DS707 Data Analytics
Project Report

SSLC Data Analysis

Masters of Technology
in
Information Technology

Submitted by

Roll No	Names of Students
MT2013025	Apoorwa Mishra
MT2013026	Arjun S Bharadwaj
MT2013140	Ankit Shah

Under the guidance of
Prof. Chandrashekar R

International Institute of Information Technology
Bangalore, Karnataka, India – 560 100

Fall Semester 2014

Contents

Experiment 1

Discretizaion + Classification

Consider the marks information:
L1_MARKS, L2_MARKS, L3_MARKS, S1_MARKS, S2_MARKS, S3_MARKS.
Consider TOTAL_MARKS as the dependent variable

- Objective:
  - Discretize subject marks into discrete attributes S (use NRC_CLASS as domain)
  - Build a classification model based on S for NRC_CLASS class variable
- Procedure followed:
  - Step 1
  - Step 2
- Results Obtained:
  - Result 1
  - Result 2
- Conclusion:
  - Conclusion 1
  - Conclusion 2

Experiment 2

Regression + Classification

Consider the marks information:
L1_MARKS, L2_MARKS, L3_MARKS, S1_MARKS, S2_MARKS, S3_MARKS.
Consider TOTAL_MARKS as the dependent variable

- Objective:
  - Determine the least number of attributes S that give the best possible regression equation (least error)
  - Build a classification model based on S for NRC_CLASS class variable
- Procedure followed:
  - Data is loaded and cleansed by removing invalid and missing rows.
  - Regression analysis is performed on the data by using the marks data.
  - Synergy/Interaction effect of all the marks are obsereved and the combination of marks having least p-value is chosen for classification.
  - Marks are rounded off for improving the speed of classification.
  - Classification is performed on the data based on the class variable combination having least p-value.
  - The results of confusion matrices are compared.
- Results Obtained:
  - All Subjects are used for classification:
    - Accuracy: 90.2%

    - 95% CI: (0.8962, 0.9076)
  - Best case:
    L1_MARKS, L2_MARKS, S2_MARKS, S3_MARKS (p-value = 0.0732) are used for classification:
    - Accuracy: 84.03%
    - 95% CI: (0.8332, 0.8472)
  - Worst case:
    L3_MARKS, S1_MARKS (p-value = 0.94523) are used for classification:
    - Accuracy: 69.46%
    - 95% CI: (0.6857, 0.7033)

- Conclusion:
  - Taking all the subjects marks for classification gives the highest accuracy.
  - Taking the combination of the subjects having low p-value offers the next highest accuracy for classification.
  - Conversely, the combination of subjects having highest p-value gives the least accuracy.

Experiment 3

Regression + Classification

Consider the marks information:
L1_MARKS, L2_MARKS, L3_MARKS, S1_MARKS, S2_MARKS, S3_MARKS.
Consider TOTAL_MARKS as the dependent variable

- Objective:
  - Determine the least number of attributes S that give the best possible regression equation (least error)
  - Build a classification model based on S for NRC_CLASS class variable
- Procedure followed:
  - Step 1
  - Step 2
- Results Obtained:
  - Result 1
  - Result 2
- Conclusion:
  - Conclusion 1
  - Conclusion 2

Experiment 4

Regression + Classification

Consider the marks information:
L1_MARKS, L2_MARKS, L3_MARKS, S1_MARKS, S2_MARKS, S3_MARKS.
Consider TOTAL_MARKS as the dependent variable

- Objective:
  - Determine the least number of attributes S that give the best possible regression equation (least error)
  - Build a classification model based on S for NRC_CLASS class variable
- Procedure followed:
  - Step 1
  - Step 2
- Results Obtained:
  - Result 1
  - Result 2
- Conclusion:
  - Conclusion 1
  - Conclusion 2

Experiment 5

Urban / Rural Characterization

What are the characteristics of students from urban and rural areas, respectively? For antecedent, try with demographic info (SCHOOL_TYPE, URBAN_RURAL, NRC_CASTE_CODE, NRC_GENDER_CODE, NRC_MEDIUM, NRC_PHYSICAL_CONDITION, CANDIDATE_TYPE) Also try with subject performance in the antecedent

