IDS Project : Students' Dataset

Raksha Ramesh
PES1201800345
Computer Science Engineering
PES University

Shreya Prabhu
PES1201800128
Computer Science Engineering
PES University

Tanishq Vyas
PES1201800125
Computer Science Engineering
PES University

I. Abstract

This is a project to extract meaning and analyze a dataset containing student grades, to come up with a hypothesis and to test if the hypothesis made, stands true or false.

The chosen dataset consists mainly of students' grades, with additional parameters like their travel time, family size, occupations of their parents, etc. as different columns.

II. Steps Involved

A. Adding NaN values at random

The dataset obtained was initially clean, so NaNs were introduced with varying percentages for majority of the columns. The chosen percentages for the introduction of the NaNs were: 3%, 1.5%, and 1%.

B. Replacing NaN

After the introduction of NaNs, the dataset now had to be cleaned. Separate lists were made of the columns which needed the NaNs to be replaced by mean, median and previous values respectively. The NaNs were then replaced with the respective form, depending on the kind of data.

C. Standardization and Normalization

The marks fields were standardized and replaced with new values. The data fields were normalized to make mean 0 and variance 1. The copy of each state of dataset was maintained.

D. Exploratory Analysis

The following graphs were plotted in order to gain insights

1) Box-Plot: For G1, G2, G3 marks. The graph shows that G3 distribution is more normally distributed than G1 and G2. Also there are a few left side outliers in G3 distribution and the G1 and G2 distributions are slightly left skewed.

2) Scatter-Plot: Marks in G3 vs Travel time in hours. The graph shows that people who score higher have a relatively higher travel time than who score less. This hints that these two columns may be related but the Pearson's coefficient turned out to be -0.130768, thus stating the visualization is faulty.

3) Pie-chart: For male vs female student distribution. This shows that there are 59% female students and only 41% male students.

4) Histogram : For G3 distribution. This shows us that G3 is normally distributed with some outliers on its left.

III. Hypothesis

With all the exploratory analysis and better understanding of the dataset, the following hypothesis was constructed.

H0 : average marks scored by female students are greater than or equal to that by male students i.e.
(Fmean - Mmean >= 0)

H1: The average marks scored by the female students are lesser than that of male students i.e.
(Fmean - Mmean <0)

Where Fmean and Mmean represent the average means for female and male students respectively.

IV. Findings

Our chosen significance level is 5% and our confidence level is 95%

The P value turned out to be 0.000529
Since P is found to be <<0.05, H0 can be rejected and H1 can be accepted.

V. Most and Least Correlated Columns

G2 and G3 are the most related columns, with Pearson's coefficient : 0.918548 whereas paid and G3 are least related columns, with Pearson's coefficient : -0.054898.

VI. Conclusion

Through this it can be inferred that the average marks scored by males was greater than that of females, 95% of the time.