Product Selection

1. Output of kNN Classifier

a. Label of each new customer as predicted by kNN (k = 3):

Type	Life-Style	Vacation	eCrecdit	Salary	Property value	Label
student	spend<saving	12	19	14.79	3.7697	C1
student	spend>>saving	29	10	16.19	2.4839	C1
student	spend<<saving	28	60	15.46	1.1885	C1
Engineer	spend>saving	15	41	21.26	1.4379	C1
Librarian	spend<saving	2	9	19.7207	0.6913	C2
Librarian	spend>saving	7	9	12.7098	1.4728	C2
Professor	spend>saving	5	10	20.883	1.3131	C2
Professor	spend<saving	3	15	16.5711	0.4792	C2
student	spend<saving	9	71	25.7	2.0947	C3
student	spend>saving	10	67	27.11	3.8391	C3
Doctor	spend>saving	7	229	30.61	7.0074	C3
Doctor	spend<saving	8	243	25.33	8.7276	C3
Professor	spend>saving	51	5	18.98	2.8944	C5
Doctor	spend>saving	34	51	19.9	3.9544	C4
student	spend>>saving	39	40	19.3	3.8317	C4
student	spend<<saving	36	57	19.61	4.888	C4
Professor	spend>>saving	34	30	20.91	2.4095	C4
Librarian	spend<<saving	48	35	20.15	2.4436	C4
Professor	spend>>saving	52	5	22.63	2.2115	C5
Engineer	spend>saving	50	17	32.59	1.2229	C5
Engineer	spend>>saving	50	15	21.78	2.0736	C5

b. The KNN java code is enclosed in the task deliverable folder.

c. Cross validation results:

```
------Testing the Product Selection: ----------
For 0th test, the correctness is 0.8888888888888888
For 1th test, the correctness is 0.8888888888888888
For 2th test, the correctness is 0.8888888888888888
For 3th test, the correctness is 1.0
For 4th test, the correctness is 0.8888888888888888
For 5th test, the correctness is 0.9444444444444444
For 6th test, the correctness is 0.9444444444444444
For 7th test, the correctness is 0.9444444444444444
For 8th test, the correctness is 0.9444444444444444
For 9th test, the correctness is 0.8333333333333334
Testing through cross validation, the accuracy is: 0.9166666666666667
```

2. **Output of Decision Tree:**
   a. Product class label as predicted by the decision tree classifier for each new customer:

Type	Life-Style	Vacation	eCredit	Salary	Property value	Label
Student	spend<saving	12	19	14.79	3.7697	C1
Student	spend>>saving	29	10	16.19	2.4839	C1
Student	spend<<saving	28	60	15.46	1.1885	C1
Engineer	spend>saving	15	41	21.26	1.4379	C2
Librarian	spend<saving	2	9	19.7207	0.6913	C2
Librarian	spend>saving	7	9	12.7098	1.4728	C2
Professor	spend>saving	5	10	20.883	1.3131	C2
Professor	spend<saving	3	15	16.5711	0.4792	C2
Student	spend<saving	9	71	25.7	2.0947	C3
Student	spend>saving	10	67	27.11	3.8391	C3
Doctor	spend>saving	7	229	30.61	7.0074	C3
Doctor	spend<saving	8	243	25.33	8.7276	C3
Professor	spend>saving	51	5	18.98	2.8944	C5
Doctor	spend>saving	34	51	19.9	3.9544	C4
Student	spend>>saving	39	40	19.3	3.8317	C4
Student	spend<<saving	36	57	19.61	4.888	C4
Professor	spend>>saving	34	30	20.91	2.4095	C4
Librarian	spend<<saving	48	35	20.15	2.4436	C4
Professor	spend>>saving	52	5	22.63	2.2115	C5
Engineer	spend>saving	50	17	32.59	1.2229	C5
Engineer	spend>>saving	50	15	21.78	2.0736	C5

   b. Decision tree model built by WEKA using its default parameter settings, and the prediction error for existing customers reported by WEKA:

