Midterm Lab Report

Mayur Bhai and Vashist Patel

CSCI 4150U: Data Mining

Dr. Kourosh Davoudi

March 5, 2021

Lab 2

Part 1:

2.3 - Most people work in prod-specialty and craft-repair

2.4 - Majority of the people are aged around 35 years old and most of them are under 50 years old.

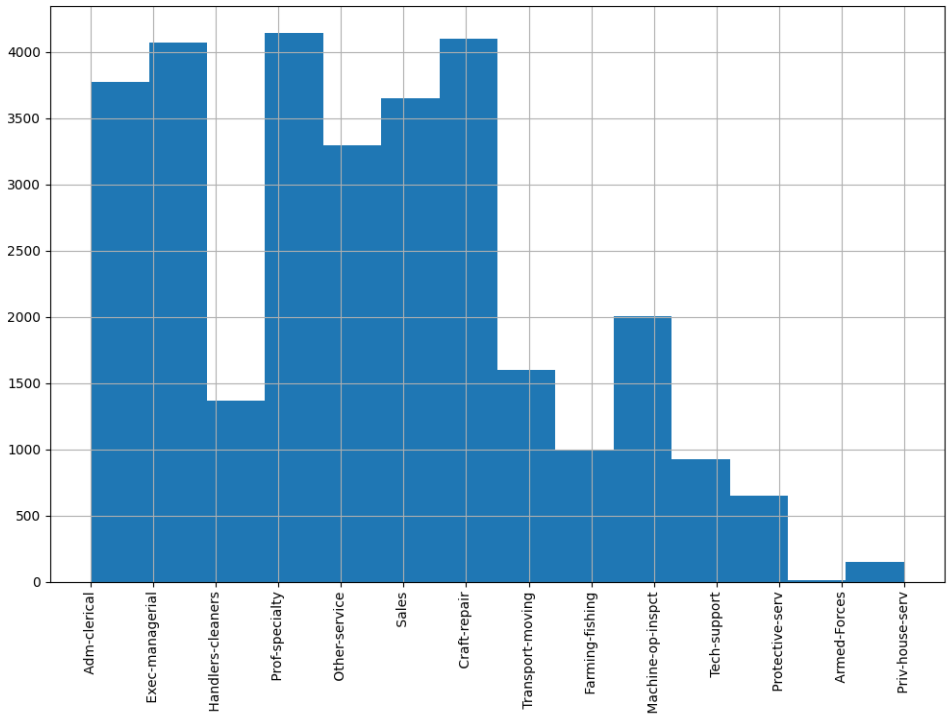
2.5

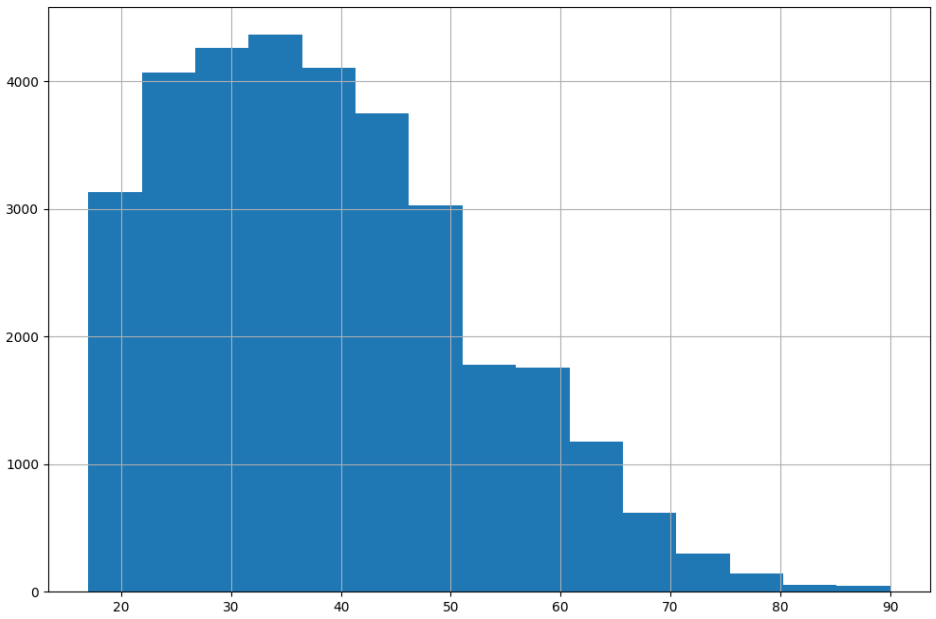
- Each education level has people in it with a variety of different ages. Education level preschool is staggered as. Less people are expected to be in the group.
- capital gain of 100000 is observed in age groups up to around 80 years of age. Between 40000 and 100000, there are almost 0. people with capital gain in that range. Most people have a capital gain between 0 and 20000.
- a lot of people with capital loss between 1000 and 3000. Every age group has someone with 0 capital loss
- Lost of people in the age group 70 to 90 dont own-child.
- white people have the most diverse range of captial gains. As the population of each group decreases, smaller the range of captial-gain. So if you are an 'other' then you are most likely to have a captial gain of 0-20000 or 100000

