PREDICTING THE RAIN PHENOMENA IN SAUDI ARABIA

NAIF ALORFI

IS IT GOING TO RAIN?

WEATHER INFORMATION CONSIDERED IMPORTANT IN MANY FIELDS SUCH AS AVIATION, AGRICULTURE, MUNICIPALITIES AND MANY OTHERS. WE NEED WATER FOR MANY REASONS AND WEATHER CONTROLS THE DISTRIBUTION OF RAIN WATER ON EARTH. THE AIM OF THIS STUDY IS TO ANALYSIS THE CORRELATION OF THE WEATHER INFORMATION AND BUILD MODEL THAT HELP TO PREDICT RAIN PHENOMENA IN SAUDI ARABIA.

USING MACHINE LEARNING ALGORITHMS

THIS PROJECT AIMS TO PREDICT THE RAIN USING MACHINE LEARNING ALGORITHMS BASED ON OTHER WEATHER ELEMENTS SO MAY IT CAN HELP FORECASTERS IN THE METEOROLOGY CENTER AND OTHER OFFICIALS FOR EXAMPLE MUNICIPALITIES NEED THE RAIN FORECAST TO GET PREPARED FOR ANY INCIDENT THAT MAY OCCUR BECAUSE OF THE RAIN SUCH AS FLASH FLOOD. I UNDERSTAND THERE IS MANY FACTORS THAT SHOULD BE CONSIDERED WHEN FORECASTING THE WEATHER. THIS PROJECT IS BASED ON THE DATASET AVAILABLE PUBLICLY AT HTTPS://WWW.KAGGLE.COM/ESRAAMADI/SAUDI-ARABIA-WEATHER-HISTORY WEBSITE WITH TITLE "SAUDI ARABIA WEATHER HISTORY".

DATA PRE-PROCESSING

- CLEANED UP THE DATASET BY REPLACING THE MISSING VALUES WITH MEAN PER CITY.

- CHANGE THAT NECESSARILY COLUMNS TO THE CORRECT DATA TYPE .

- DEALING WITH OUTLIERS TO THE MINIMUM AND MAXIMUM VALUES USING INTERQUARTILE RANGE.

- ADDRESSING IMBALANCED DATASETS BY OVERSAMPLE THE MINORITY CLASS.