- Objective:
  Identify association rules with URBAN_RURAL in the consequent of the rule
- Procedure followed:
  - Data is loaded and cleansed by removing invalid and missing rows.
  - The values of all the columns are factored so that it's suitable for association rules analysis.
  - Apriori algorithm is run on the data by forcing URBAN_RURAL=Rural rule in consequent.
  - Apriori algorithm is run on the data by forcing URBAN_RURAL=Urban rule in consequent.
  - The rules generated with high confidence and lift are compared for both the cases.
- Results Obtained:
  For URBAN_RURAL = Urban, the following were the results:

Antecedant	Support	Confidence	Lift
SCHOOL_TYPE = Unaided, NRC_MEDIUM = English	0.1305375	0.8029823	1.883977
SCHOOL_TYPE = Unaided, NRC_MEDIUM = English, NRC_PHYSICAL_CONDITION=Normal	0.1297800	0.8025108	1.882871
SCHOOL_TYPE = Unaided, NRC_MEDIUM = English, NRC_CASTE_CODE=General	0.1130235	0.8002575	1.877584

For URBAN_RURAL = Rural, the following were the top three results:

Antecedant	Support	Confidence	Lift
SCHOOL_TYPE = Government, NRC_GENDER_CODE=Boy, NRC_MEDIUM = Kannada, CANDIDATE_TYPE=Regular Fresher, L1_RESULT=Pass, L2_RESULT=Pass, S2_RESULT=Pass, S3_RESULT=Pass	0.1018423	0.8611325	1.500797
SCHOOL_TYPE = Government, NRC_GENDER_CODE = Boy, NRC_MEDIUM = Kannada, CANDIDATE_TYPE = Regular Fresher, L1_RESULT = Pass, L2_RESULT = Pass, L3_RESULT = Pass, S1_RESULT = Pass, S3_RESULT = Pass	0.1015393	0.8609969	1.500561
SCHOOL_TYPE = Government, NRC_GENDER_CODE = Boy, NRC_MEDIUM = Kannada, CANDIDATE_TYPE = Regular Fresher, L1_RESULT = Pass, L2_RESULT = Pass, L3_RESULT = Pass, S1_RESULT = Pass, S2_RESULT = Pass	0.1012969	0.8609323	1.500448

- Conclusion:
  - Students in Urban area mainly belong to Unaided English medium schools.
  - Students in Rural area are mainly boys who belong to Government Kannada medium schools.

Experiment 6

Regression + Classification

Consider the marks information:
L1_MARKS, L2_MARKS, L3_MARKS, S1_MARKS, S2_MARKS, S3_MARKS.
Consider TOTAL_MARKS as the dependent variable

- Objective:
  - Determine the least number of attributes S that give the best possible regression equation (least error)
  - Build a classification model based on S for NRC_CLASS class variable
- Procedure followed:
  - Step 1
  - Step 2
- Results Obtained:
  - Result 1
  - Result 2
- Conclusion:
  - Conclusion 1
  - Conclusion 2

Experiment 7

Regression + Classification

Consider the marks information:
L1_MARKS, L2_MARKS, L3_MARKS, S1_MARKS, S2_MARKS, S3_MARKS.
Consider TOTAL_MARKS as the dependent variable

- Objective:
  - Determine the least number of attributes S that give the best possible regression equation (least error)
  - Build a classification model based on S for NRC_CLASS class variable
- Procedure followed:
  - Step 1
  - Step 2
- Results Obtained:
  - Result 1
  - Result 2
- Conclusion:
  - Conclusion 1
  - Conclusion 2

Experiment 8

Regression + Classification

Consider the marks information:
L1_MARKS, L2_MARKS, L3_MARKS, S1_MARKS, S2_MARKS, S3_MARKS.
Consider TOTAL_MARKS as the dependent variable

- Objective:
  - Determine the least number of attributes S that give the best possible regression equation (least error)
  - Build a classification model based on S for NRC_CLASS class variable
- Procedure followed:
  - Step 1
  - Step 2
- Results Obtained:
  - Result 1
  - Result 2
- Conclusion:
  - Conclusion 1
  - Conclusion 2