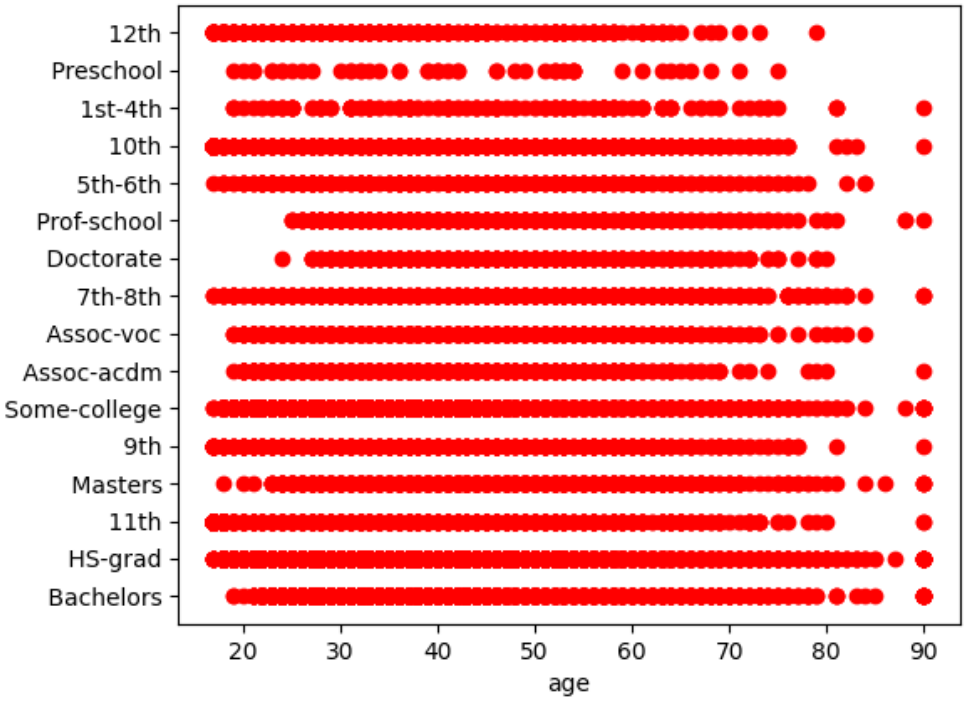
2.6

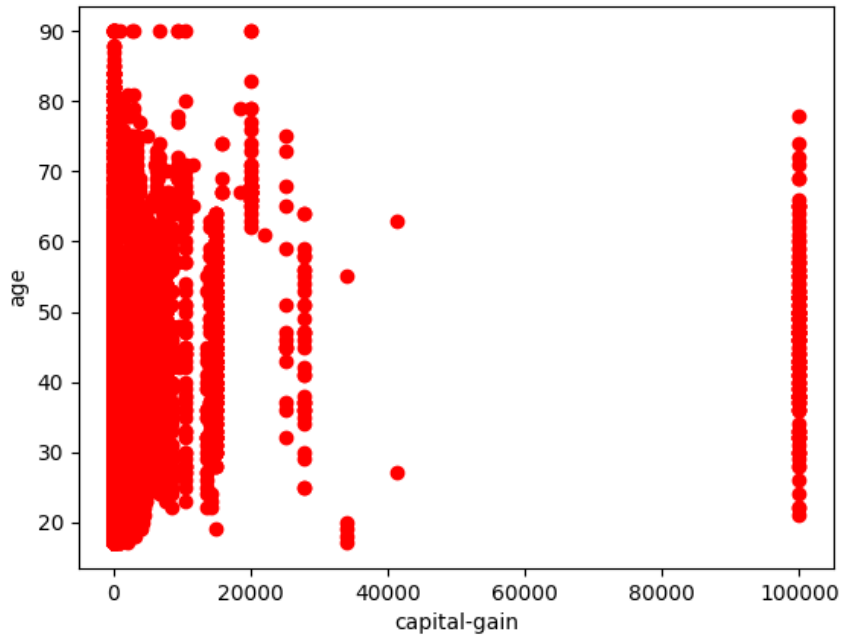
- Younger the person, less hours they work. A lot more males with high hours-per-week then females.
- White race has higher capital gain

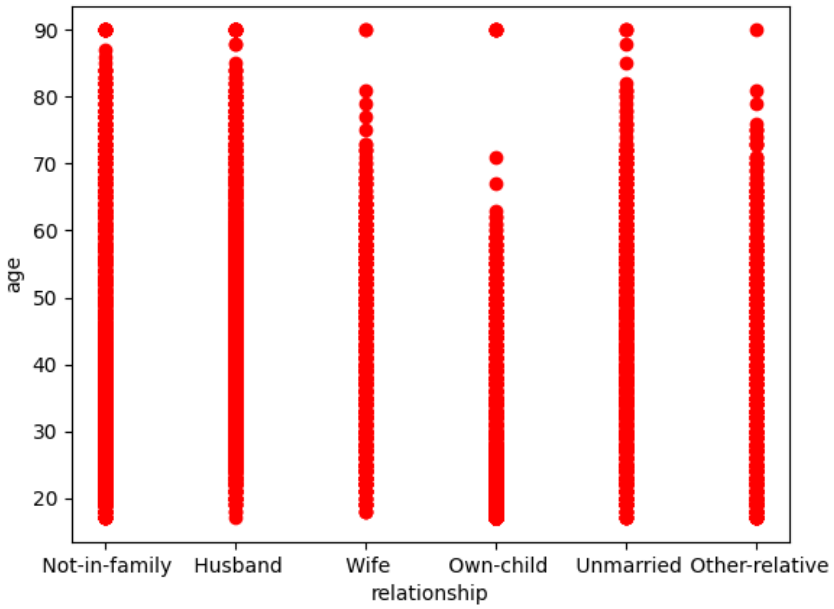
```
C:\Users\Mayor\AppData\Local\Programs\Python\Python38-32\python.exe "C:/Users/Mayor/IdeaProjects/Python Project/lab2_part1.py"
age:
     Mean = 38.58
     Standard deviation = 13.64
     Minimum = 17.00
     Maximum = 90.00
fnlwgt:
     Mean = 189778.37
     Standard deviation = 105549.98
     Minimum = 12285.00
     Maximum = 1484705.00
education-num:
     Mean = 10.08
     Standard deviation = 2.57
     Minimum = 1.00
     Maximum = 16.00
capital-gain:
     Mean = 1077.65
     Standard deviation = 7385.29
     Minimum = 0.00
     Maximum = 99999.00
capital-loss:
     Mean = 87.30
     Standard deviation = 402.96
     Minimum = 0.00
     Maximum = 4356.00
hours-per-week:
     Mean = 40.44
     Standard deviation = 12.35
     Minimum = 1.00
     Maximum = 99.00
```

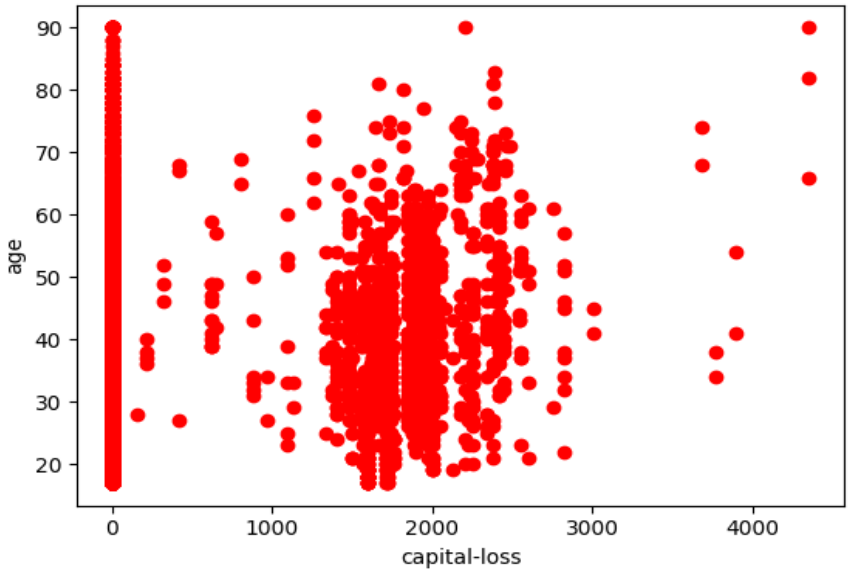

4000
3500
3000
2500
2000
1500
1000
500
0
Adm-clerical
Exec-managerial
Handlers-cleaners
Prof-specialty
Other-service
Sales
Craft-repair
Transport-moving
Farming-fishing
Machine-op-inspct
Tech-support
Protective-serv
Armed-Forces
Priv-house-serv