- FEATURE SCALING IS IMPLEMENTED BEFORE TRAINING THE MODELS.
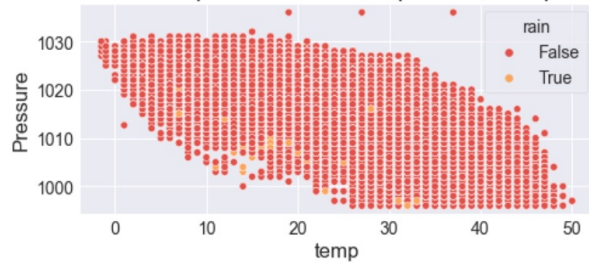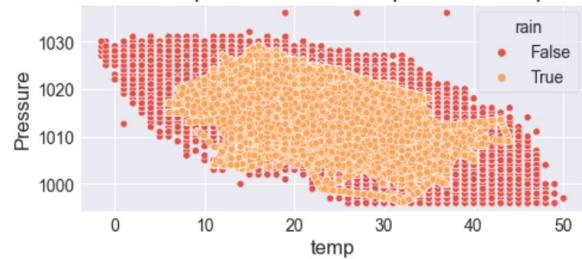
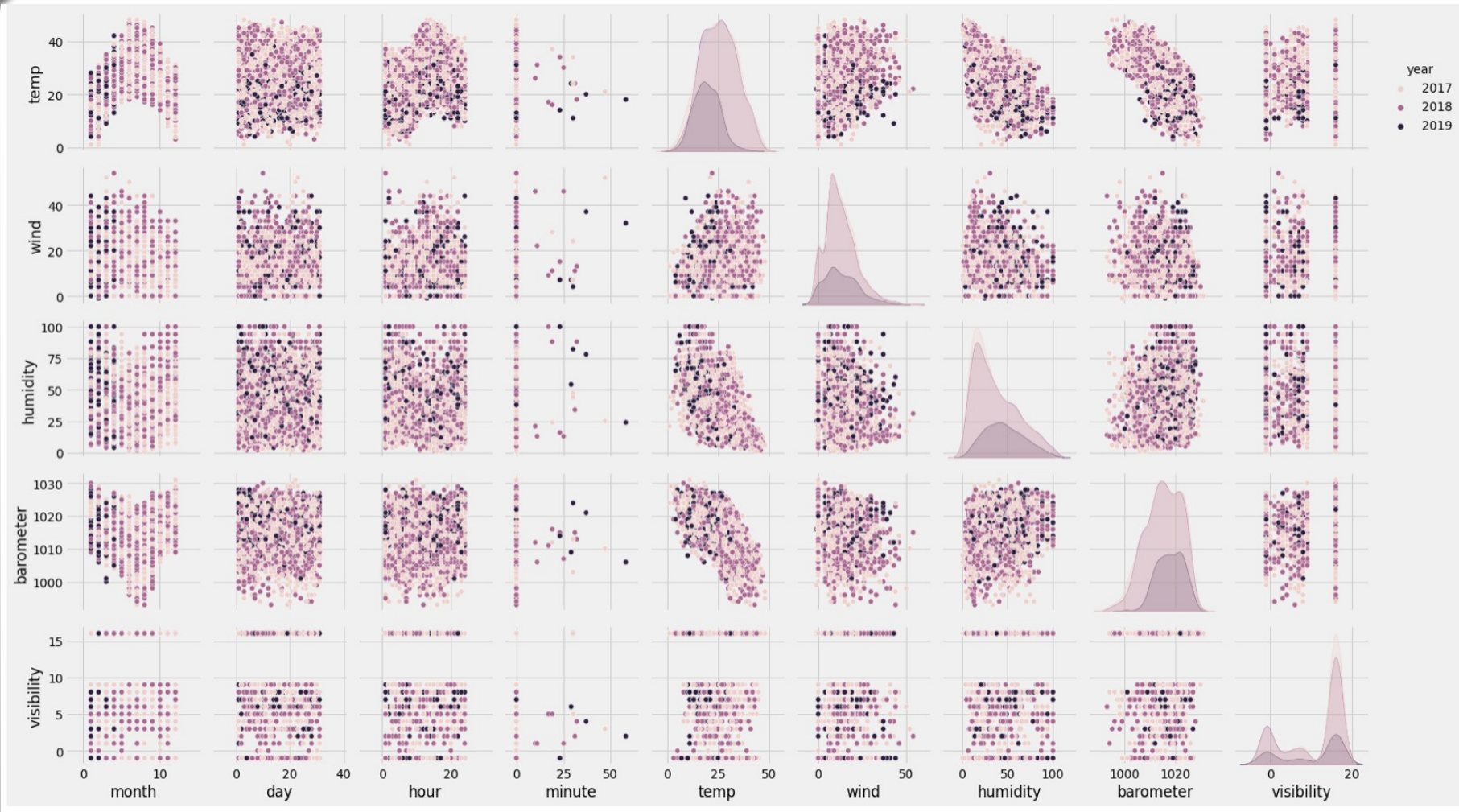# DEALING WITH IMBALANCE DATA USING "SMOTE"

