February 27, 2021

Students Performance on Exams Data Set

Peter Massarello
CS 4300
University of Missouri-St. Louis

Contents

1 Introduction

I chose this specific data set originally because my last pick didn't feel like it would work best for what this project aims to help teach us. It was a data set on the end game of Tic-Tac-Toe and felt very limited in what I had to work with. But with this set, I feel as though there are many interesting parts I will get to graph and analyze.

This alone made it more interesting and worthwhile to spend my time on than the previous data set.

Google Colab:
`https://colab.research.google.com/drive/1GuOZw3IUxS1Kn5rlFa_aT28RvnypCsK9`

2 Data Set Description

The data set, "Students Performance in Exams" was obtained from Kaggle Data Science[1] This data set comes from the attempt to understand how certain factors like, parental background, test, preparation, lunch plan, etc have an effect on a students exam score. In this data set there are three output values, Math test score, Reading test score, and Writing test score. I decided to only focus on one of these outputs, that being the Math test score.

After cleaning the data set I changed much of the text data into integers in a range so that those values would be able to be graphed properly. The numerical values that replaced the input values range from 0-N, where N is the total number of inputs minus 1.
EX: Gender - Male, Female $\rightarrow 0, 1$

The total list of inputs are as follows:

1. Gender (Male or Female)
2. Race/Ethnicity (Group A-E)
3. Parental Level of Education (Some High School to Master's Degree)
4. Lunch (Free/Reduced or Standard)
5. Test Preparation (None or Completed)

2.1 Visualization of the Distribution of each input

Histogram plots of each input displaying the range of values each input possesses. (see **Figure 4**).

