BITS F464 – MACHINE LEARNING

PROJECT – MACHINE LEARNING FOR SUSTAINABLE DEVELOPMENT GOALS (SDGs)

Due Date: Saturday, November 20th, 2023, at 23:59 IST

Marks: 20 marks (10%)

Assessment type: Written evaluation with model results in a Google Colab Notebook and Live Demonstration

Overview:

In this project, you must showcase your knowledge of machine learning models learnt in the course. The project will allow you to apply your learning throughout the course and use what you learnt for conducting an ML for SDG project.

You will be assigned to an ML for SDG project in which you must answer the problem statement to help solve an issue plaguing the world. The respective dataset will be given to you, along with the information about the respective features. You must appropriately apply three models taught in class and a model found in the research literature.

Assessment criteria:

This assessment will measure your ability to:

- Apply the appropriate machine learning methods for the respective problem statement.
- Analyse data quality and quantity of a dataset to determine if the data is sufficient.
- Compare, interpret, and communicate outputs from machine learning models.
- Suggest ways to improve the models based on performance across all models.

Assessment details:

This assessment aims to help you gain hands-on experience by developing and applying appropriate machine learning models to help resolve ML for good problems. Your team will be randomly assigned to one of six problem statements. Each scenario is linked to an SDG and contains the relevant data for modelling. Use the dataset with four appropriate models coded and built from scratch **(*three should be from those covered in class and not covered as part of the two assignments and one that you discover from research literature*).**

The following are the project descriptions. Please refer to the one that your team has been assigned to.

1. **Early Detection of Heart Disease Using Machine Learning** (Dataset and Feature Description)

   Description: Heart disease is a leading cause of mortality globally. Early detection and accurate diagnosis are crucial for improving patient outcomes. Develop a machine learning model that can predict the presence or absence of heart disease in patients based on their clinical and demographic information using the dataset.

Insight Deliverable: A predictive model that accurately classifies patients into those with heart disease and those without it.

2. **Identification of Portable Water** (Dataset and Feature Description)

   Description: Safe and readily available water is essential for public health and usage. Improved water supply and sanitation and better management of water resources can boost countries' economic growth and contribute significantly to poverty reduction. Contaminated water and poor sanitation cause transmission of cholera, diarrhoea, dysentery, hepatitis A, typhoid, and polio. Inadequate sanitation services expose individuals to preventable health risks. Using the water quality dataset, create a machine learning model to distinguish between potable and non-potable water.

   Insight Deliverable: Distinguish between potable and nonpotable water.

3. **Quality of Air** (Dataset and Feature Description)

   Description: Clean air has a significant positive impact on several SDGs, such as Good Health and Well-Being, Affordable and Clean Energy and an indirect impact on others. Air pollution can contribute to global mortality rates as it causes respiratory and cardiovascular diseases. Building sustainable cities and communities relies on safe levels of particulate matter pollution, achieving universal access to sustainable energy and shifting away from the dirty fuels that cause air pollution. Solutions to improving air quality, such as switching to affordable and cleaner energy, cooking, and lowering transport emissions, will also address the climate emergency. Air pollution can harm ecosystems and biodiversity. Hence, clean air will help protect these species and maintain healthy ecosystems. Using the given dataset, create machine learning models to identify the air quality at any given point in time.

   Insight Deliverable: Identify the air quality at any given point in time based on historical data.

4. **Prediction of wildfires** (Dataset and Feature Description)

   Description: Wildfires are a natural phenomenon but are becoming more dangerous and affecting substantial areas. Wildfires are predicted to worsen in the coming years and decades, the United Nations Environment Programme (UNEP) has warned in its annual Frontiers report released February 17, 2022. There has been a rapid expansion of cities towards forest areas in many regions in recent decades. This wildland-urban interface is the area where wildfire risks are most pronounced. The given dataset gives you an insight into the forest fires in the Northeast region of Portugal. Using the dataset, identify the probability of future fires and the extent of damage.

   Insight Deliverable: Predict future fires and the extent of damage.

5. **Emissions of Carbon Dioxide** (Dataset and Feature Description)

   Description: Carbon dioxide emissions are a primary driver of global climate change. The world needs to reduce emissions to reduce climate change's impacts urgently. How the responsibility will be shared between regions, countries, and individuals has been an endless point of contention in international discussions. The debate arises from the numerous ways in which emissions are compared, such as annual emissions by country, emissions per person, historical contributions, and whether they adjust for traded goods and services. These metrics can tell entirely different stories. Using the given dataset from the Canadian Government of the amount of CO2 emissions by a vehicle, identify the influence of different variables on the emission of CO2 and if there will be any difference in the CO2 emissions when fuel consumption is considered separately.

Insight Deliverable: Identify the safest/ best car models based on their emissions.

6. **Analysing Economic Growth Trends** (Dataset and Feature Description)

    Description: This project aims to analyse historical economic growth trends using the German Credit Data dataset.

    Insight Deliverable: The insights from this analysis can include identifying factors that influence economic growth, predicting future economic trends, and assessing the impact of credit data on economic stability.

Evaluation requirements:

You will be evaluated based on the

- Preprocessing of the dataset (2.5 marks),
- Selection and application of 4 ML (3 taught and not covered in assignments + 1 from research literature) models (10 marks),
- Presentation and Readability of your markdown file (2.5 marks),
- Error-free run of the code during the live demonstration (2.5 marks),
- Ability to answer individual questions related to the project (2.5 marks).

Please ensure that you have explained the steps of your code using comments and markdown text to interpret the results obtained in your markdown file. Appropriate references should be included in the markdown file to showcase your literature review and selection of an appropriate new model.

Submission format:

Use the supplied ipynb notebook to add your code and insights. After running and testing the entire code, please submit both **ipynb and PDF versions of your** code. In the header of the notebook, add markdown text containing your team's information, such as team number, full name of all members and their id numbers. Also, add your team no to the name of the files (For example, **TeamXX_Project.pdf and TeamXX_Project.ipynb**). Files to be submitted in a single folder (named **TeamXX_Project**) on Google Classroom before the specified deadline are as follows:

a. ipynb notebook
b. pdf version of ipynb notebook

IMPORTANT NOTE:

- ***USE ONLY THE PROVIDED PYTHON NOTEBOOK FORMAT TO SUBMIT YOUR PROJECT***

- ***ONLY ONE SUBMISSION PER TEAM NEEDS TO BE DONE. DO NOT MAKE MULTIPLE SUBMISSIONS.***
- ***NO MAKEUPS AND LATE SUBMISSIONS WILL BE ACCEPTED AND MARKED.***
- ***ANY KIND OF PLAGIARISM WILL LEAD TO SEVERE PENALIZATION.***

Contact for clarifications:

In case of any queries, please email the course's Teaching Assistants (TAs), and any other communication is invalid. It would be best if you wrote a mail to **all** the following TAs for clarification.

1. Pranjali Attarde, p20220018@hyderabad.bits-pilani.ac.in
2. Priya Baju, p20230804@hyderabad.bits-pilani.ac.in
3. S Shashank, p20210412@hyderabad.bits-pilani.ac.in