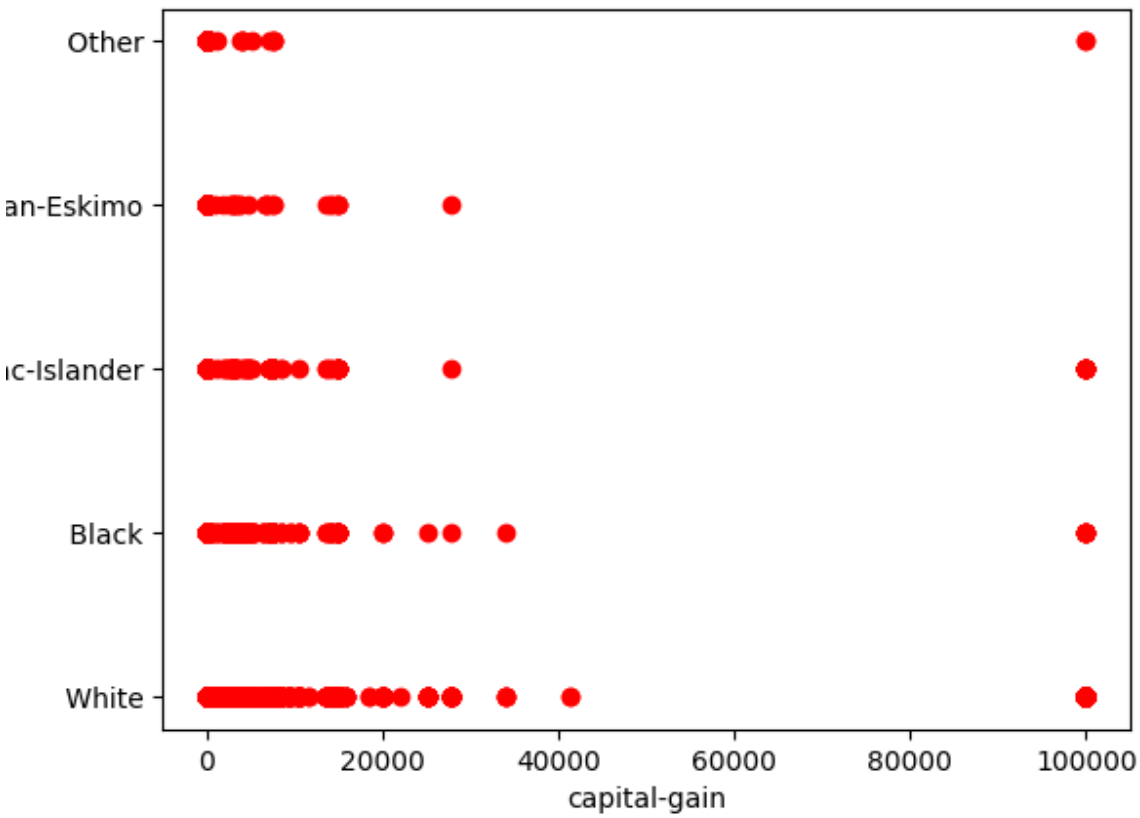

4000
3000
2000
1000
0
20
30
40
50
60
70
80
90


12th
Preschool
1st-4th
10th
5th-6th
Prof-school
Doctorate
7th-8th
Assoc-voc
Assoc-acdm
Some-college
9th
Masters
11th
HS-grad
Bachelors
20
30
40
50
60
70
80
90
age

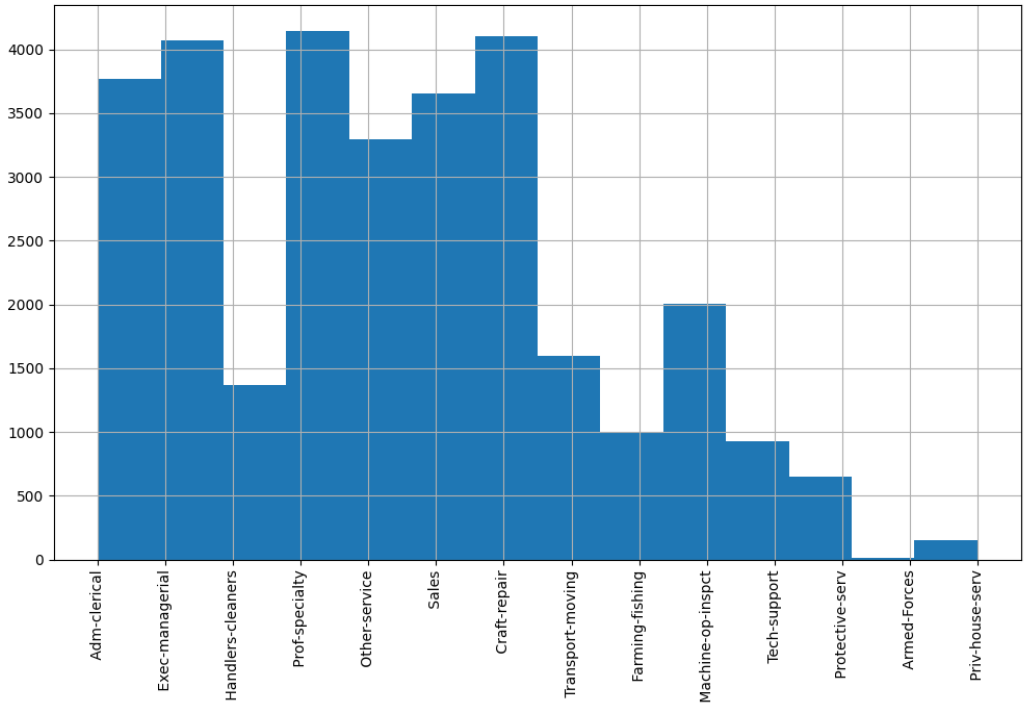

90
80
70
60
50
40
30
20
age
0
20000
40000
60000
80000
100000
capital-gain

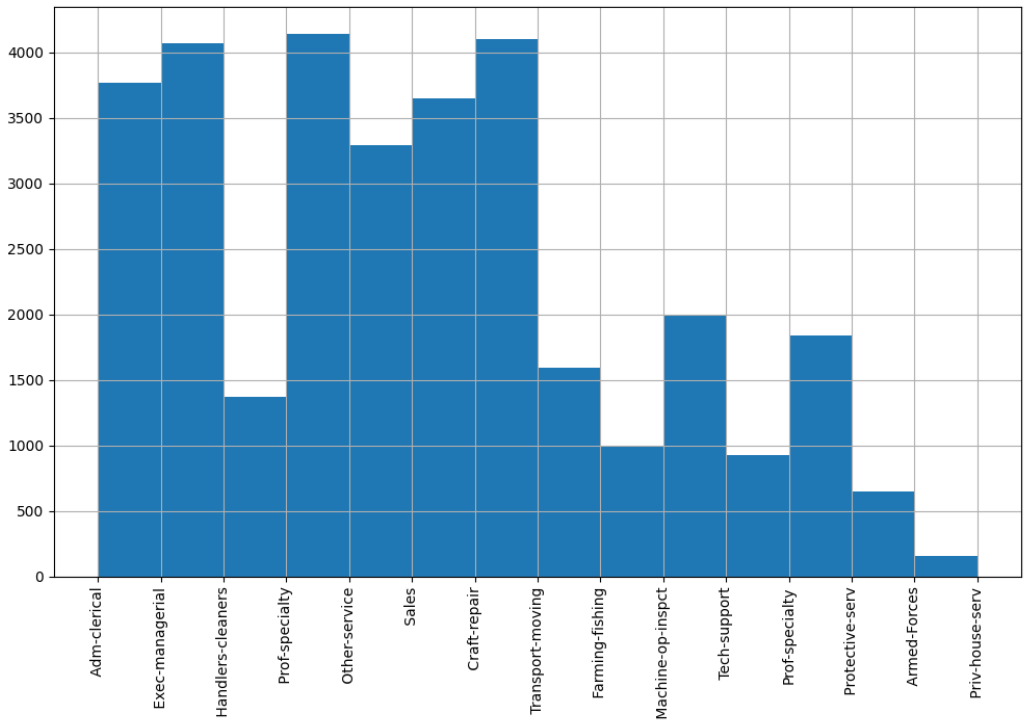

90
80
70
60
50
40
30
20
age
Not-in-family
Husband
Wife
Own-child
Unmarried
Other-relative
relationship