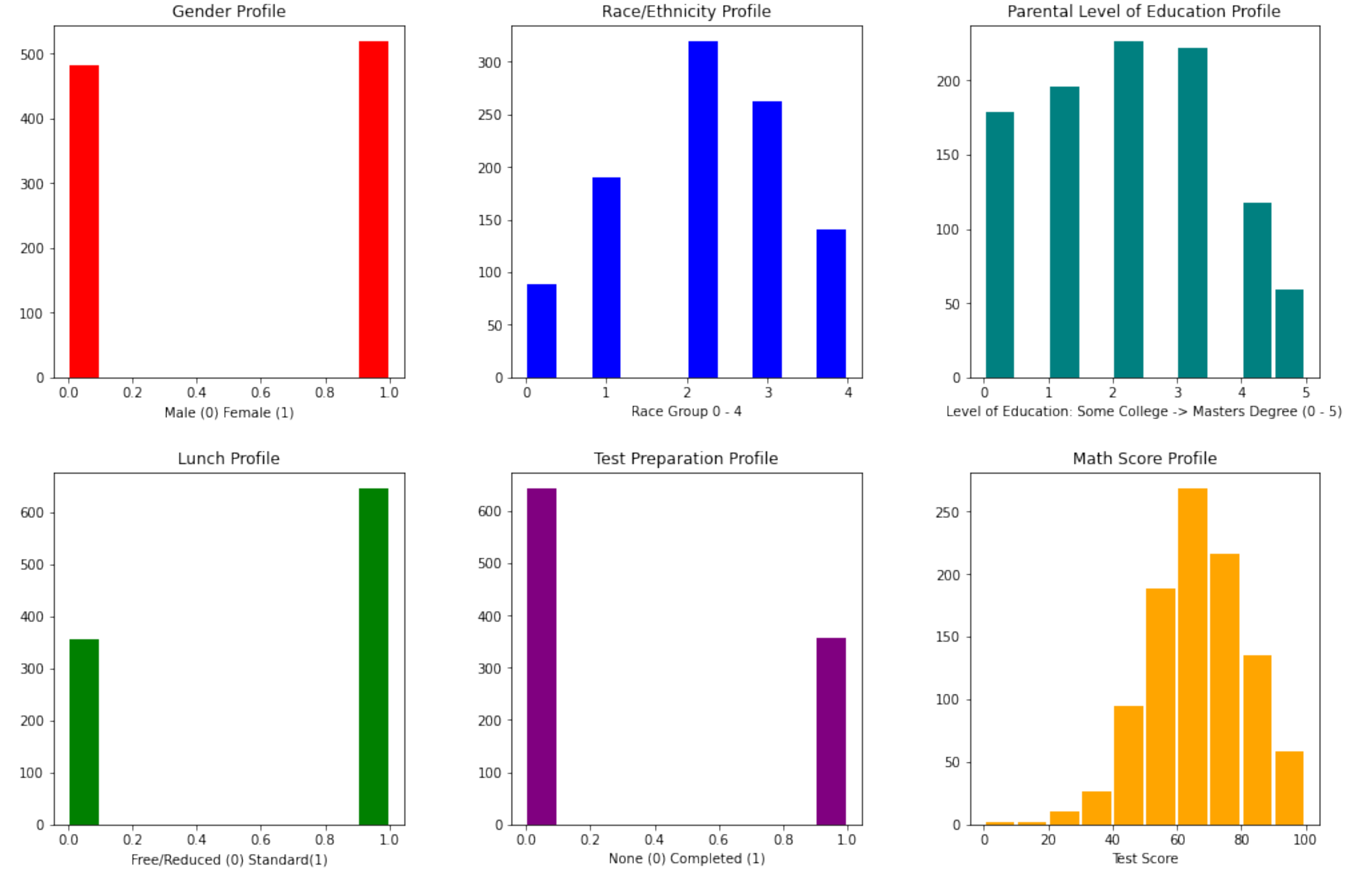
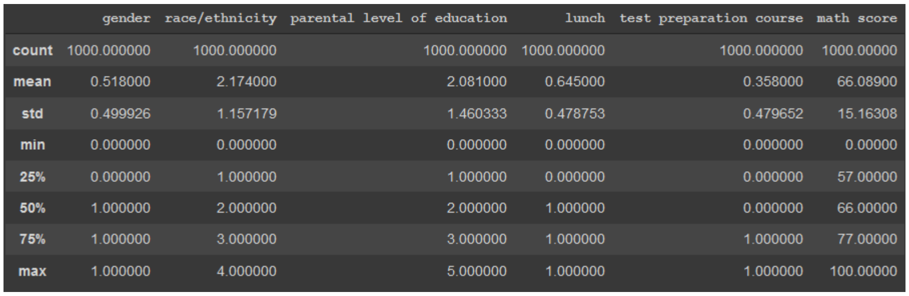


Figure 1: Histogram's of each Input feature and the output feature (Prior to normalization



	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score
count	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.00000
mean	0.518000	2.174000	2.081000	0.645000	0.358000	66.08900
std	0.499926	1.157179	1.460333	0.478753	0.479652	15.16308
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000
25%	0.000000	1.000000	1.000000	0.000000	0.000000	57.00000
50%	1.000000	2.000000	2.000000	1.000000	0.000000	66.00000
75%	1.000000	3.000000	3.000000	1.000000	1.000000	77.00000
max	1.000000	4.000000	5.000000	1.000000	1.000000	100.00000

Figure 2: Table of Values showing data points such as mean, min, max, etc

2.2 Distribution of the Output Labels

Overall most data was either pretty well divided or had a proper curve to it. Except for two graphs, Lunch Profile and Test Preparation Profile. These seemed to have a larger disparity between the binary values show below (see **Figure 4**). From what we can take from **Figure 2** we can determine that around 75 percent of Lunch profile is that of Standard lunch plan, while the remainder 25 percent is that of Free/Reduced. In Test Preparation we can see nearly the opposite. Where around 75 percent comes from the No preparation category and the remainder 25 percent comes from the Completed category