THIS APPROACH FOR ADDRESSING IMBALANCED DATASETS BY OVERSAMPLE THE MINORITY CLASS. `RAIN` WAS IMBALANCED WITH 245696 `FALSE` AND 3327 `TRUE` OBSERVATION. SO I GOT 245696 FOR EACH AFTER IMPLEMENTING SMOTE.
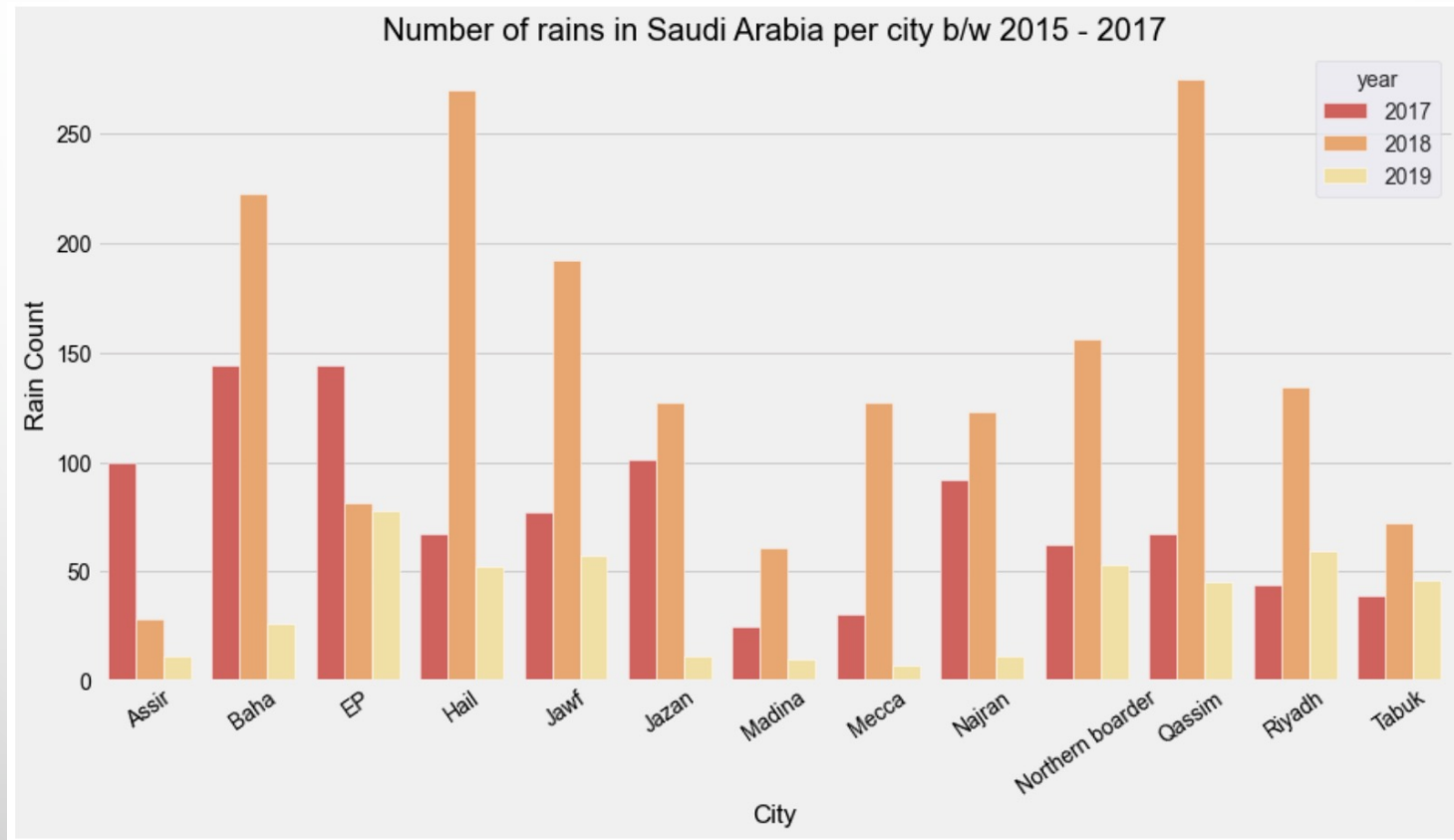
EXPLORATORY DATA ANALYSIS (EDA)

