Anomaly Detection Challenge

CHALLENGE 1:

Cotton Soil Classification in Satelite Images

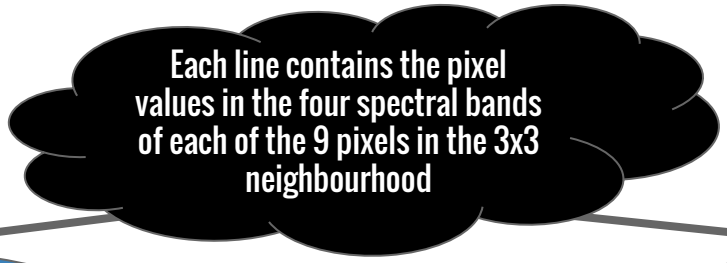
Team: Abracadata
Ishmeet Kaur(03677735)
Mustika Rizki Fitriyanti(03667399)

1.

Introduction about the Data

I. Introduction

- The data that is a sub-area of a scene that consist with 82 x 100 pixels.



Training Data

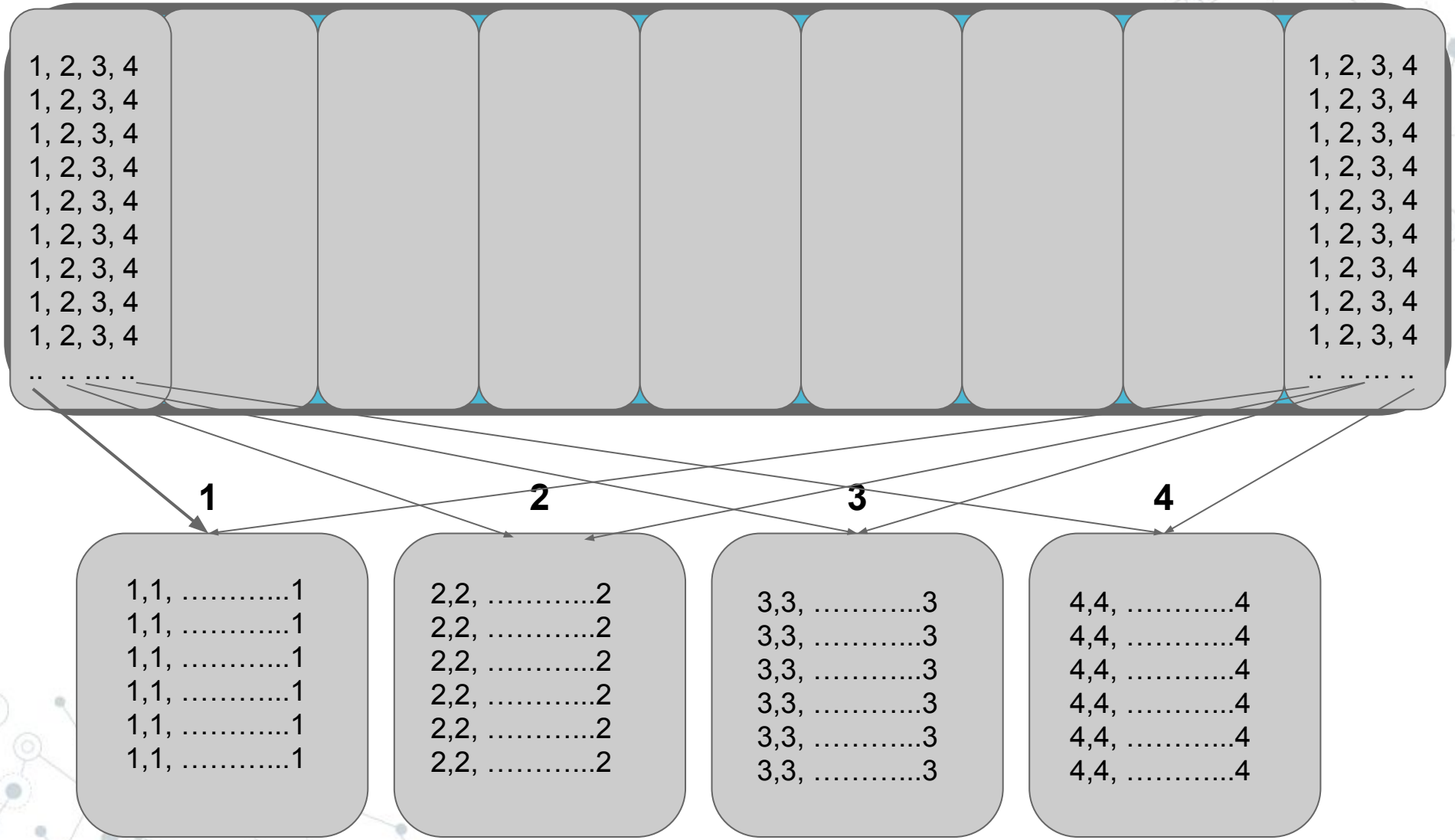
Number Features	36
Number Rows	4435
Classify as Cotton Soil (Y=1)	479
Classify as Normal Soil (Y=0)	3956