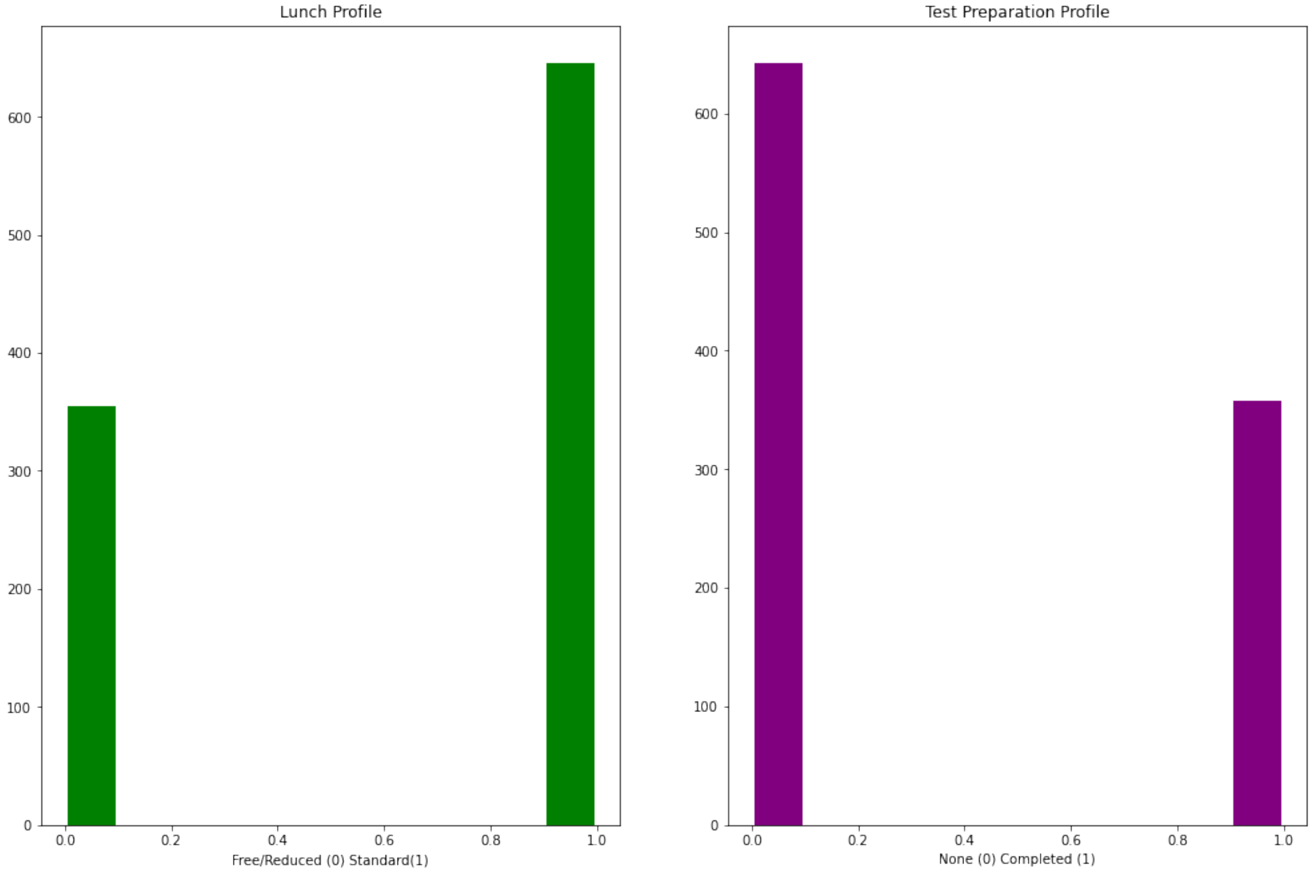


Figure 3: Sub-section of graphs from section 2.1

3 Data Processing

3.1 Data Normalization

Before we can begin data processing and learning, we have to perform some initial changes to the data. This is done through normalization techniques. Neural networks learn better when the numbers range from around 0 - 1. Therefore, for the input profiles which do not contain that range, we will normalize as to prep them for the later steps. The method I chose is Re-scaling Normalization. It goes as follows:

