_f=statistics.mean(fullreport2["Fmeasure_train"])
mean_test_f=statistics.mean(fullreport2["Fmeasure_test"])
mean_train_f_dt=statistics.mean(fullreport2["Fmeasure_train_dt"])
mean_test_f_dt=statistics.mean(fullreport2["Fmeasure_test_dt"])
mean_time=statistics.mean(fullreport2["time_required"])
mean_time_dt=statistics.mean(fullreport2["time_required_dt"])
mean_test_f,mean_train_f,mean_test_f_dt,mean_train_f_dt,mean_time,mean_time_dt
```

```
]: (0.8138759148571306,
    0.8882819538602845,
    0.7000124466271592,
    0.7203750977211254,
    0.0732461929321289,
    0.0010149002075195313)
```

The fmeasure for test data is less than fmeasure for train data, which is expected. We see the same phenomenon for the decision tree. mean time required is quite less for decision tree.



Here the k value of 5 gives the highest f measure for KK classifier. So the value of k will be 5 for this data set and for kk classifier. But for decision tree we see that the fmeasure is quite low for all the depths, which wasn't the case for previous dataset. For this data set KK classifier is better choice as a classifier.