**PROJECT 1 - EXPOLOTARY DATA ANALYSIS –**

Titanic: Machine Learning from Disaster

AIM OF WORKING CHOOSEN DATASET

In my first project, the purpose of my intention to work with this data set was that it is being widely studied. I have seen that every person on the way to becoming a Data Scientist has encountered or worked on samples of this data set at least once. In my first data science project, I started working with this data set to apply what I know and make inferences.

WHAT DID I DO WHEN REVIEWING AND ANALYSING THIS DATA SET

As you know, it plays an important role in understanding and interpreting the data set you are working on. At this stage, I first worked to understand my data set.

I have 12 columns in my data set. I can explain the data in it as follows. The most important column in our data set that will allow us to infer is the **'Survived'** column. The **survived** persons can be thought of as 1, the **non-survived** as 0.

Then there's the **Pclass** column in our dataset. This column contains information about the level at which people continue their journey. And **Age** column contains age information of passengers. Afterwards, **Sibsp** column includes;

Sibling = brother, sister, stepbrother, stepsister

# Spouse = husband, wife (mistresses and fiancés were ignored). So Sibsp can be equal to 1 or more than 1 who has siblings/ spouse. Also it can be 0 who has non-sibling / non-spouse.

And the dataset defines family relations in this way **Parch** column. If parch value is equal or more than 1, we can say that passengers travel with their families. Some children travelled only with a nanny, therefore parch=0 for them.

The **Fare** column contains the values of tickets. The value **Embarked** is where passengers go. Other values in dataset are easily understandable.

ANALYSING THIS DATA SET

After uploading my train and test data to my workspace, I had the opportunity to review the data types clearly in terms of **Data Quality**. In this step, I examined dataset size, shape of dataset and dimension of dataset.

My next step to find count, mean or average, standart deviation, min, 25th percentiles,50th percentiles,75th percentiles, max values for columns which contains integer values. Actually in this time, PassengerId column wasn’t contains data we were going to use. Because it was a column that counted passenger numbers.

In my last data quality step shows relationship betwwen attributes is examined. If the correlation coefficient is negative, there is an inverse relationship between the two variables, ie "one of the variables increases while the other is decreasing". If the correlation coefficient is positive, it is interpreted that "one of the variables increases while the other increases". If 0.8>, it is interpreted that there is a very high correlation. I continued my work, observing these results.

Within the Scope of **Missing Value Check** step, I saw that my columns had some empty values. And I thought about whether the empty values in these columns were important and what I could do with those values. Age, Cabin, Embarked columns had empty values. I've displayed this numerically and percentile.

In Cabin values, 687 missing values obtained and this corresponds %77. In Age vaues, 177 missing values viewed and it corresponds %19. And in Embarked column, 2 missing values obtained and it corresponds %0.0022. I've seen this with heatmap too.

So I thought that the Embarked value would not help me analyze it, and I decided to remove these values from my dataset because they have very few empty values. I decided to remove missing Cabin Values them because there were so many empty values (%77) in the cabin column. I know there are empty values left in the age column, I thought about filling them according to the titles(Mr.,Mrs,Miss) in the names column. Then I added the new column my dataset as **Salutation** which includes titles as Mr, Mrs, Miss. Then I filled missing age values with grouping salutation and sex values and applied. So in this step, i have no missing values in the dataset.

After that, I wanted to see effects of Age groups on Survived/Not Survived. Therefore, I gropuped Age Values Babies=<2, Little Children= <7, Children 0-15, Adults 15-45, Elder 45+. I would say it is good that I do this grouping to visualize the survivors' ages by grouping them.

Then,t o update embarked values, I used basic dictionary, "C ->"Cherbourg", "S ->"Southampton", "Q->"Queenstown".

In **Target Visualization Step**, As I mentioned earlier, the survivors and the non-survivors are the most important column in which we will analyze our data.Then I want to see relationship between Survived or Not Survived and Sex(male /female). In this analising, I've noticed that women have a higher survival rate than men. I analyzed that men have a higher rate of non-survival.

Then I want to see, Survived or Not Survived rates relationships with Embarked(S-C-Q). But I couldn't do much analysis from here. Because there are data who is non-survived with Embarked=Southampton and also survived with Embarked= Southampton.

Moreover, I want to see Survived or Not Survived rates relationships with Pclass(1-2-3). As we can interpret plot, All passengers who did not survive are third-class passengers. First class passengers are the most surviving passengers.

Then, I want to analyse Survived or Not Survived rates relationships with Parch. In this plot, clearly we can see that Parch values are not effect this survived or not survived event.

Besides, I want to analyse Survived or Not Survived rates relationships with Sibsp. In this plot, clearly we can see that Sibsp values are effect this survived or not survived event. People who have not Sibsp or have 1 Sibsp can survived more than other people who have more than 1+Sibps.

Lastly I drop,useless columns for analysing as Embarked, Cabin, PassengerId, Sex.

So I did the analysis and I filled missing values in this EDA project scope.

Aleyna Sarışın