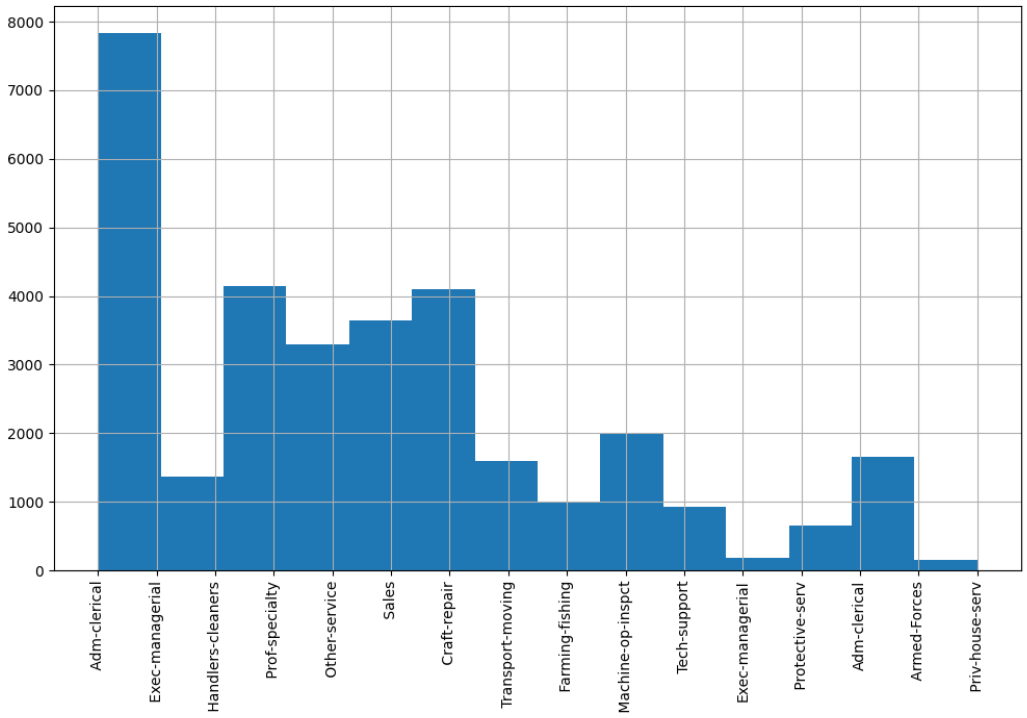

90
80
70
60
50
40
30
20
age
0
1000
2000
3000
4000
capital-loss


Other
an-Eskimo
ıc-Islander
Black
White
0
20000
40000
60000
80000
100000
capital-gain

Part 2:


4000
3500
3000
2500
2000
1500
1000
500
0
Adm-clerical
Exec-managerial
Handlers-cleaners
Prof-specialty
Other-service
Sales
Craft-repair
Transport-moving
Farming-fishing
Machine-op-inspct
Tech-support
Protective-serv
Armed-Forces
Priv-house-serv


4000
3500
3000
2500
2000
1500
1000
500
0
Adm-clerical
Exec-managerial
Handlers-cleaners
Prof-specialty
Other-service
Sales
Craft-repair
Transport-moving
Farming-fishing
Machine-op-inspct
Tech-support
Prof-specialty
Protective-serv
Armed-Forces
Priv-house-serv


8000
7000
6000
5000
4000
3000
2000
1000
0
Adm-clerical
Exec-managerial
Handlers-cleaners
Prof-specialty
Other-service
Sales
Craft-repair
Transport-moving
Farming-fishing
Machine-op-inspct
Tech-support
Exec-managerial
Protective-serv
Adm-clerical
Armed-Forces
Priv-house-serv

Lab 3

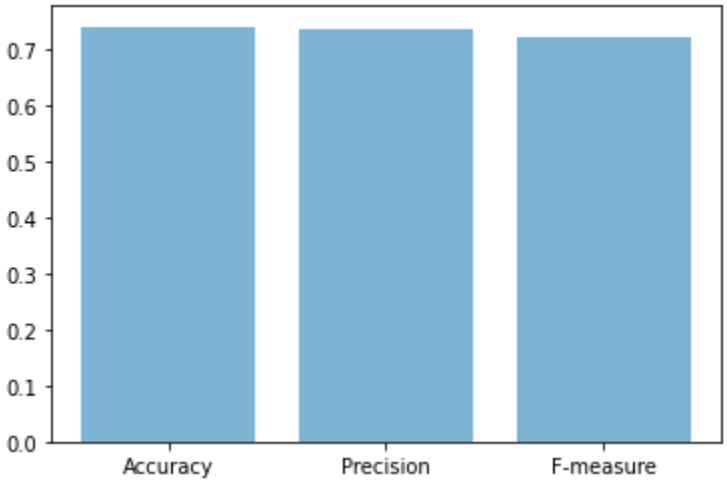
German dataset

Part 1

Holdout

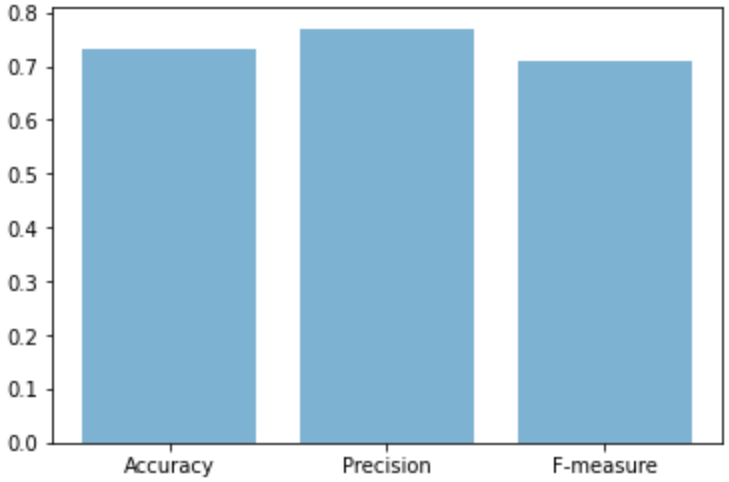
```
Accuracy Average: 0.74
Precision Average: 0.74
F-Measure Average: 0.72
```

	pass1	pass2	pass3	pass4	pass5
Accuracy	0.79	0.71	0.68	0.79	0.74
Precision	0.782936194240542	0.7058610954263129	0.6681912681912682	0.7679221927497789	0.7546666666666667
F-measure	0.7856575212866603	0.707862606035322	0.6730158730158731	0.765869673354336	0.6900060277275466



Cross-validation

	pass1	pass2	pass3	pass4	pass5	pass6	pass7	pass8	pass9	pass10
Accuracy	0.8	0.71	0.72	0.72	0.75	0.72	0.68	0.74	0.75	0.68
Precision	0.8378378378378378	0.759493670886076	0.7763157894736842	0.7625	0.7586206896551724	0.7282608695652174	0.7317073170731707	0.7972972972972973	0.7848101265822784	0.796875
F-measure	0.7956349206349206	0.6931701539676273	0.7101978691019787	0.7013333333333334	0.714116427195971	0.657959714100065	0.6526315789473685	0.7343253968253967	0.735491512041058	0.6873812754409769

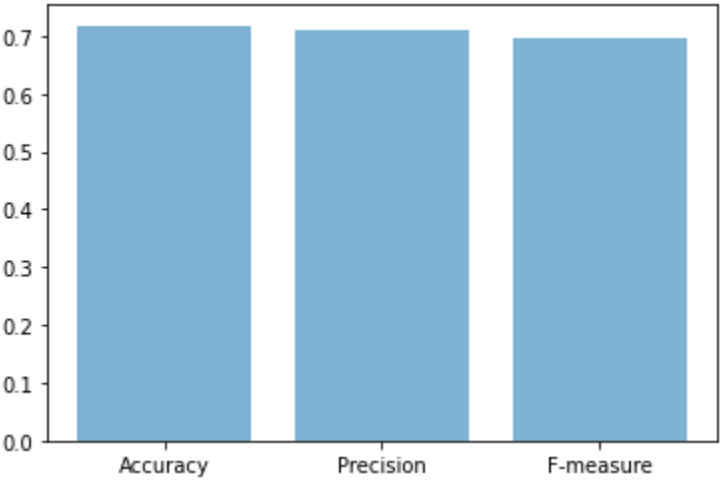


Part 2

Holdout

```
Accuracy Average: 0.72
Precision Average: 0.71
F-Measure Average: 0.70
```

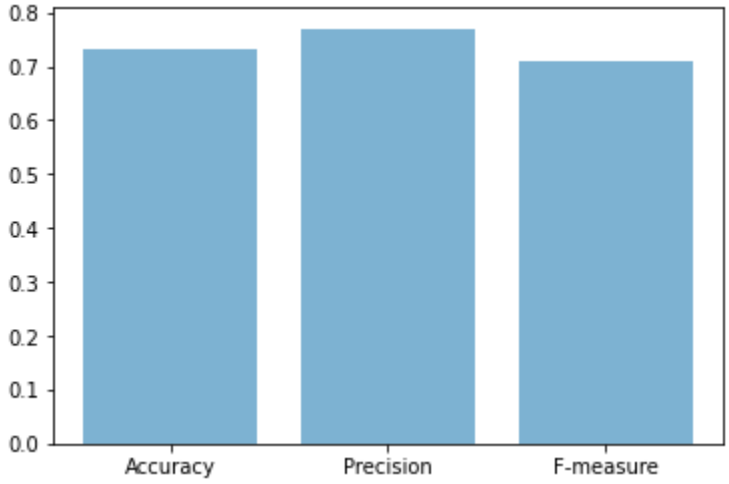
	pass1	pass2	pass3	pass4	pass5
Accuracy	0.71	0.74	0.68	0.75	0.71
Precision	0.7078749999999999	0.7346179401993355	0.6945845390657297	0.7403846153846154	0.6733720930232558
F-measure	0.6839921482026745	0.704325239977414	0.6848896434634975	0.7344827586206897	0.6732721078358762



