Assignment #3

**Classification-**

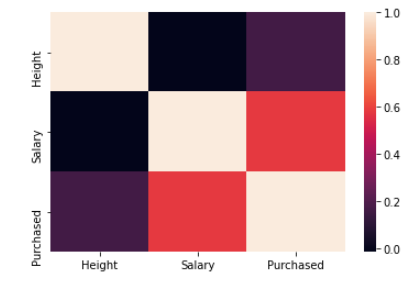
**Code explains**

**First stage – Data preparation**

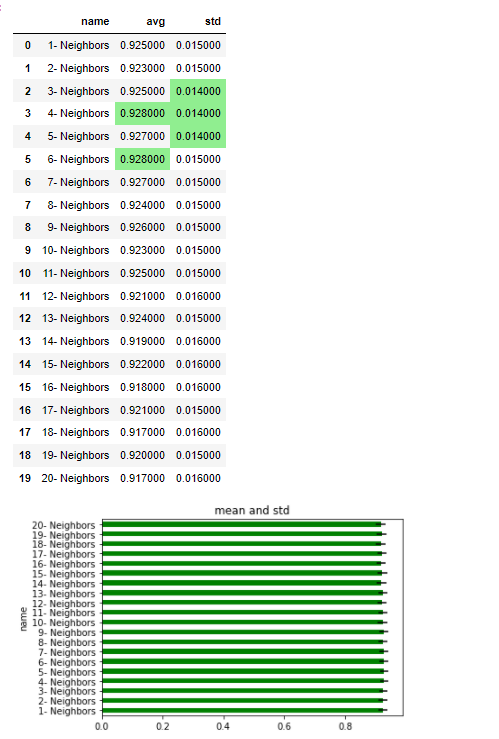
First I read the data from the CSV file, then I maid a counterplot of Purchased (0 or 1 ) and the percentage from total

****

and then I checked there is no correlation between the features



**second stage k-nn-**



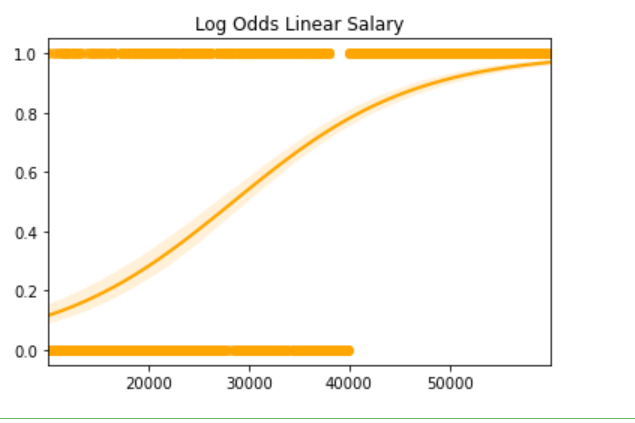
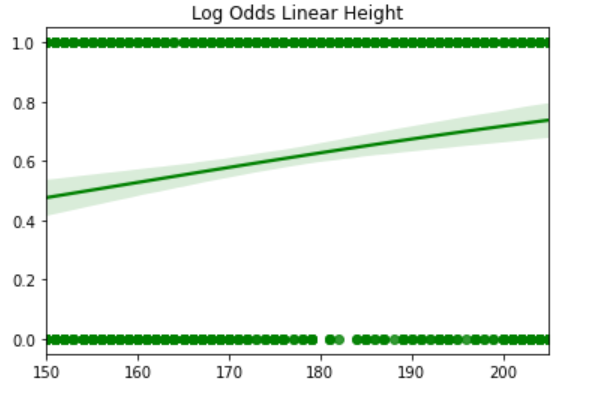
For the knn**, I will pick 4 neighbors** because this hyperparameter has the highest mean of f1 score and lowest standard deviation

**third stage logistic regression-**

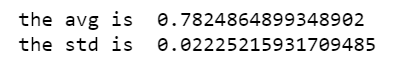
first I checked for multicollinearity (VIF that is greater than 5 )

according to VIF, there is multicollinearity

so I removed Height because the feature does not look log-linear oods, so the only feature for the logistic regression is salary

Then I checked that variables are linearly related to the log odds :

Then I applied to train and test and calculated the f1 score



The mean f1 score is 0.782

And the std is 0.022

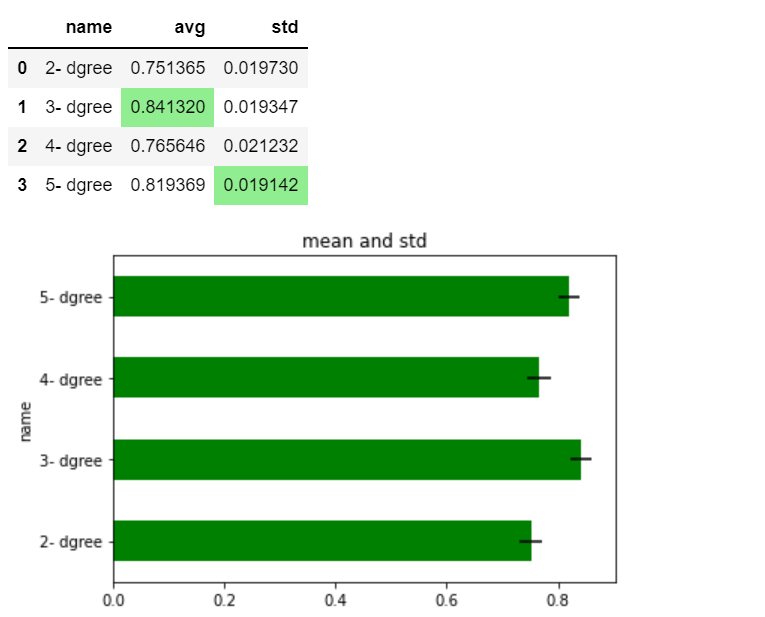
**fourth stage svc linear-**



The mean f1 score is 0.814

And the std is 0.0210

**fifth stage svc poly regression-**

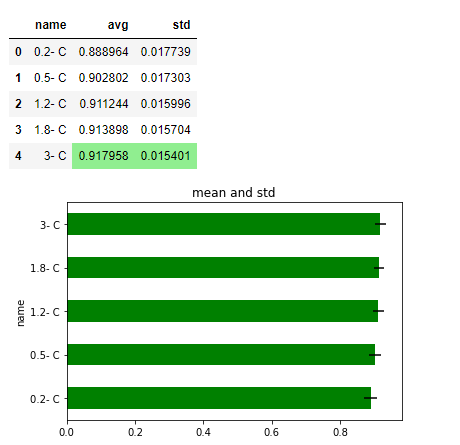


**The best degree is -3**

The mean f1 score is 0.841

And the std is 0.019  
  
which are better than the linear svc and logistic regression but still not so good

**six-stage svc gaussian-**



**Best C is -3**

The mean f1 score is 0.917

And the std is 0.015  
  
which is very good!