# Title page

Studio 1 – 6

Group Number: 1

Theme: 5 (Electrical Engineering / Telecom Engineering)

Group members:

Dang Khoa Le

BAO Nguyen

Nguyen Hung Nguyen

Trung Kien Nguyen - 104053642

Saw Ko Ko Oo

PROJECT BRIEF

# 1. Introduction

## Background and Motivation

Nowadays, there are more and more options to enhance network performance thanks to the development of 5G technology. The utilisation of machine learning techniques is important in order to optimise network operations, given the abundance of available network performance data. The goal of this study is to forecast future network performance by clustering geographical zones depending on their performance using 5G network performance data. Network operators would greatly benefit from this in terms of enhanced service quality and effective resource management.

The primary consumers of these AI models will be network administrators and telecom engineers who are in charge of keeping an eye on and improving the performance of 5G networks. In order to deploy resources proactively, they will utilise the model to forecast performance concerns and identify zones with weak connectivity.

## Project Objectives

This project's main objective is to create a machine learning model that will use network performance data to categorise and forecast traffic levels (High, Medium, or Low). With this model, we expect the network operators will be able to manage and optimise the 5G network more efficiently, as they can now anticipate future traffic conditions and understand traffic patterns in real-time.

## Project Schedule

The project is going to kick off from the start of Week 7, and expected to finish at the end of Week 13, based on the Semester 2 – 2024 calender of Swinburne University of Technology.

* Week 7: Propose appropriate datasets, and get feedback from the Tutor
* Week 8: Finalize the dataset selection and begin the data exploration process.
* Week 9: Perform feature selection and preprocess the dataset to prepare for training
* Week 10: Train and validate (test) time-series forecasting models.
* Week 11: Do the evaluation of the model(s) and come up with ideas about final report(s) and presentation
* Week 12: Present the project demonstration
* Week 13: Finalize and submit the implementations, and the final report.

# Data

## Data source

The chosen dataset for our project contains 5G network performance data in CSV format, providing a comprehensive overview of the network's performance in different aspects, including various performance metrics. The data spans multiple geographical zones (identified by latitude and longitude) and is recorded over time. There is a total of 74 features, here is the summary of some significant ones:

* Input features
  + Date: This is the timestamp associated with each observation.
  + Duplexing Type: The duplexing mode (FDD/TDD), which influences network performance.
  + Site Id and Sector Id: These ids represent the location of the network site and its sector, which are key for geographical clustering.
  + Carrier Number DL (earfcn): The downlink carrier frequency, used for analyzing performance differences between frequencies.
  + Call Setup Success Rate (%) is a critical feature representing the success rate of calls, indicating network reliability.
  + CSFB Preparation Success Rate: Indicate the effectiveness of circuit-switched fallback preparations.
  + DL/UL Traffic (GB): The amount of data transferred, representing network load.
  + Throughput Metrics: These include downlink and uplink user throughput (Mbps), which are core features for predicting network performance.
  + CSSR (Call Setup Success Rate) Metrics: These are further broken down into RRC and eRAB success rates, critical for understanding network setup and reliability.
  + ERAB Drop Rate: This feature measure the network session reliability, indicating the frequency of dropped sessions.
* Potential target feature (Label): This is the **Traffic Level**, classified into three categories:
  + High Traffic
  + Medium Traffic
  + Low Traffic

## Data processing

For this task, we aim to process the dataset to be as clean and clear as much as possible. Here is the brief plan of the data preprocessing task:

* **Collecting and merging features**
* **Handling the missing values in some features.** Temporarily, our plan is to remove any rows with excessive missing data if the missing information exceeded a predetemined threshold percentage.
* **Labelling data**: The traffic levels (High, Medium, Low) were labeled based on the volume of DL/UL Traffic (GB) and Throughput (Mbps). The thresholds for this task will also be predetermined
* **Selecting appropriate features**. There is a total of 74 features in the original dataset, so we will refine it by removing some inappropriate features
* **Scaling/Normalizing features**: The techniques used is Min-Max Scaling

# Requirements

## Must-have functionalities/features

* **Traffic Level Prediction**: The system (model) must be able to predict precisly the level of network traffic (High/Low/Medium) based on the given input features. The accuracy rate is expected to archieve at least 85%.
* **System (Model) Evaluation**: This include an assessment of the machine learning models used for classification and clustering, incorporating measures like accuracy, precision, recall, and F1-score for traffic prediction. The output visualization(s) must be provided.
* **Easy-to-use and understandable UI:** It is expected to develop a simple and user-friendly UI to train/test/evaluate the model. The plan is to develop a console-based UI, providing a easy approach to input and modify any parameters
* **Data preprocessing Pipeline:** The system must include a robust data preprocessing pipeline, covering everything, but not limited to, the features in section 2.2.

## Optional features

There are a wide range of desirable for a predicting ML system. At this stage, we expect that our deliverable can perform some additional tasks as follow:

* **Processing the data based on a real-time dataset**. Our system might be able to handle real-time data streams. In other words, it allows dynamic and live prediction of traffic levels based on real-time network input data.