Testing Data

Number Features	36
Number Rows	2000

2.
Data Preprocessing

2.1 Arranging the Data



2.2 Handling Missing Values

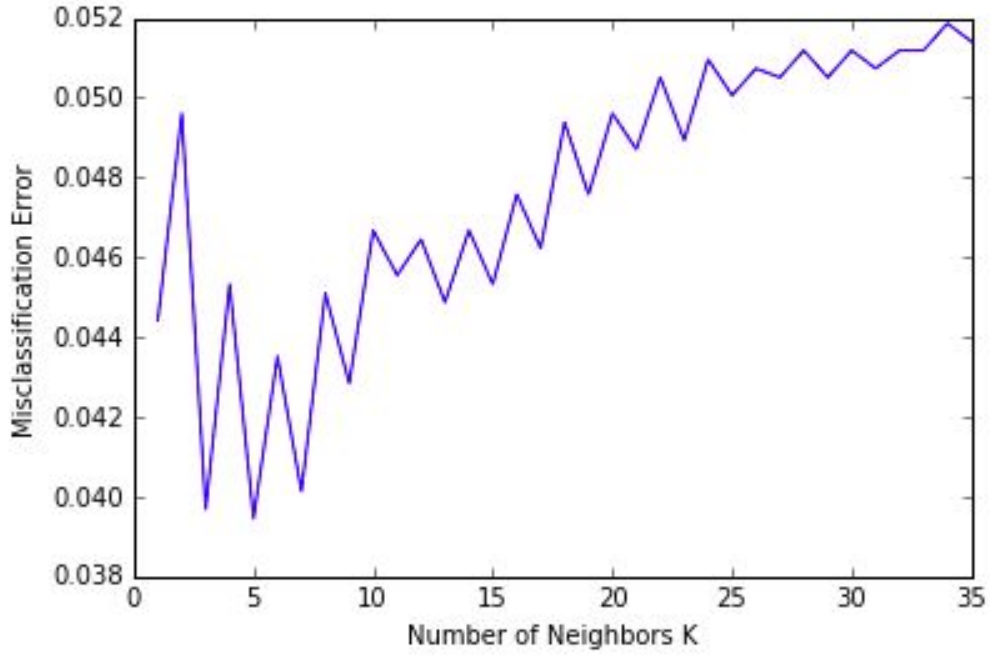
Replacement Method	Description
Mean	Replacement of the missing values was done both row and column wise.
Median	The missing values were replaced by the median value both along the row and column.
Mode	The missing values were replaced by the most frequent occurring element in both row and column wise.
Min/Max	Minimum and maximum values for both row and column were used to replace missing values.
Interpolation	The missing values were estimated by interpolation.
Per Spectral Value	The mean , median and mode values for every spectral of the image were used to replace the missing values per spectral.