$$X_{new} = \frac{X}{X_{max}} \tag{1}$$

3.2 Normalized Data

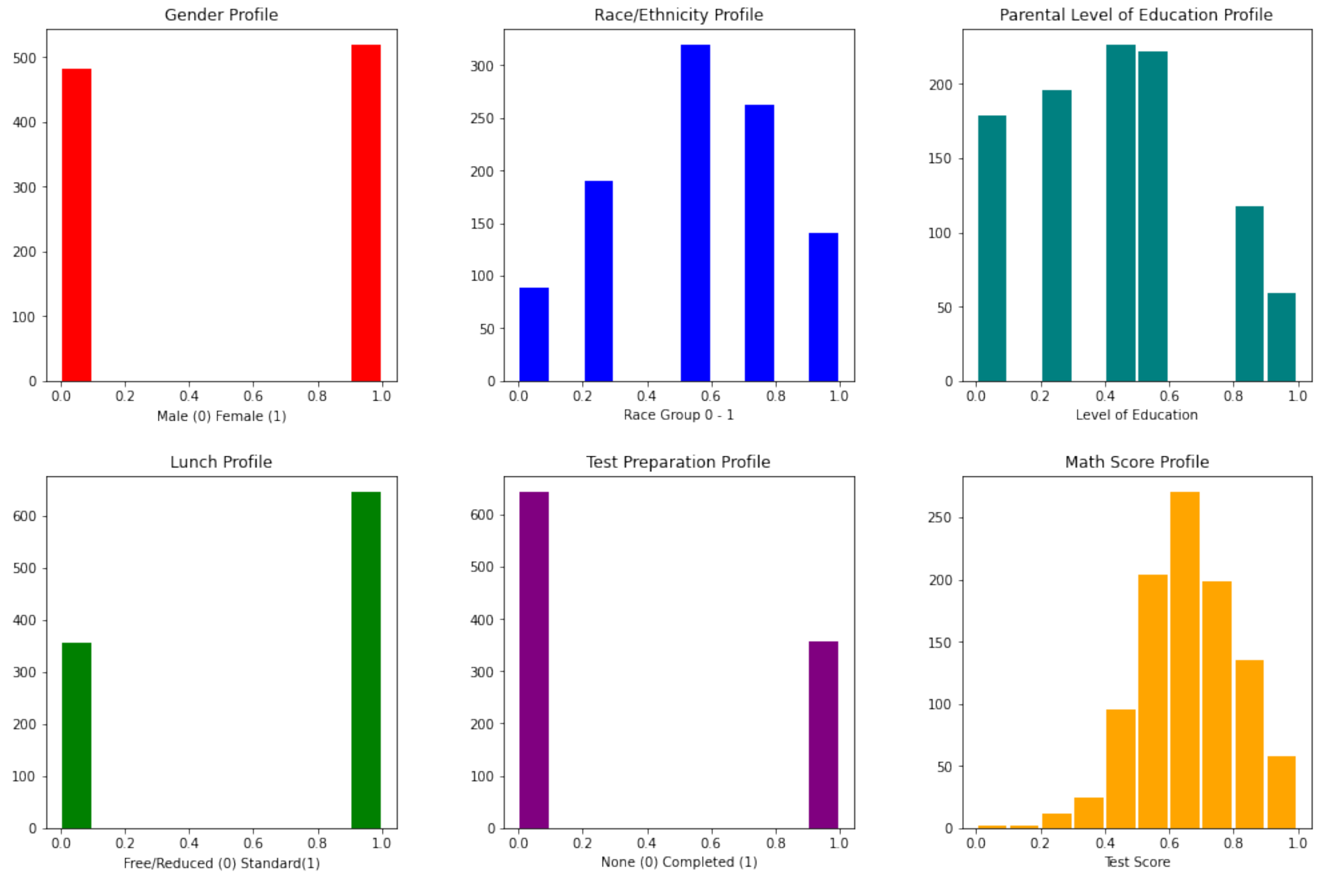


Figure 4: Input data after Normalization (Re-Scaling)

References

[1] Jakki Seshapanpu. Students Performance in Exams Marks secured by the students in various subjects. 2018.