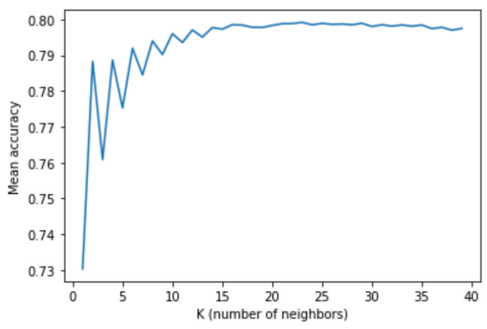
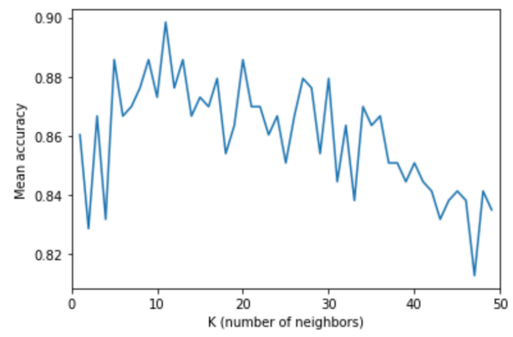
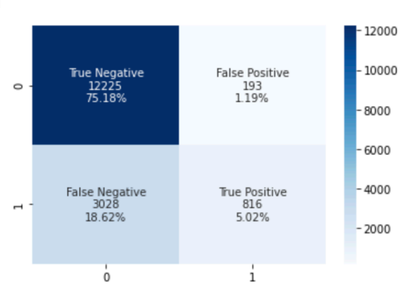
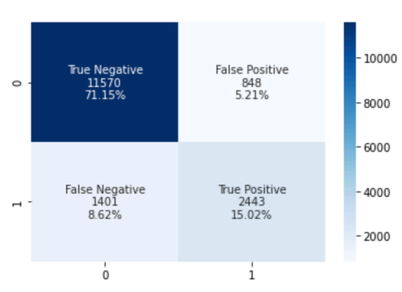
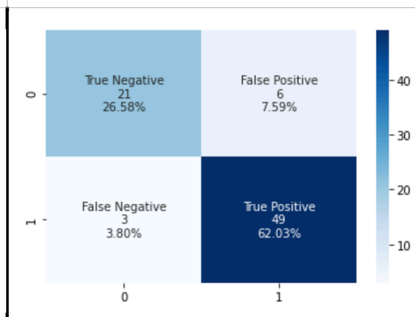
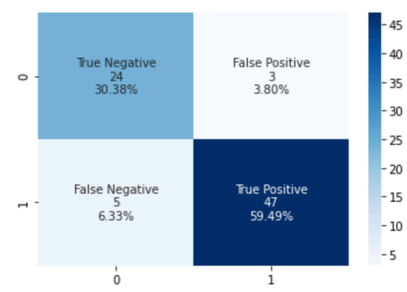
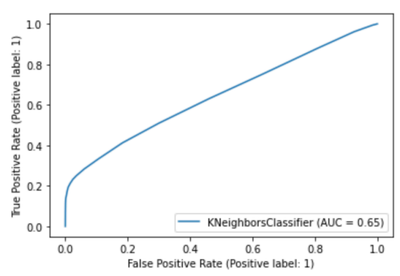
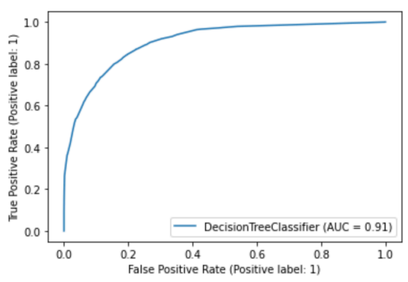
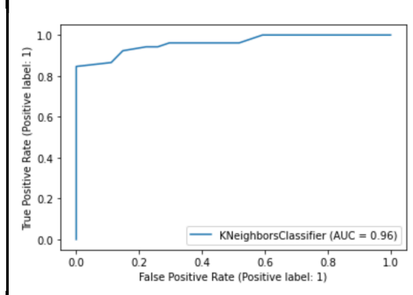
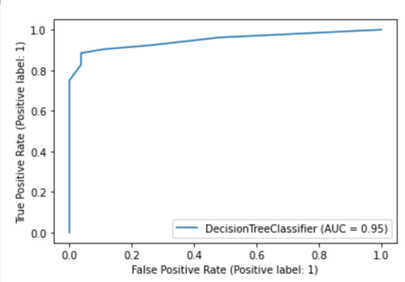
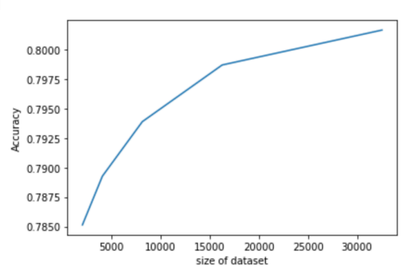
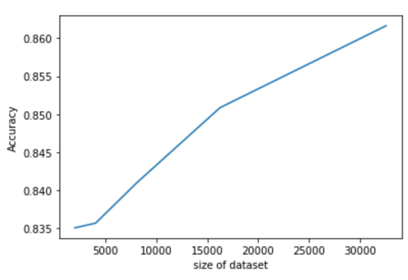
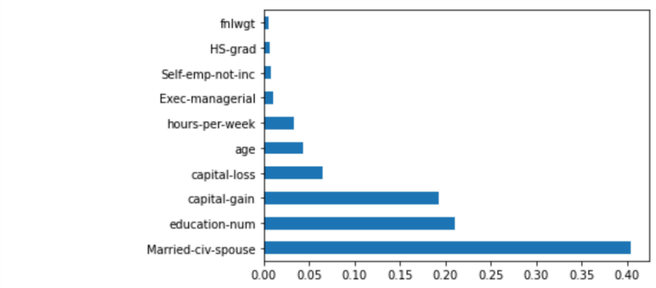
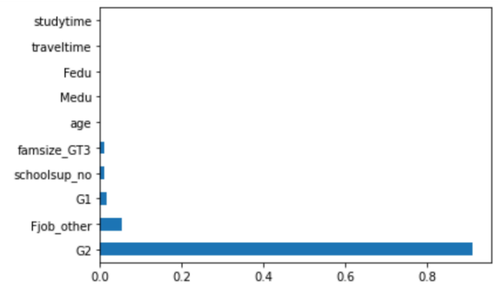
	Dataset 1		Dataset 2	
	KNN	Decision Tree	KNN	Decision Tree
Dataset size	(32561, 15)		(395,32)	
Dataset size after cleaning	(32534, 15)			
After onehot encoding	(32534, 105)		(395,58)	
Best Cross-validation accuracy	79.9%	86%	89.84%	93.99%
Hyperparameters	K=23	Max depth = 17, Min samples leaf = 4, Min impurity= 0.00014	K=15	Max depth = 3, Min samples leaf = 10, Min impurity= 0
Accuracy-K curve				
Test dataset size	(16262, 105)		(79,58)	
Testing Accuracy	**80.19%**	**86.17%**	**88.61%**	**89.87%**
AUC	0.8511	0.7836	0.8600	0.8964
Confusion Matrix				
ROC Curve				
Dataset Size Experiment	The accuracy is 0.7851 when dataset size is 2033 The accuracy is 0.7893 when dataset size is 4066 The accuracy is 0.7939 when dataset size is 8133 The accuracy is 0.7987 when dataset size is 16267 	The accuracy is 0.8351 when dataset size is 2033 The accuracy is 0.8357 when dataset size is 4066 The accuracy is 0.8410 when dataset size is 8133 The accuracy is 0.8509 when dataset size is 16267 		
Different data imputation techniques	Do nothing - Impute missing data with constant 0 Imputation Using k-NN with k=1 But the accuracy didn't change….			
Feature selection experiments				
	remove the most importance 3 features--Married-civ-spouse, education-num, capital-gain		remove G1 G2 G3	
	final accuracy=76.44% when K=38 (drop from 80.19%)	final accuracy = 83.69% when Max depth = 17, Min samples leaf = 4, Min impurity= 0.00014 (drop from 86.17%)		AUC = 0.5705128205128205

It's weird that the AUC printed by metrics.auc(fpr, tpr) is different from the one on the ROC curve