3. Creation and Evaluation of the Models

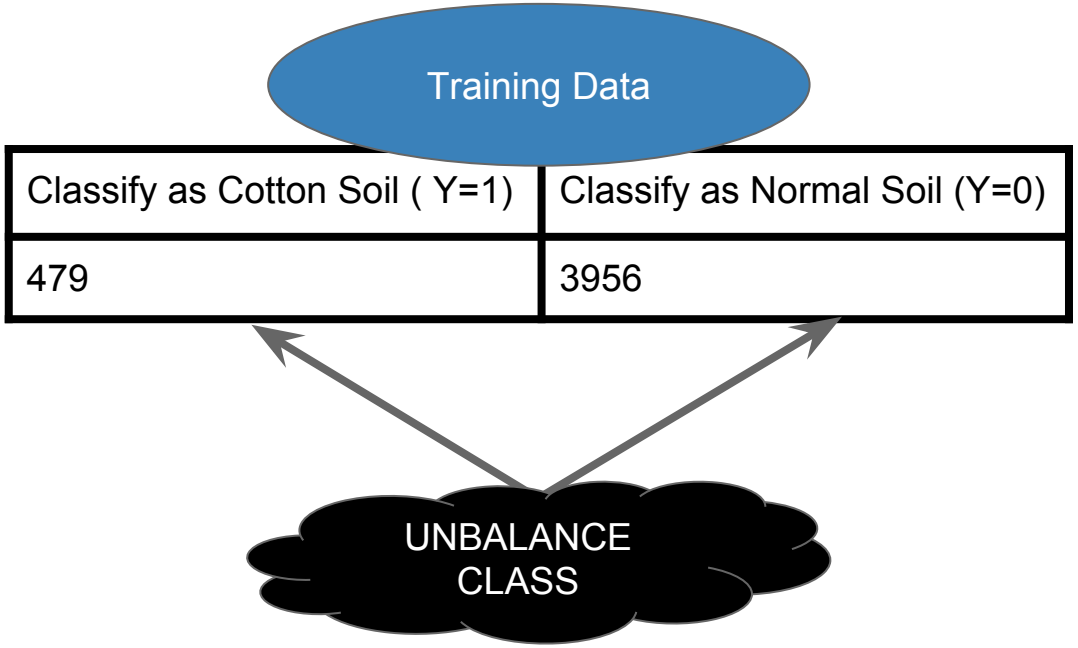
Naive Bayes
Decision Trees
K-Nearest Neighbours
LogisticRegression
Random Forest

3.2 Result 1: Implementation of Classification Algorithms

Calculating the misclassification error, we found that the best accuracy for K-Nearest Neighbors could be achieved when the k=5.