Correlation Heatmap of Rain in Saudi Arabia Dataset
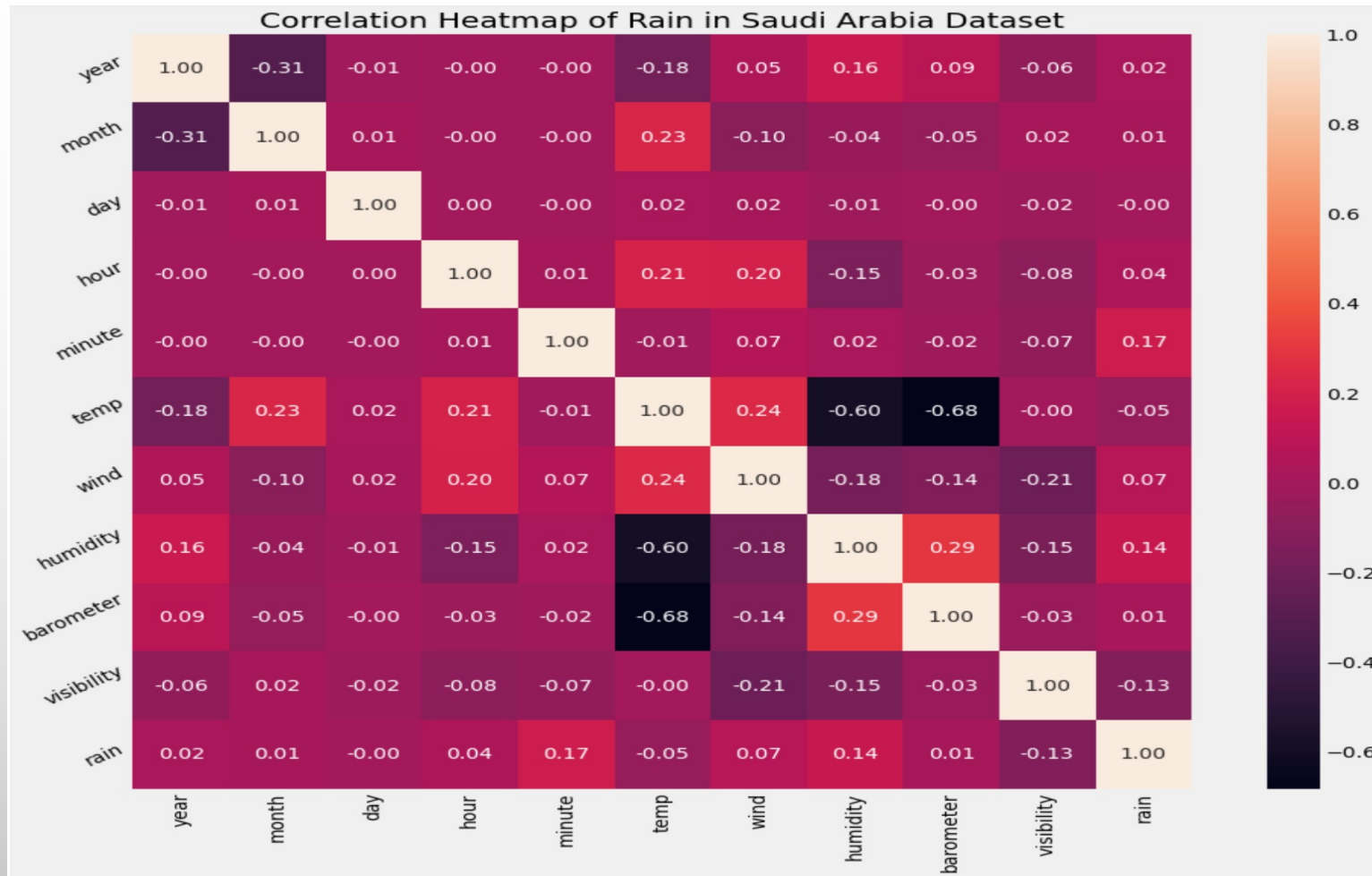
EXPLORATORY DATA ANALYSIS (EDA)

