Team Qualifier 2024
Technical Document Specification

A national initiative of the Department of Science and Innovation and implemented by the CSIR.
science & innovation
Department: Science and Innovation
REPUBLIC OF SOUTH AFRICA
CSIR

DIRISA SDC
STUDENT DATATHON CHALLENGE

Team Qualifier 2024
Technical Document Specification

Table of Contents

DIRISA SDC
STUDENT DATATHON CHALLENGE

Team Qualifier 2024
Technical Document Specification

1. Problem Overview

South Africa is a country with a high youth unemployment rate, which is currently at 46.3%. Youth unemployment and under-employment are pressing challenges in developing countries, serving as crucial indicators of economic health. South Africa is not spared from these challenges. As of the first quarter of 2023, South Africa's youth unemployment rate, focusing on jobseekers between the ages of 15 and 24, rose to 62.1%, marking the highest rate in a year. The total number of unemployed youth (ages 15-34) increased by 241,000, reaching 4.9 million. These statistics show that youth unemployment is a serious problem, as it can lead to poverty, crime, and social unrest.

2. Essential Skills and Knowledge

The following skills and knowledge are required:

- Proficiency in Pre-Processing, involving the preparation and cleansing of data.
- Expertise in Statistical Analysis to extract meaningful insights from datasets.
- Adeptness in Data Visualization for effective communication of findings.
- Competence in Machine Learning to build robust forecasting models.

Participants have the flexibility to utilise any language, framework, development environment, library, or API in their forecasting and modelling endeavours.

3. Challenge Description

Successfully forecasting youth unemployment rates and predicting individual employment statuses necessitates a comprehensive strategy integrating data analysis, economic comprehension, and social considerations.

Participants are tasked with precision in forecasting regional unemployment rates based on the provided dataset. Moreover, they are required to utilise historical data to predict individual employment statuses. Emphasis should be placed on the models' predictions and pivotal features affecting the predicted unemployment rate and individual employment status.

4. Challenge Datasets

The datasets required for this challenge can be downloaded at: https://sdcdata.dirisa.ac.za/index.php/s/LS64EmGLP8Hpj7G

There are two types of datasets available on the SDC Next Cloud that need to be used for forecasting and for the prediction.

Forecasting:

Source of dataset - https://data.worldbank.org/indicator/SL.UEM.1524.ZS

Dataset - API_SL.UEM.1524.ZS_DS2_en_csv_v2_5996635

Predict employment status:

Source of dataset - https://www.datafirst.uct.ac.za/

Dataset - ghs-2012-person-1.0 to ghs-2022-person-v1

5. Challenge Marking Rubric

You will be assessed on the following elements:

- Data understanding and exploration.
- Classification model.
- Forecasting model.
- Documentation.
- Bonus points

6. Challenge Submission

- You are required to submit your code by **8am on 13 August 2024**.
- You are required to record a 15 minutes presentation of the work you have done – each member of the team needs to present.
- Your submission must be done on the NextCloud platform - https://sdcdata.dirisa.ac.za/index.php/login
- Only **one** team member needs to upload the relevant submission that comprises of:
    - Notebook with code, comments and references to packages used if needed.
    - Video recording

- Presentation slides with references.

The team member needs to create a shareable link on NextCloud. This link must be shared with Dr Nobubele Shozi via Slack