Dataset Description:

The data has been collected from the regional weather stations and the backhaul radio tower sensors [24] by the organization network experts. The first set of data contains the entry from January 2019 to December 2019. Then, we have collected data from January 2020 to March 2020. We have trained our model with the 2019 data only.First We have tested our model performance for 2019 test data and then we tested with completely new data from 2020. The features presented in the 2020 data are the same as the 2019 data. There are six subsets in the original data set that represents different data for different scenarios. The brief description of the subsets are following,

## Met-Forecast Data
The weather forecast data contains entry from '02-01-2019' to '31-12-2019' of 96 weather stations with weather-station, date-time, report-time and weather,tempmax,humidity-max, humidity-min, wind-dir, wind speed for day 1 of the date-time to day 5 of the date time.The features are described briefly in the following,
• The 96 weather station name for each date. E:g : WS-17038
• Date and time from January 02 to December 31 of 2019 with time
• The time when the weather is measured for forecasting. E.g : morning.
24
• The type of weather of the day is taken for particular weather station.
E.g : scattered clouds, rain.
• The maximum and the minimum temperature of the day enlisted in different features.
• The maximum and the minimum Humidity of the day enlisted in different features.
• The wind direction and speed are taken in two different features.

## Met - Station
The data-set contains spatial features of the weather stations. The features are , weather stations name, ground height which has the value for the distance of the weather station from the sea level and the clutter class which indicates the type of environment of the weather station location.

## Met-Real
The weather forecast data contains entry from '01-01-2019' to '31-12-2019' of 112 weather stations with of each hour of the day with station-no, date time,

measured-date, measure-hour, temp (max,min,current), wind-dir (max,min), wind-speed (max,min), humidity,precipitation, precipitation-coeff, pressure, pressure-sea-level. Here measure data is the same date as the date time and the measure-hour is the time of the day that the measurement is taken.

# RL- Sites

The data-set contains spatial features of the Radio link sites. The features are site-no, ground height which has the value for the distance of the radio link site from the sea level and the clutter class which indicates the type of environment of the site location. There are 1387 sites entry in the Rl-sites dataset.

# RL - KPIS

Each of the row in the Performance Indicator (Kpis) data-set represents the features that are used to measure the performance of the radio links. The measurement of each link is taken for every date of 2019 for each site. The features for each links are following,

• **Type:** It represents the radio link equipment vendor type. There are two equipment provider companies which are ENK wireless INC. and NEC enterprise. The radio link failure event has no correlation with the vendor equipment as link failure event occurred for both type of equipment vendor.
*Correlation:* The number of failure event occurred more for ENK type than the NEC type as the amount for ENK type is approximately 40 more than NEC type.

• **Date-Time :** The date and timestamps from 2nd January of 2019 to 31st December 2019. The measurements are taken at 12.00 am for each date and for the most of the radio link sites

• **Tip:** This feature represents the end-points of the link in the site. The end-point is the entry and exit point of the radio link to the sites. It is in different locations such as near the station or far from the station. The link access entry points has no significant effect on the radio link failure event

• **Mlid:** The mini-link id for each site RL. It is a market-leading microwave family for cost-efficient mobile transport networks. It's a microwave cell that enables the 5g ubiquitous transport. This feature represents the id of the microwave mini-link device of the site no. A site-no may have multiple mini-link device. This is the reason behind the multiple entry of the site-id in the dataset where Mlid are unique

for each mini link devices . E.g : K7DZ, K7ER.

• **Mw-connection-Id:** This feature represents the internal Microwave connection Id. This is the unique connection Id for mini-link devices for each of the site every day.

• **Site No :** This feature represents the site no for the 5g radio link sites. There are approximately 1387 site no in the dataset. The sites are the area where the 5g network antennas and the devices are situated. *Correlations:* The sites are correlated with the weather stations in terms of the distance between a WS and a 5g network site area

• **Polarization:** Radio Link antenna polarization. This feature contains vertical and horizontal polarizations. This form of antenna polarization has horizontal elements. It picks up and radiates horizontally polarized signals, i.e. electromagnetic waves with the electric field in the horizontal plane.

• **Card Type :** A modem card is an internal type of modem that is plugged into the PCI slot. A Radio Link modem is a communications device that allows antenna to transmit data wirelessly. There are 5 different types of modem cards in the dataset. The feature doesn't have any correlation with the radio link failure event as the different card types represents the different card provider.

• **Adaptive Modulation :** Adaptive Modulation is a technique which allows a radio link to change its speed as conditions in the network changes. The adaptive modulation for the RL sites are enabled or disabled which doesn't have the correlation with the radio link failure.

• **Frequency Band :** This features represents the standalone 5g (millimeter wave) frequency bands which are divided into five bands (f1,f2,f3,f4,f5).
The frequencies are 26,24,39,28 (GHz) and commonly named (LMDS,Kband, ka-band, ka-band) respectively.

• **Link Length :** This is the distance in kilometer between two RL sites.

• **Severely Error seconds :** This feature represents the count of the 1 second period time where the RL error occurs with more than 30 % damaged of the RL frame.

• **Error Second :** Count of one second period with error in the link.

• **Unavailable Second :** The RL operation disruption in seconds. The duration of the RL unavailable operation duration in second.

• **Available time :** The total operation time of the network for one day in seconds. The measurement of the error seconds are deducted from the available time.

• **BBE :** Background Bit error (BBE) represents the number of bit error in a unit time. It is the number of bit error divided by the total transmitted bit for each link. This is the indicator of the performance degradation in each link in the dataset.

*Correlation :* The high value of the unavailable RL operation and the BBE combined results the radio link failure event.

• **Radio Link Power Level :** This represents the transmission energy level that is received by the radio links.

• **Scalability score :** This is the modulation scalability score only for the links where adaptive modulation operation is enabled.

• **Capacity :** This feature is the RL transmission capacity of each link which is represented in Mb/s.

• **Modulation :** This feature represents the Quadrature Amplitude Modulation (QAM) that is deployed for the radio links. It is a method of combining two amplitude-modulated (AM) signals into a single channel, thereby doubling the effective bandwidth. The microwave carrier modulation ranges from 16 QAM to 2048 QAM.

• **Radio Link Failure :** It represents the radio link failure event that is occurred for each link. It is shown by True or False.

The target variable for the dataset = Radio Link Failure event

## Distance Matrix

The distance matrix represents the distance between the weather station and the radio link tower site. The data has the entry in terms of distance in kilometers for each weather stations and the site id. The distance matrix is the only correlation found between the weather stations and the radio link tower site in terms of the pre-processing of the dataset for the learning model.