PES UNIVERSITY

-RR CAMPUS

STATISTICS DATA SCIENCE

DATASET CHOSEN : STALEMATE

TEAM MEMEBERS:

CHINMAY KULKARNAL(PES1UG19CS127),

BRUNDHA.P(PES1UG19CS116),

DEBADITYA RAY(PES1UG19CS133),

ANUDEEP CVS(PES1UG19CS070).

SEMESTER :3.
SECTION :B

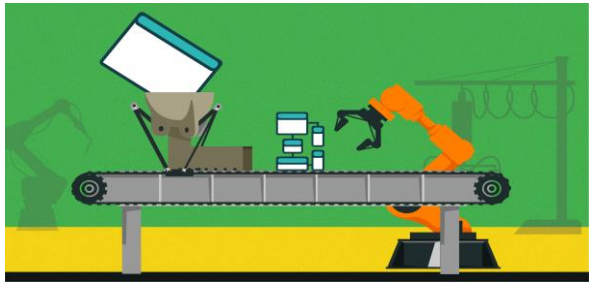
INTRODUCTION TO DATASET: STALEMATE.

- THIS DATASET CONTAINS 4.5% OF NAN VALUES.
- THE NUMBER OF WINNERS IN THE OF BLACK AND WHITE SIDES OF THE CHESS.
- EACH OF THE WHITE AND BLACK SIDE PLAYERS OF THE CHESS AS A UNIQUE ID FOR EACH NAMED AS WHITE_ID AND BLACK_ID.
- THE MOVES MADE ARE ALSO NOTED IN THE DATSET AND THE TIMING OF THE PLAYERS IS RECORED.
- THE STATUS OF VICTORY IS FILLED BY VICTORY_STATUS IN THE FORM OF RESIGN, MATE AND OTHERS.
- THE RATING OF WHTE AND BLACK PART IS ALSO IN THE FORM WHITE_RATING AND BLACK_RATING.

DATA CLEANING

- THE FACTORS THAT LEAD TO CLEAN IN THE DATASET:
  - CONSTRAINTS
  - VALIDITY
  - RANGE
  - DATA TYPES
  - DUPLICATION
  - REGULAR PATTERNS

NORMALIZATION



- Normalization is a technique often applied as part of data preparation for machine learning. The goal of normalization is to change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values.
- Here, normalization doesn't mean normalizing data, it means normalizing residuals by transforming data. So normalization of data implies to normalize residuals using the methods of transformation.

EFFECTS OF NORMALIZATION:

- It often refers to rescaling by the minimum and range of the vector, to make all the elements lie between 0 and 1 thus bringing all the values of numeric columns in the dataset to a common scale.
- Similarly, the goal of normalization is to change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values. For machine learning, every dataset does not require normalization. It is required only when features have different ranges.

CORRELATION

- **Correlation is the co-relationship or association of two variables.**
- Usage of the correlation is to represent linear relationship between two variables. There is no difference in dependent and independent variables. The main objective is to find a numerical value expressing the relationship between the variables. It indicates the extent to which two varibles move together.
- Correlation is a statistical measure. Correlation explains how one or more variables are related to each other. These variables can be input data features which have been used to forecast our target variable. ... It means that when the value of one variable increases then the value of the other variable(s) also increases


Thank
You