**Stellar Classification**

**Milestone: Data Collection, Data Visualization, Data Exploration and Data Processing**

Group 16

Student 1 Sanidhya Karnik

Student 2 Digvijay Raut

617-407-1206 (Tel of Student 1)

857-492-3195 (Tel of Student 2)

[karnik.san@northeastern.edu](mailto:karnik.san@northeastern.edu)

[raut.di@northeastern.edu](mailto:raut.di@northeastern.edu)

**Percentage of Effort Contributed by Student 1: 50%**

**Percentage of Effort Contributed by Student 2: 50%**

**Signature of Student 1: Sanidhya Karnik**

**Signature of Student 2: Digvijay Raut**

**Submission Date: 02/16/2024**

**Data Collection:**

The dataset originates from the Sloan Digital Sky Survey (SDSS), comprising 100,000 entries. Each entry is characterized by 17 features and 1 target class, culminating in a dataset with 100,000 rows and no missing values. Below are the statistics of the raw data:

A screenshot of a computer

Description automatically generated

**Data Reduction:**

Upon analysis, we determined that the columns 'obj\_ID', 'run\_ID', 'rerun\_ID', 'cam\_col', and 'spec\_obj\_ID' do not influence the target class in a meaningful scientific manner. Notably, 'rerun\_ID' remains constant across the dataset, rendering its presence in the model redundant. By excluding these columns, we are left with 13 relevant columns for our analysis.

**Data Exploration:**

The correlation matrix serves as a pivotal tool for understanding the interdependencies among variables, thereby guiding data interpretation and decision-making. Through this analysis, we discovered a correlation among the 'u', 'g', 'r', 'i', 'z' columns.

A colorful squares with numbers

Description automatically generated

Fig 1

A blue dot diagram with numbers

Description automatically generated with medium confidenceA blue blot of dots

Description automatically generated

A blue dotted line graph

Description automatically generatedA blue dotted graph with numbers

Description automatically generated

Fig 2

In Fig 2, we visualized scatter plots between ‘g’ vs ‘u’, ‘g’ vs ‘z’ and ‘i’ vs ‘r’ which showed a strong positive correlation whereas ‘readshift’ vs ‘r’ showed a weak positive correlation.

A subsequent box-plot analysis revealed an outlier with an unusually low value (refer Fig 3a), leading to its removal. Fib 3b shows a better readable boxplot after the outlier was removed. Further outlier removal was conducted using sklearn.neighbours library, resulting in a refined dataset of 89,999 records.

A white rectangular object with black lines

Description automatically generatedA diagram of different colored rectangles

Description automatically generated

Fig 3 (a, b)

Given the presence of three target variables, we employed a bar graph to visualize their distribution, identifying a predominant class among them.

A graph with different colored squares

Description automatically generated

Fig 4

A Principal Component Analysis (PCA) was executed on the dataset and the first 7 components that encapsulate ~95% of variance, were picked to be the principal components. This process effectively reduced the dataset from 17 features to 7 and from 100,000 records to 89,999.

Variance distribution: [49.62369147 12.06979656 9.07481334 8.22269234 7.51893126 7.07511107 6.41496395]

Variance captured: 95.22%

Statistics of the final dataset acquired after performing Principal component analysis can be seen in Fig 5 below:

A screenshot of a computer

Description automatically generated

Fig 5

In the final stage, we examined a correlation matrix of the principal components, revealing negligible correlation among them, indicating successful dimensionality reduction.

A screenshot of a graph

Description automatically generated

Fig 6