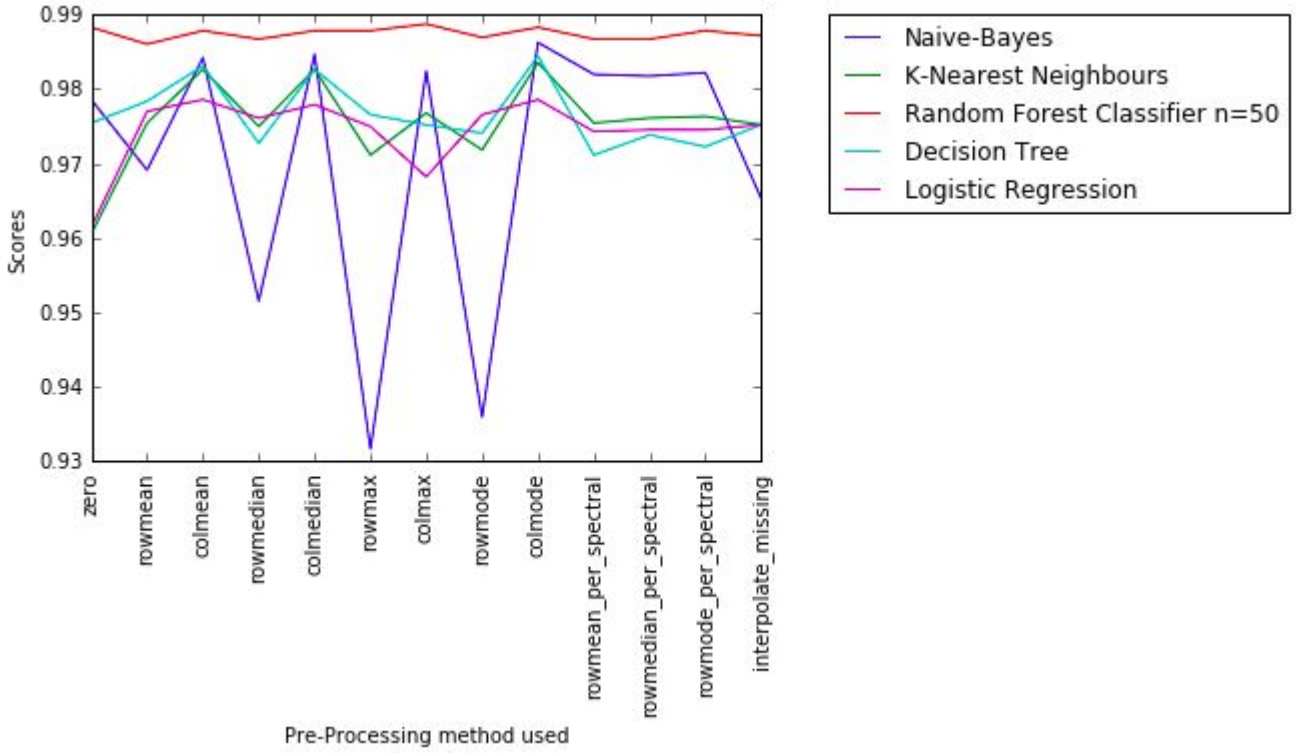


3.1 Using Stratified K-Cross Validation



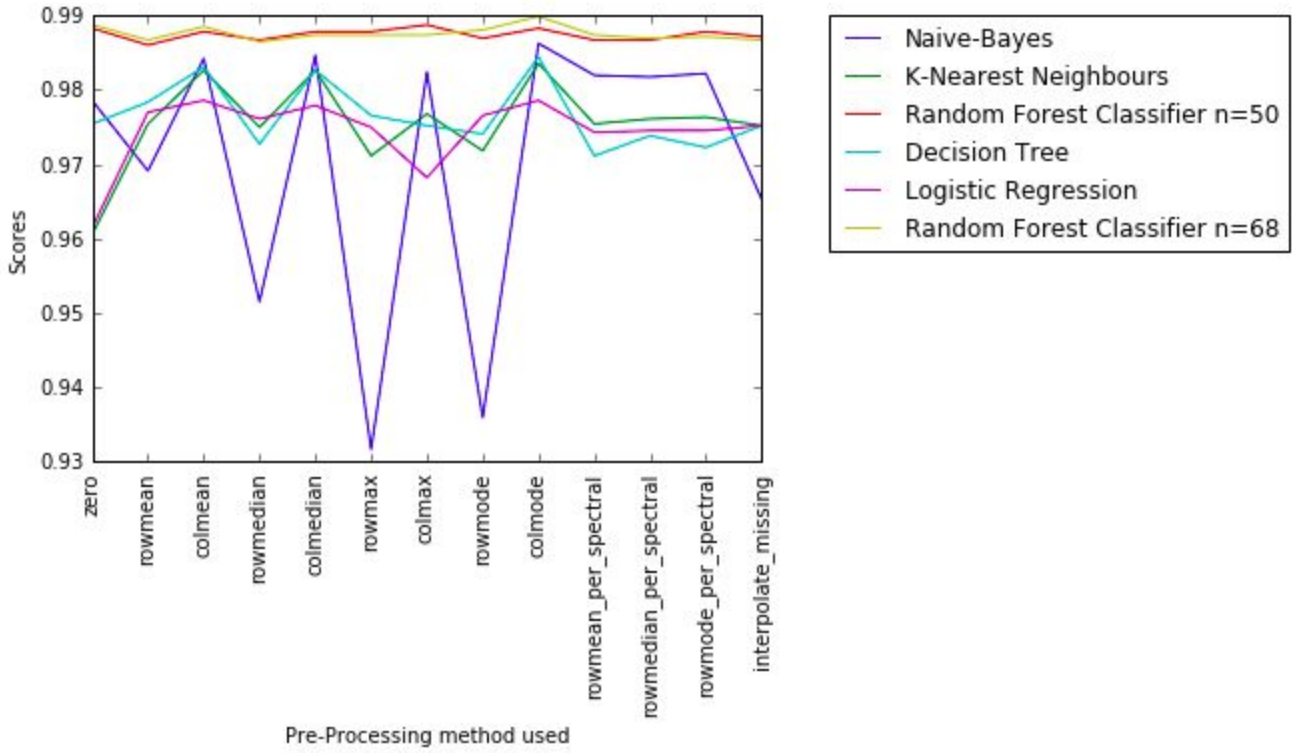
Stratification will ensure that the percentages of each class in the entire data will be the same or close within each individual fold.

3.2 Result 1: Implementation of Classification Algorithms



4.
Result Summary

4.1 Final Accuracy Comparison



4.2 Kaggle Score

Method	Public Score	Private Score
Random Forest colmode n=68	0.98121	0.98953
Random Forest colmedian n=68	0.97693	0.99421
Random Forest colmean n=50	0.97693	0.99009