Cross-validation

```
Final Average Accuracy of the model: 0.73

Final Average precision of the model: 0.77

Final Average F-measure of the model: 0.71
```

	pass1	pass2	pass3	pass4	pass5	pass6	pass7	pass8	pass9	pass10
Accuracy	0.76	0.71	0.73	0.72	0.75	0.72	0.7	0.76	0.75	0.69
Precision	0.7446808510638298	0.759493670886076	0.7792207792207793	0.7625	0.7586206896551724	0.7916666666666666	0.7272727272727273	0.8026315789473685	0.7848101265822784	0.8
F-measure	0.697560975609756	0.6931701539676273	0.7185983827493261	0.7013333333333334	0.714116427195971	0.7171442447790188	0.6528028933092225	0.7515981735159818	0.735491512041058	0.6961823361823362



Part 3

When tree depth is 3:

- Accuracy Average: 0.72
- Precision Average: 0.72
- F-Measure Average: 0.71

When tree depth is 5:

- Accuracy Average: 0.74
- Precision Average: 0.74

- F-Measure Average: 0.73

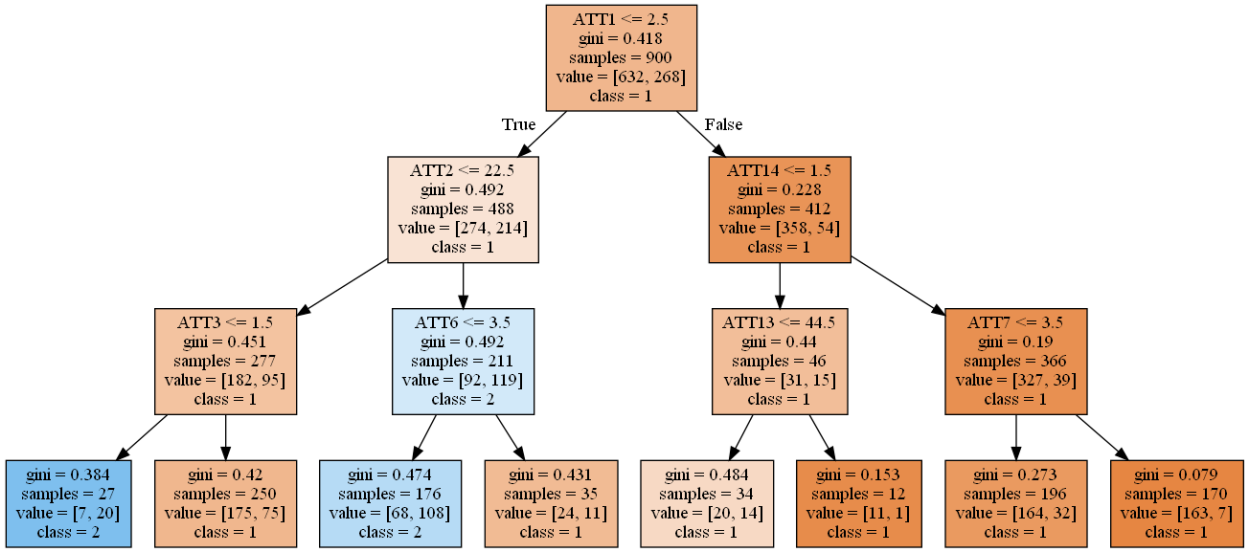
When tree depth is 10:

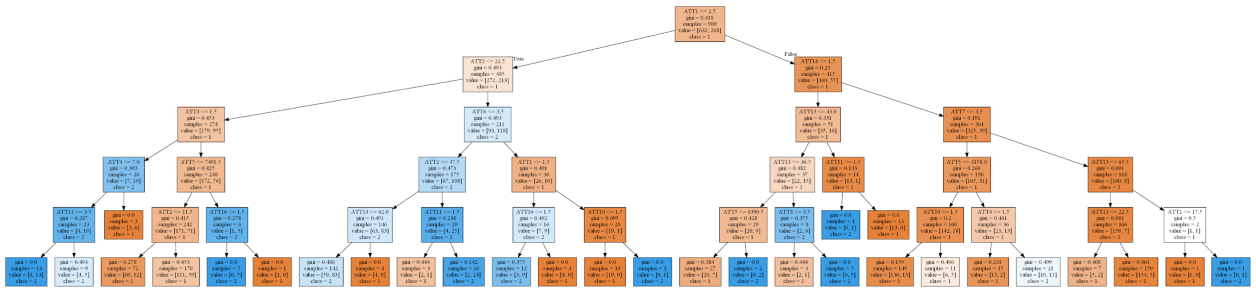
- Accuracy Average: 0.70
- Precision Average: 0.71
- F-Measure Average: 0.70

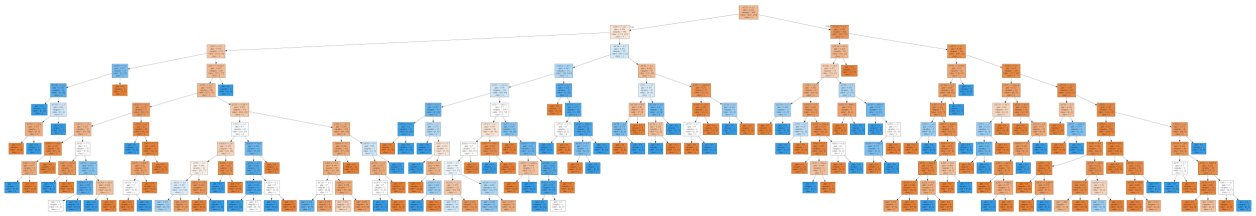
When tree depth is 50:

- Accuracy Average: 0.68
- Precision Average: 0.68
- F-Measure Average: 0.67

Therefore, after a tree depth of 5, the accuracy of the decision tree model decreases, making the model less accurate. Tree depth of 1 to 5 is increasing in accuracy making tree depth of 5 with the best accuracy.







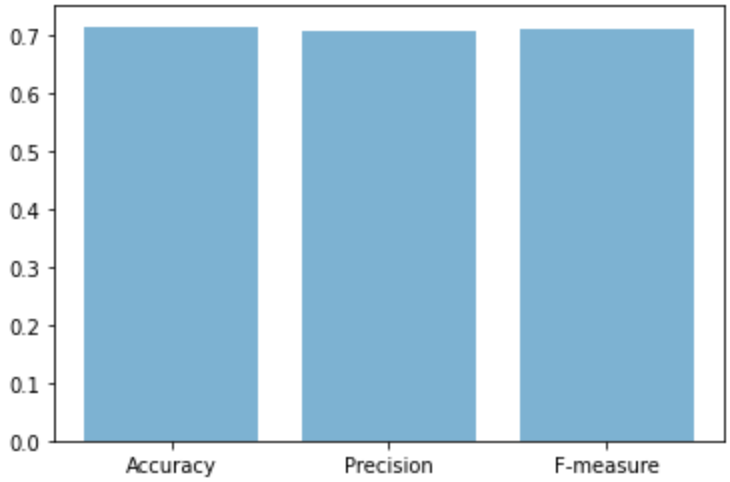
WaveForm dataset

Part 1

Holdout

```
Accuracy Average: 0.72
Precision Average: 0.71
F-Measure Average: 0.71
```