EXPLORATORY DATA ANALYSIS (EDA)

Number of rains in Saudi Arabia per city b/w 2015 - 2017
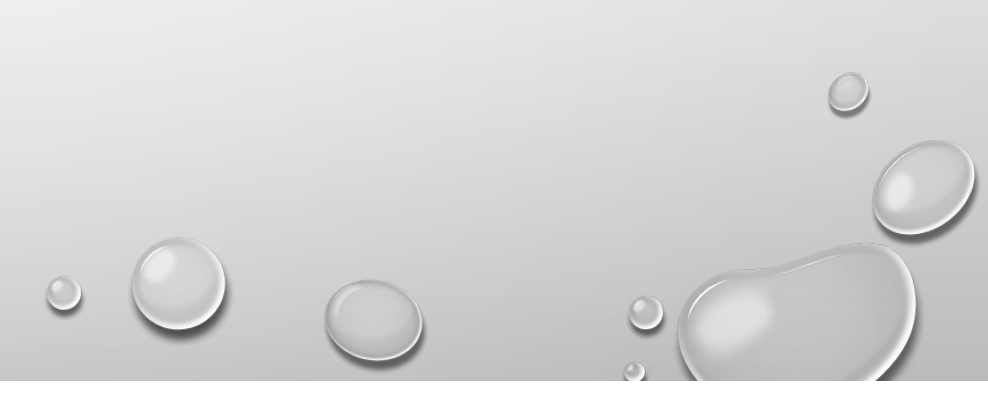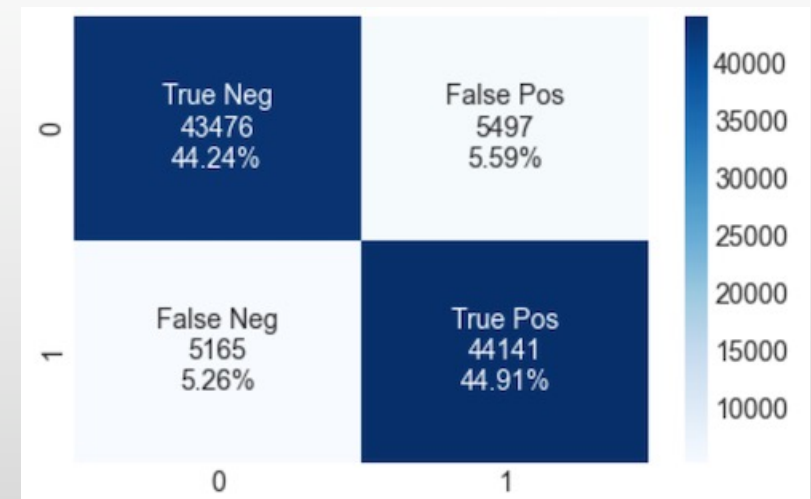
MACHINE LEARNING MODELS

- 1- LOGISTIC REGRESSION.

- 2- ARTIFICIAL NEURAL NETWORK (ANN).
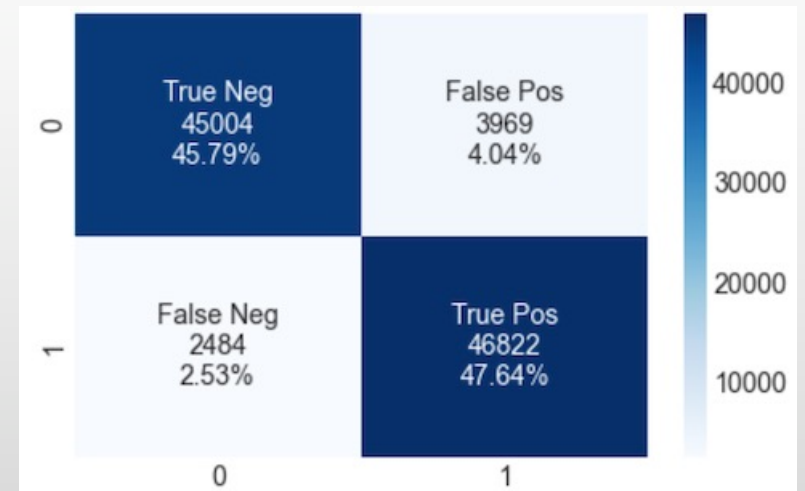
- 3- RANDOM FOREST CLASSIFIER.

