**Artificial Intelligence & Machine Learning | Batch 1 - Day 18 Overview**

**DAY 18 AGENDA**

* Boxplot
* Boxon Plot
* Dealing with Outliers
* Correlation
* Dealing with different correlation
* Feature Selection
* Heatmap
* PairPlot
* BarPlot
* ScatterPlot
* PointPlot

**Boxplot :**

1. data.bloxplot() will give the blox plot of the data.
2. Data.bloxplot(figsize=()) The arguments can be passed in figsize to define the size of the figure.
3. Bloxplot also helps in the detection of outliers.
4. We can explore the interquartile ranges.
5. Look at the 5 point summary.
6. The data which falls out the plot are noises or outliers.
7. Pandas and Seaborn both can plot boxplots.
8. Seaborn bloxplots are much more interactive than the One plotted using Pandas.
9. We can infer out the features which have maximum, moderate, and least outliers.
10. BoxPlot as based on the 5 point summary
11. distplot plots a distribution plot

**Boxon plot :**

1. It is a seaborn graph.
2. It is more interactive than bloxplot.

**Dealing with  outliers:**

1. when some features have max outliers in that case we can either delete or use those outliers.
2. Deleting more outliers we could loose originality of data.
3. Removing many outliers will make the model incapable of its task.
4. Having the domain knowledge you can decide which rare cases can be eliminated.
5. We should often accept moderate outliers.
6. We can directly accept the least outliers.

**Correlation:**

1. It is a mathematical coefficient which gives the relation between two cases or features.
2. Using correlation we can define how much one feature affects the other feature.
3. The range is between -1 to +1.
4. -1 says the features have a strong negative correlation.
5. +1 says the features have a strong positive correlation

**Dealing with different correlation:**

1. -.2 tp +0.2 is considered as no correlation.
2. -0.3 to -0.6 and +0.3 to +0.6 is considered as moderate negative/positive correlation.
3. -0.6to -0.8 and +0.6 to +0.8 is a good correlation.
4. More than -0.8 to +0.8 is a strong correlation

**Feature Selection :**

1. When there is no correlation simply drop these correlations.
2. Based on the correlation coefficient we select the appropriate features.
3. We should check all other coefficients of relation in data and then we should decide what to do with that feature

**Heatmap :**

1. This helps us to plot the correlation.
2. Sns.heatmap(data.corr())  displays the heatmap.

**Pairplot**

1. sns.pairplot(data) will plot the pairplot for the data.
2. 2 Paired samples are used to plot one data
3. We can decide the correlation lookin at the graph
4. sns.pairplot(data,hue=”Outcome”) Displays the pairplot in 2 colours

**Bar plot :**

1. sns.barplot(x=data[],y=data[]) Gives you barplot between specified features

**Scatter Plot:**

1. sns.scatterplot(x=data[],y=data[]) Gives the scatterplot between the specified features

**Point Plot :**

1. sns.pointplot(x=data[],y=data[]) Gives the pointplot between the specified features

**Example EDA of a dataset :**

1. WE can use data[].value\_counts will give you the number of records in each feature.
2. pd.crosstab(x=,y=) Gives a cross table for the specified features .It can two values and the first value can be string or categorical .
3. Sns.lmplot(x=,y=)
4. You can drop features use x=data.drop(axis=0/1) The axis defines either row or column that has to be deleted of the feature