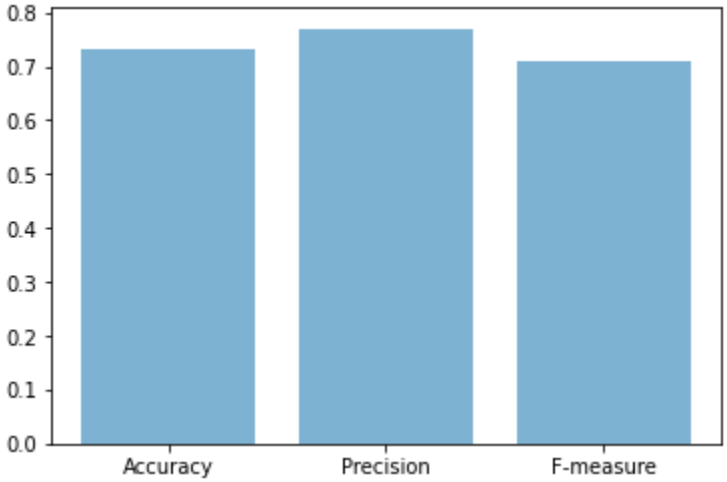
	pass1	pass2	pass3	pass4	pass5
Accuracy	0.672	0.706	0.722	0.72	0.758
Precision	0.6676526312454456	0.6893535425351991	0.7297257095257927	0.7206509901665844	0.7323906240131716
F-measure	0.6695864020057434	0.69960121019288	0.7191817054223583	0.7177569707968983	0.7526256713627822



Cross-validation

```
Final Average Accuracy of the model: 0.73

Final Average precision of the model: 0.77

Final Average F-measure of the model: 0.71
```

	pass1	pass2	pass3	pass4	pass5	pass6	pass7	pass8	pass9	pass10
Accuracy	0.8	0.71	0.72	0.72	0.75	0.72	0.68	0.74	0.75	0.68
Precision	0.8378378378378378	0.7594936708860076	0.7763157894736842	0.7625	0.7586206896551724	0.7282608695652174	0.7317073170731707	0.7972972972972973	0.7848101265822784	0.796875
F-measure	0.7956349206349206	0.6931701539676273	0.7101978691019787	0.7013333333333334	0.714116427195971	0.6579597141000965	0.6526315789473685	0.7343253968253967	0.735491512041058	0.6873812754409769

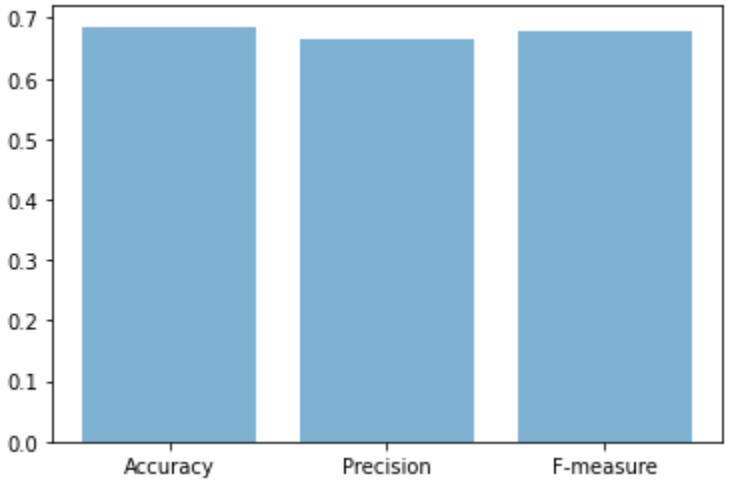


Part 2

Holdout

```
Accuracy Average: 0.69
Precision Average: 0.67
F-Measure Average: 0.68
```

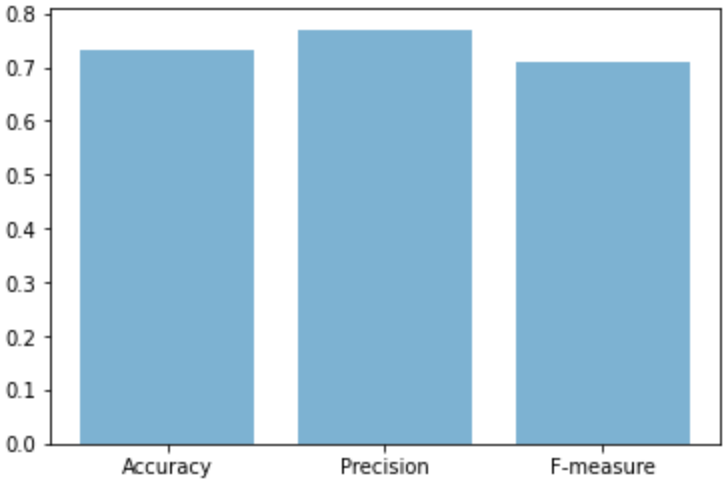
	pass1	pass2	pass3	pass4	pass5
Accuracy	0.68	0.68	0.666	0.694	0.71
Precision	0.65666918264136	0.6623671595116796	0.6304083834180922	0.6963109216395365	0.6851352503402176
F-measure	0.667847238455175	0.6754791593202738	0.6595935072463767	0.6844832389795372	0.70227038568857



Cross-validation

```
Final Average Accuracy of the model: 0.73

Final Average precision of the model: 0.77

Final Average F-measure of the model: 0.71
```

	pass1	pass2	pass3	pass4	pass5	pass6	pass7	pass8	pass9	pass10
Accuracy	0.76	0.71	0.73	0.72	0.75	0.72	0.7	0.76	0.75	0.69
Precision	0.7446808510638298	0.759493670886076	0.7792207792207793	0.7625	0.7586206896551724	0.7916666666666666	0.7272727272727273	0.8026315789473685	0.7848101265822784	0.8
F-measure	0.697560975609756	0.6931701539676273	0.7185983827493261	0.7013333333333334	0.714116427195971	0.7171442447790188	0.6528028933092225	0.7515981735159818	0.735491512041058	0.6961823361823362



Part 3

When tree depth is 3:

- Accuracy Average: 0.72
- Precision Average: 0.70
- F-Measure Average: 0.71

When tree depth is 5:

- Accuracy Average: 0.76
- Precision Average: 0.77
- F-Measure Average: 0.76

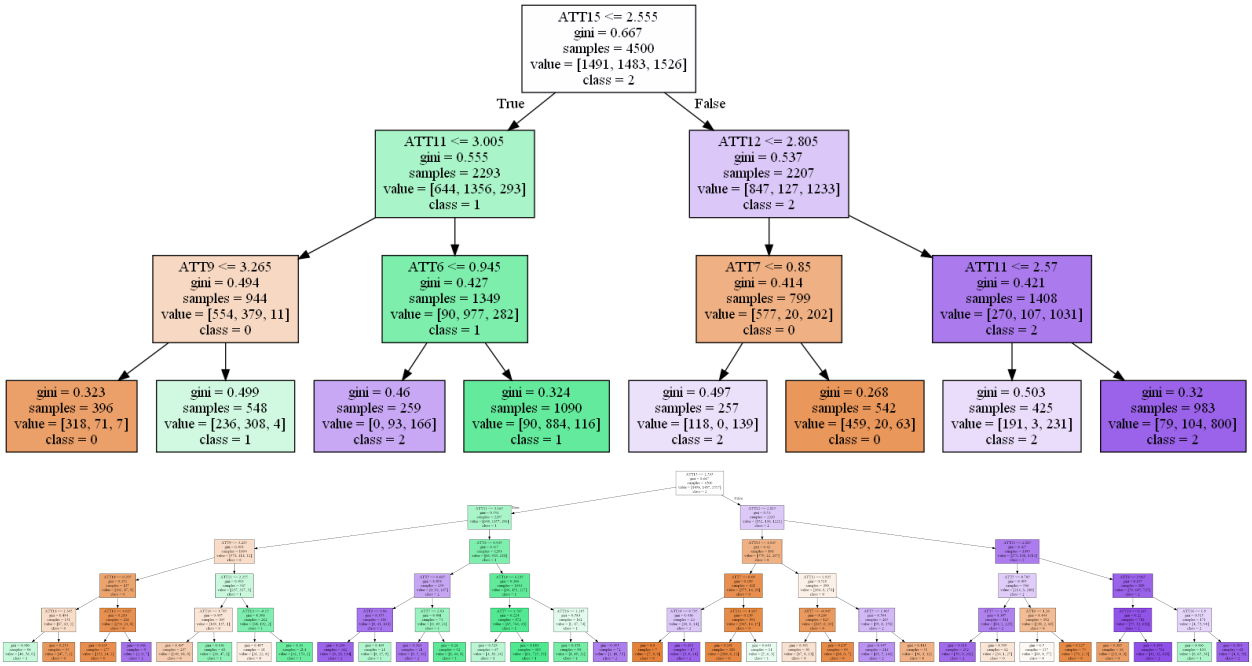
When tree depth is 10:

