

**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

|   |                                |   |
|---|--------------------------------|---|
| 1 | Discretizaion + Classification | 1 |
| 2 | Regression + Classification    | 2 |
| 3 | Regression + Classification    | 4 |
| 4 | Regression + Classification    | 5 |
| 5 | Regression + Classification    | 6 |
| 6 | Regression + Classification    | 7 |
| 7 | Regression + Classification    | 8 |
| 8 | Regression + Classification    | 9 |

# 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 observed 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

## 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 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