Naïve Bayes Classifier

Team members:

- Bassant Ehab Moustafa 2017170106 SC
- Ayaalla Mohamed Eltabey 2017170103 SC
- Asmaa Ali El-shiekh 2017170070 SC

1. Part 1 (Predict individual's income):

The model's priori: first we count the number of each option in the income column then we divide it over the count of all options to get probabilities, the sum of all probabilities must be equal to 1.

10-50K	50-80K	GT 80K
7232	1132	646

10-50K	50-80K	GT 80K
0.80266371	0.12563818	0.07169811

The model's conditional probabilities: we calculate the conditional probability for each input feature (age, gender, educ) by get the count then divide it over the row count to get the ratio. The sum of each row must be equal to 1.

- Conditional probability between income and age it is the probability of age given the income:

	Age			
Income		20-30	31-45	GT 45
	10-50K	1504	2492	3236
	50-80K	94	450	588
	GT 80K	44	220	382

	Age			
Income		20-30	31-45	GT 45
	10-50K	0.20796460	0.34457965	0.44745575
	50-80K	0.08303887	0.39752650	0.51943463
	GT 80K	0.06811146	0.34055728	0.59133127

- Conditional probability between income and gender it is the probability of gender given the income:

	Gender		
Income		F	M
	10-50K	3470	3762
	50-80K	325	807
	GT 80K	133	513

	Gender		
Income		F	M
	10-50K	0.4798119	0.5201881
	50-80K	0.2871025	0.7128975
	GT 80K	0.2058824	0.7941176

- Conditional probability between income and education it is the probability of education given the income:

	Education			
Income		College	Others	Prof/Phd
	10-50K	1778	5350	104
	50-80K	561	501	70
	GT 80K	348	191	107

	Education			
Income		College	Others	Prof/Phd
	10-50K	0.24585177	0.73976770	0.01438053
	50-80K	0.49558304	0.44257951	0.06183746
	GT 80K	0.16563467	0.29566563	0.16563467

Use naive bayes classifier on the training data then predict the output for testing data compare it with the actual one and create the model's confusion matrix:

	Predicted			
Actual		10-50K	50-80K	GT 80K
	10-50K	787	0	6
	50-80K	127	0	5
	GT 80K	67	0	8

The model predict (127+67+6+5 =205) wrong in the testing data, so the accuracy of this model = 79.5%

Accuracy = 0.795

2. Part 2 (Predict individual's gender)

The model's priori: first we count the number of each option in the gender column then we divide it over the count of all options to get probabilities, the sum of all probabilities must be equal to 1.

F	M
3928	5082

F	M
0.43596	0.56404

The model's conditional probabilities: we calculate the conditional probability for each input feature (age, income, educ) by get the count then divide it over the row count to get the ratio. The sum of each row must be equal to 1.

- Conditional probability between gender and age it is the probability of age given the gender:

		Age		
Gender		20-30	31-45	GT 45
	F	708	1365	1855
	M	934	1797	2351

		Age		
Gender		20-30	31-45	GT 45
	F	0.1802444	0.3475051	0.4722505
	M	0.1837859	0.3536009	0.4626131

- Conditional probability between gender and income it is the probability of income given the gender:

		Income		
Gender		10-50K	50-80K	GT 80K
	F	3470	325	133
	M	3762	807	513

		Income		
Gender		10-50K	50-80K	GT 80K
	F	0.88340122	0.08273931	0.03385947
	M	0.74025974	0.15879575	0.10094451

- Conditional probability between gender and education it is the probability of education given the gender:

	Education			
Gender		College	Others	Prof/Phd
	F	1262	2581	85
	M	1425	3461	196

	Education			
Gender		College	Others	Prof/Phd
	F	0.32128310	0.65707739	0.02163951
	M	0.28040142x	0.68103109	0.03856749

Use naive bayes classifier on the training data then predict the output for testing data compare it with the actual one and create the model's confusion matrix:

	Predicted		
Actual		F	M
	F	106	321
	M	97	476

The model predict(97+321 =418) wrong in the testing data, so the accuracy of this model = 58.2%

Accuracy = 0.582

The accuracy doesn't improve. Still getting a large number of wrong prediction.

3. Part 3 (Balance your data):

As I divide the training data into male and female select randomly 3500 record from each of them ad combine them in one new training data as 7000 record.

The model's priori: first we count the number of each option in the gender column then we divide it over the count of all options to get probabilities, the sum of all probabilities must be equal to 1.

F	M
3500	3500

F	M
0.5	0.5

The model's conditional probabilities: we calculate the conditional probability for each input feature (age, income, educ) by get the count then divide it over the row count to get the ratio. The sum of each row must be equal to 1.

- Conditional probability between gender and age it is the probability of age given the gender:

		Age		
Gender		20-30	31-45	GT 45
	F	622	1217	1661
	M	650	1236	1614

		Age		
Gender		20-30	31-45	GT 45
	F	0.1777143	0.3477143	0.4745714
	M	0.1777143	0.3477143	0.4745714

- Conditional probability between gender and income it is the probability of income given the gender:

		Income		
Gender		10-50K	50-80K	GT 80K
	F	3099	285	116
	M	2593	561	346

		Income		
Gender		10-50K	50-80K	GT 80K
	F	0.88542857	0.08142857	0.03314286
	M	0.74085714	0.16028571	0.09885714

- Conditional probability between gender and education it is the probability of education given the gender:

	Education			
Gender		College	Others	Prof/Phd
	F	1117	2310	73
	M	960	2393	147

	Education			
Gender		College	Others	Prof/Phd
	F	0.02085714	0.66000000	0.02085714
	M	0.04200000	0.68371429	0.04200000

Use naive bayes classifier on the training data then predict the output for testing data compare it with the actual one and create the model's confusion matrix:

	Predicted		
Actual		F	M
	F	369	58
	M	412	161

The model doesn't classify the testing data so well.

Repeating random selecting of the record doesn't improve the model's performance. The conditional probabilities change every time but we still training the model on the same number of male and female records so it doesn't affect the accuracy of the model.

The conclusions from this whole task is that each input feature has a different effect on the result of model performance and large training data is better than small training data.