- Accuracy Average: 0.77
- Precision Average: 0.78
- F-Measure Average: 0.77

When tree depth is 50:

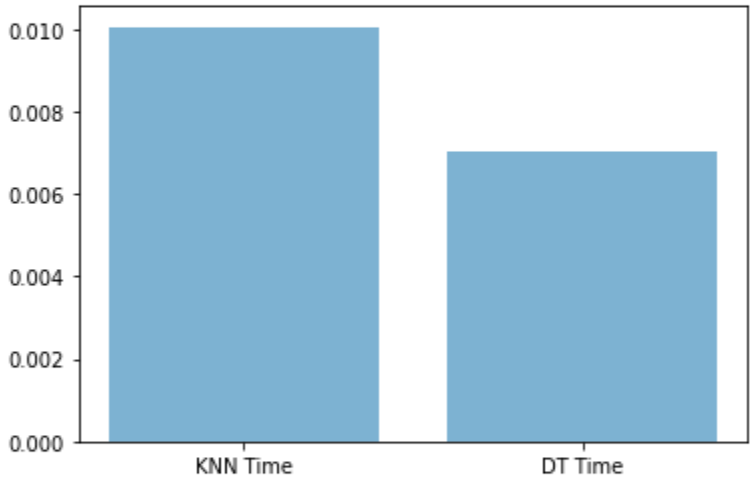
- Accuracy Average: 0.75
- Precision Average: 0.76
- F-Measure Average: 0.75

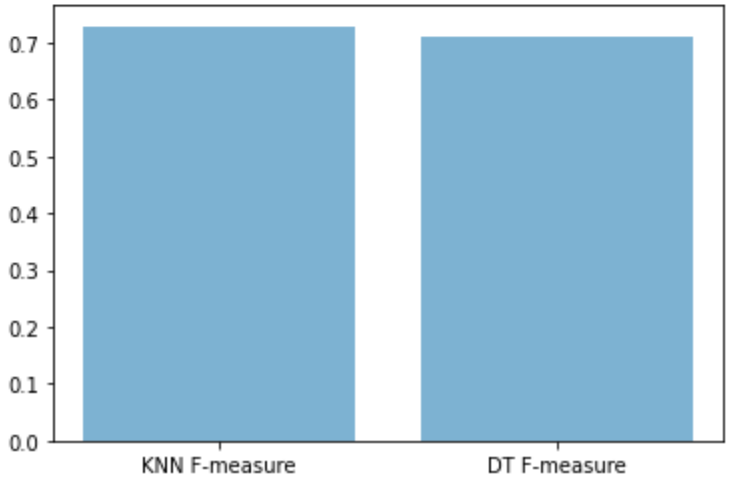
Therefore, after a tree depth of 10, the accuracy of the decision tree model decreases, making the model less accurate. Tree depth of 1 to 10 is increasing in accuracy slightly making tree depth of around 10 with the best accuracy.



Lab 4 German dataset
Part 1

```
Running KNN Algorithim
Accuracy of the model on Testing Sample Data: 0.71
Accuracy of the model on Testing Sample Data: 0.83
Accuracy of the model on Testing Sample Data: 0.73
Accuracy of the model on Testing Sample Data: 0.71
Accuracy of the model on Testing Sample Data: 0.67
F-Measure Average: 0.729
Time Average: 0.0101

Running DT Algorithim
Accuracy of the model on Testing Sample Data: 0.692
Accuracy of the model on Testing Sample Data: 0.74
Accuracy of the model on Testing Sample Data: 0.701
Accuracy of the model on Testing Sample Data: 0.751
Accuracy of the model on Testing Sample Data: 0.659
F-Measure Average: 0.709
Time Average: 0.0070
```

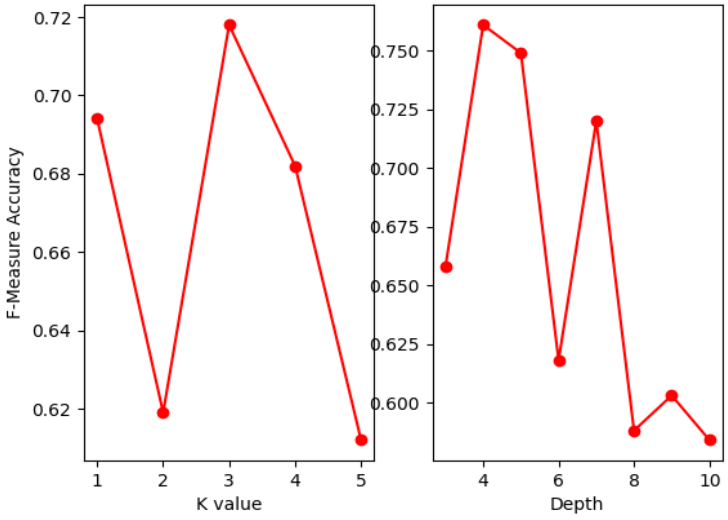




When comparing the average F_measureof the KNN model to a decision tree model, we see that KNN has a slightly higher F-measure then a decision tree model. When comparing the average time it takes to run KNN and a DT model, we see that KNN takes a longer time to compute than DT model. DT model takes 2 thirds of the time of KNN.

Part 2

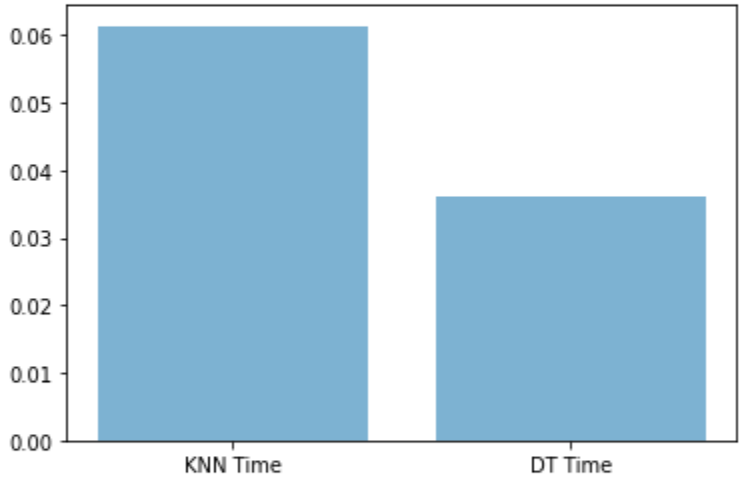
```
C:\Users\Mayor\AppData\Local\Programs\Python\Python38-32\python.exe "C:/Users/Mayor/IdeaProjects/Python Project/Lab 4_Part2.py"
Running KNN Algorithim
K = 1
F_Measure of the model on Testing Sample Data:  0.732
Accuracy of the valdiation:  0.694
K = 2
F_Measure of the model on Testing Sample Data:  0.698
Accuracy of the valdiation:  0.619
K = 3
F_Measure of the model on Testing Sample Data:  0.674
Accuracy of the valdiation:  0.718
K = 4
F_Measure of the model on Testing Sample Data:  0.663
Accuracy of the valdiation:  0.682
K = 5
F_Measure of the model on Testing Sample Data:  0.692
Accuracy of the valdiation:  0.612

Running DT Algorithim
Depth = 3
Accuracy of the model on Testing Sample Data: 0.697
Accuracy of the valdiation:  0.658
Depth = 4
Accuracy of the model on Testing Sample Data: 0.788
Accuracy of the valdiation:  0.761
Depth = 5
Accuracy of the model on Testing Sample Data: 0.745
Accuracy of the valdiation:  0.749
Depth = 6
Accuracy of the model on Testing Sample Data: 0.714
Accuracy of the valdiation:  0.618
Depth = 7
Accuracy of the model on Testing Sample Data: 0.703
Accuracy of the valdiation:  0.72
Depth = 8
Accuracy of the model on Testing Sample Data: 0.725
Accuracy of the valdiation:  0.588
Depth = 9
Accuracy of the model on Testing Sample Data: 0.728
Accuracy of the valdiation:  0.603
Depth = 10
Accuracy of the model on Testing Sample Data: 0.631
Accuracy of the valdiation:  0.584
```