LOGISTIC REGRESSION

- FIRSTLY, I HAVE TRAINED THE MODEL WITH LOGISTIC REGRESSION WITH PARAMETERS `SOLVER= 'LIBLINEAR' , C=1000` AND I GOT **THE ACCURACY 89.22 %**
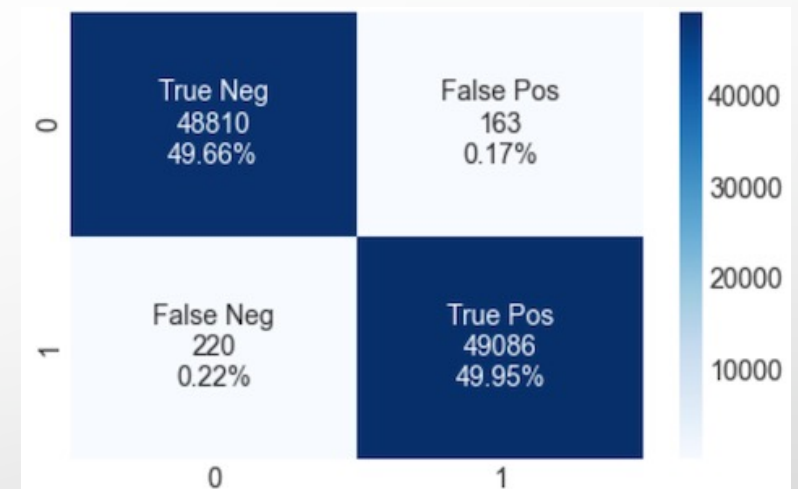
ARTIFICIAL NEURAL NETWORK (ANN)

- SECOND, I HAVE TRAINED THE MODEL WITH ARTIFICIAL NEURAL NETWORK WITH PARAMETERS `BATCH_SIZE = 3200, EPOCHS = 100`

- AND GET **THE ACCURACY 93.24 %**

RANDOM FOREST CLASSIFIER

- LASTLY, I HAVE TRAINED THE MODEL WITH LOGISTIC REGRESSION WITH PARAMETERS `SN_ESTIMATORS = 10, CRITERION = 'ENTROPY', RANDOM_STATE = 42`

- AND GET **THE ACCURACY 99.58 %**

CONCLUSION

- I HAVE CLEANED UP THE DATASET BY REPLACING THE MISSING VALUES WITH MEAN PER CITY.

- THEN, I HAVE RESOLVED THE OUTLIERS TO THE MINIMUM AND MAXIMUM VALUES THEN I HAVE BALANCED THE DATA WITH SMOTE AND THE FEATURE SCALING IS IMPLEMENTED.

- FINALLY, I HAVE TRAINED THE TRAINING DATASET WITH THREE MODELS AND GOT THE ACCURACY, RECALL, PRECISION AND F1-SCORE:

- 1- LOGISTIC REGRESSION.

- 2- ARTIFICIAL NEURAL NETWORK (ANN).

- 3- RANDOM FOREST CLASSIFIER.

- IT SEEMS THAT, `RANDOM FOREST CLASSIFIER` WAS THE BEST MODEL FIT WITH ACCURACY SCORE **99.5 %**.

- FUTURE WORK: WORK IN DATASET FROM NCM ALSO CONSIDER RADAR IMAGES AND OTHER FACTORS.

THANK YOU