J48 pruned tree

Vacation <= 15
| property <= 1.9857: C2 (26.0)
| property > 1.9857
| | salary <= 20.65: C1 (14.0/2.0)
| | salary > 20.65: C3 (39.0)
Vacation > 15
| eCredit <= 26
| | Vacation <= 46
| | | Type = student: C1 (3.0)
| | | Type = engineer: C1 (1.0)
| | | Type = librarian: C4 (3.0)
| | | Type = professor: C4 (1.0)

```
|   |   |   Type = doctor: C1 (0.0)
|   |   Vacation > 46: C5 (37.0/1.0)
|   eCredit > 26
|   |   Type = student
|   |   |   salary <= 17.6074: C1 (5.0)
|   |   |   salary > 17.6074: C4 (8.0)
|   |   Type = engineer
|   |   |   LifeStyle = spend<<saving
|   |   |   |   salary <= 20.17: C1 (3.0)
|   |   |   |   salary > 20.17
|   |   |   |   |   property <= 3.2804: C1 (2.0)
|   |   |   |   |   property > 3.2804: C4 (5.0/1.0)
|   |   |   LifeStyle = spend<saving: C4 (2.0)
|   |   |   LifeStyle = spend>saving
|   |   |   |   Vacation <= 41: C1 (6.0)
|   |   |   |   Vacation > 41: C4 (2.0)
|   |   |   LifeStyle = spend>>saving
|   |   |   |   property <= 5.2348: C4 (5.0/1.0)
|   |   |   |   property > 5.2348: C1 (2.0)
|   |   Type = librarian: C4 (6.0)
|   |   Type = professor: C4 (7.0)
|   |   Type = doctor: C4 (9.0)

Number of Leaves  :         22

Size of the tree :    35


=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         157               84.4086 %
Incorrectly Classified Instances        29               15.5914 %
```

Prediction error as reported by Weka in the default setting is 15.5914%.

c. After tuning the parameter *ConfidenceFactor* :

We tuned the Confidence Factor from the range of 0.1 to 1.0 by an increment of 0.05.

Confidence Factor	Correctly Classified Instances (%)
0.05	80.6452
0.10	81.7204
0.15	82.2581
0.20	82.7957
0.25	84.4086
0.30	84.4086
0.35	84.4086
0.40	**84.9462**
0.45	84.9462
0.50	84.9462
0.55	84.4086
0.60	84.4086
0.65	84.4086
0.70	84.4086
0.75	84.4086
0.80	84.4086
0.85	84.4086
0.90	84.4086
0.95	84.4086
1.00	84.4086

From the above table it can be observed that accuracy of the J48 classifier increased with an increase in confidence factor from 0.05 to 0.50. After 0.50 the accuracy became constant.
As it can be seen from the above table that accuracy is maximum at 0.40. So we further checked the accuracy for range of confidence factor between 0.35 and 0.40. We observed that the accuracy changed from 84.4086% to 84.9462 at a confidence factor of 0.37. We further checked the accuracy in the interval of 0.36 and 0.37. The accuracy changed at a confidence factor of 0.367. Thus we choose our new confidence factor to be **0.367**.

We tested the training data for several values of confidence factor to find the most appropriate value for our particular training set and chose **0.367** to be the new confidence factor as the prediction error observed at this confidence factor was the lowest.

d. Cross-validation results:

At a confidence factor of 0.367 the cross validation results as observed are:

Correctly Classified Instances	158	84.9462 %
Incorrectly Classified Instances	28	15.0538 %
Kappa statistic	0.8102	
Mean absolute error	0.0622	
Root mean squared error	0.224	

Relative absolute error 19.5988 %
Root relative squared error 56.2237 %
Total Number of Instances 186

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.639	0.08	0.657	0.639	0.648	0.842	C1
	0.962	0.013	0.926	0.962	0.943	0.975	C2
	0.951	0.007	0.975	0.951	0.963	0.982	C3
	0.787	0.079	0.771	0.787	0.779	0.902	C4
	0.944	0.013	0.944	0.944	0.944	0.975	C5
Weighted Avg.	0.849	0.041	0.849	0.849	0.849	0.932	

=== Confusion Matrix ===

a	b	c	d	e	<-- classified as
23	2	1	9	1	a = C1
1	25	0	0	0	b = C2
2	0	39	0	0	c = C3
9	0	0	37	1	d = C4
0	0	0	2	34	e = C5