***Machine Learning Assignment # 1***

***Maida Nadeem Butt MSDSF22M001***

***Dated: March 3rd, 2023 (Friday)***

***---------------------------------------------------------------------------------------***

***Data Description:***

Due to the growing need of educated and talented individuals, especially in developing countries, recruiting fresh graduates is a routine practice for organizations. Conventional recruiting methods and selection processes can be prone to errors and in order to optimize the whole process, some innovative methods are needed.

***Attributes’ Description:***

|  |  |
| --- | --- |
| *Column name* | *Description* |
| Sr# | Serial Number (Ordered Number sequence). (Int) |
| gender | The Gender of the candidate. (String) |
| ssc\_percentage | Senior secondary exams percentage (10th Grade). (Float) |
| ssc\_board | Board of education for ssc exams. (String) |
| hsc\_percentage | Higher secondary exams percentage (12th Grade). (Float) |
| hsc\_borad | Board of education for hsc exams. (String) |
| hsc\_subject | Subject of study for hsc .(String) |
| degree\_percentage | Percentage of marks in undergrad degree. (Float) |
| undergrad\_degree | Undergrad degree majors (String) |
| work\_experience | Past work experience. (Boolean) |
| emp\_test\_percentage | Aptitude test percentage. (Float) |
| specialization | Postgrad degree majors - (MBA specialization). (String) |
| mba\_percent | Percentage of marks in MBA degree. (Float) |
| status (TARGET) | Status of placement. Placed / Not Placed. (String) |

***Insights of EDA:***

* Looking at our categorical variables we managed to get some insights into our database.
* Most of our data is from male workers.
* Generally, our data is from the central.
* We have little data related to Arts and more data related to Commerce and Science.
* We can see that most of the data is from Communication & Management.
* Most of the data is from people with no experience.
* When we look at our Target variable, we can see that we have more data in Placed than Not Placed.
* Now talking about our continuous variables, we can see that we have some variables like ssc\_percentage , hsc\_percentage and degree\_percentage that the data is practically in a normal distribution, with most of the data in the middle and distributed practically equally left and right, and others that the data is more distributed like emp\_test\_percentage and mba\_percent.
* Looking at our boxplots we can see that we have no outliers that can influence the model.
* Looking at our target variable, we can see that we have a tendency for women not to be reallocated, another variable that we also see a different behavior is in the work\_experience variable, generally those with experience are reallocated more easily.

***Hypothesis***:

1. Now let's analyze the “ssc\_percentage” variable to see which profile of employees have the highest score, generally they are employees who are not from the central and more connected to technology and science, we can see that employees linked to art are the ones with the lowest score.
2. Now comparing with our continuous variables, we can see that generally when the employee has a high “ssc\_percentage” score and a high score in the other values, he almost always gets the Placed target.
3. Comparing our target variable with the continuous variables, we can see that we have a lot of difference between some variables, in the variables ssc\_percentage, hsc\_percentage and degree\_percentage we can see that those with higher grades are allocated more easily than those with lower grades.

***Suggestions:***

1. Dataset is small giving little flexibility for manipulations. So, Dataset should be expanded.
2. We have nonexistent Null Values. So, records or attributes are not dropped easily.

***Summary:***

* Looking at the correlation we can see that there is no strong correlation between our data, when looking at our Exploratory Analysis we can see that we have no outliers, when we look at our categorical variables we can have some conclusions about our database, most of our data are from men, usually from Central, and more focused on Commerce and Science, most of our data has no experience and when we look at our target variable, most of the data is from Placed, When we look at our continuous variables, most tend towards the mean and others are well distributed across all values.
* When we compare our categorical variables with our Target variable, we can see that the Not Placed result is usually for women, and that generally those with previous experience manage to be relocated, looking at our continuous variables, we can see that those employees who have a higher grade are more likely to be placed.
* When we take the ssc\_percentage variable to analyze, we can see that employees who are not from the central and who are linked to science are more likely to have a higher grade, when compared to our continuous variables, practically in most cases, those who have a high grade are always Placed, but when compared to our Work\_experience variable, this employee does not necessarily need to have had previous experience, just having good grades increases your chance a lot.