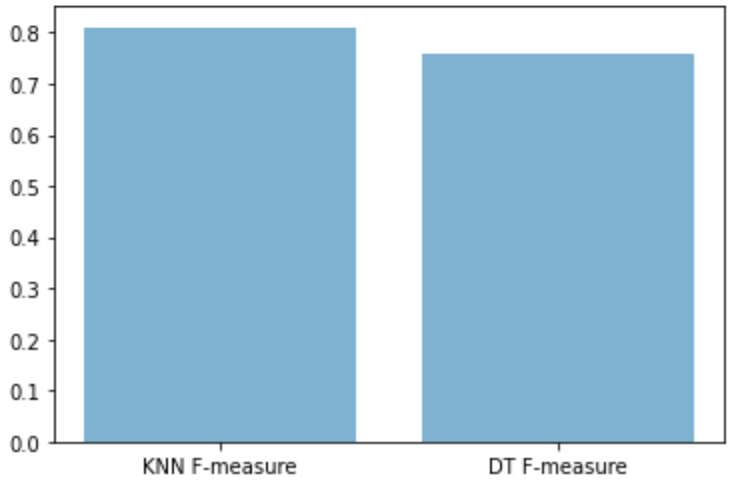


Therefore, the best k for KNN is 3 and the best depth is 4 in order to obtain the most accurate F-measure. The two graphs side by side show how the accuracy changes depending on the K/depth value.

Waveform data set
Part 1

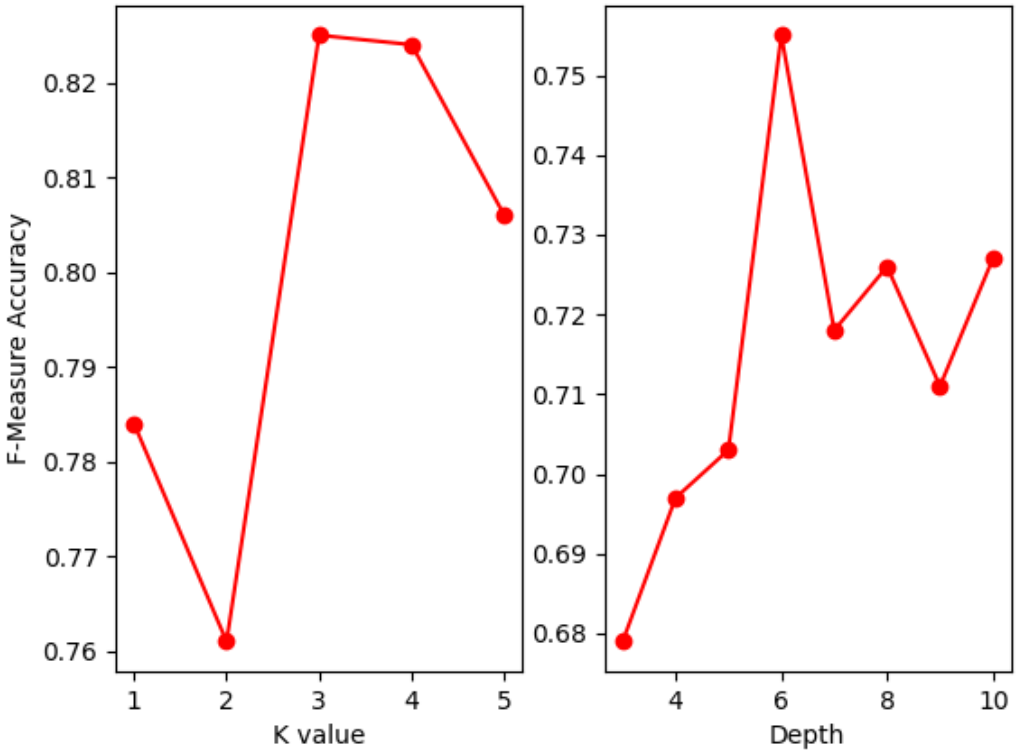
```
Running KNN Algorithim
Accuracy of the model on Testing Sample Data: 0.8
Accuracy of the model on Testing Sample Data: 0.8
Accuracy of the model on Testing Sample Data: 0.82
Accuracy of the model on Testing Sample Data: 0.81
Accuracy of the model on Testing Sample Data: 0.83
F-Measure Average: 0.811
Time Average: 0.0613

Running DT Algorithim
Accuracy of the model on Testing Sample Data: 0.725
Accuracy of the model on Testing Sample Data: 0.773
Accuracy of the model on Testing Sample Data: 0.79
Accuracy of the model on Testing Sample Data: 0.75
Accuracy of the model on Testing Sample Data: 0.754
F-Measure Average: 0.758
Time Average: 0.0361
```





When comparing the average F_measure of the KNN model to a decision tree model, we see that KNN has a slightly higher F-measure then a decision tree model. When comparing the average time it takes to run KNN and a DT model, we see that KNN takes a longer time to compute than DT model. DT model takes 2 thirds of the time of KNN.

Part 2



Therefore, the best k for KNN is 3 and the best depth is 6 in order to obtain the most accurate F-measure. The two graphs side by side show how the accuracy changes depending on the K/depth value.

Lab 5

```
Running KNN Model
KNN Stats:
F-Measure Average: 0.79977
Accuracy Average: 0.80000
Precision Average: 0.84369
F-Measure Validation Average: 0.823
Time Average: 0.07676
---------------------------------------

Running Decision Tree Model
DT Stats:
F-Measure Average: 0.97096
Accuracy Average: 0.97100
Precision Average: 0.97243
F-Measure Validation Average: 0.963
Time Average: 0.05214
---------------------------------------

Running NaiveBayes Model
NaiveBayes Stats:
F-Measure Average: 0.92962
Accuracy Average: 0.93000
Precision Average: 0.95900
F-Measure Validation Average: 0.939
Time Average: 0.05190
---------------------------------------
```

First we decided to build the KNN model with a k value of 3 because based on our previous experiments we found out that it gives the best F-measure score. Next we built a decision tree model with tree depth of 5. We picked 5 because it gave us the highest accuracy in our previous experiments. And finally we built NaiveBayes for our final model.

When comparing the data from the three models, let's look at the results of the three average F-measure scores. With a 0.97 decision tree model has the best score compared to KNN and NaiveBayes. This suggests that the dating site should use the decision tree model to improve its recommendations. The decision tree model will help the site understand how the clients are categorized and predict the clients actions. This will make it so they can give a better recommendation.