EXPLORE | DIGITAL SKILLS

Classification Predict
Instructions

Your Mission: Share your Work with the World

Regression gave us an **introduction** to various aspects of machine learning within the EGAD framework.
For the **Classification Sprint**, things get a **bit tougher** as we begin to explore unstructured data. Get ready!

Within this sprint's **Predict project**, we will be **learning essential skills** which will enable you to become productive as future data scientist driving change wherever you are. Here there are **three main tasks that we will be working on**:

Compete to **solve a complex sentiment analysis challenge** using classification techniques

Task 1) Compete in a Hackathon Challenge on Kaggle

In this sprint, you will be introduced to Kaggle, a fantastic online platform for doing data science projects.

Kaggle is an AirBnB for Data Scientists – this is where they spend their nights and weekends. It's a crowd-sourced platform to attract, nurture, train and challenge data scientists from all around the world to solve data science, machine learning and predictive analytics problems. It has over 536,000 active members from 194 countries and it receives close to 150,000 submissions per month.

All Competitions

Active (Not Entered)   Completed   InClass

Getting Started ▾   Default Sort ▾

Titanic: Machine Learning from Disaster
Start here! Predict survival on the Titanic and get familiar with ML basics
Getting Started • Ongoing • 23363 Teams

Knowledge

Digit Recognizer
Learn computer vision fundamentals with the famous MNIST data
Getting Started • Ongoing • 3180 Teams

Knowledge

Real or Not? NLP with Disaster Tweets
Predict which Tweets are about real disasters and which ones are not
Getting Started • Ongoing • 1706 Teams

Kudos

Connect X
Connect your checkers in a row before your opponent!
Getting Started • Ongoing • Simulation Competition • 386 Teams

Knowledge

Micro-Courses

Python
Learn the most important language for data science.

Intro to Machine Learning
Learn the core ideas in machine learning, and build your first models.

Intermediate Machine Learning
Learn to handle missing values, non-numeric values, data leakage and more. Your models will be more accurate and useful.

Data Visualization
Make great data visualizations. A great way to see the power of coding!

Your challenge for this sprint will be to compete in your first Kaggle competition by training a Classification model to predict the sentiment of Tweets related to climate change.

EXPLORE DIGITAL SKILLS

Task 1) Climate change tweet classification Competition Summary

Use this link to enter the competition on Kaggle.

The summary below is taken directly from the competition page:

Many companies are built around lessening one's environmental impact or carbon footprint. They offer products and services that are environmentally friendly and sustainable, in line with their values and ideals. They would like to determine how people perceive climate change and whether or not they believe it is a real threat. This would add to their market research efforts in gauging how their product/service may be received.

With this context, EDSA is challenging you during the Classification Sprint with the task of creating a Machine Learning model that is able to classify whether or not a person believes in climate change, based on their novel tweet data.

Providing an accurate and robust solution to this task gives companies access to a broad base of consumer sentiment, spanning multiple demographic and geographic categories - thus increasing their insights and informing future marketing strategies.

Task 1) Instructions

1. Sign up to Kaggle and **create your own personal profile** (every group member must do this).

2. Enter the Climate Change Belief analysis competition.

3. Register yourself for the competition.
   - Be sure to use name yourself as :<First>_<Surname>_<custom-text> so we know which submission belongs to you!
   - For example: "damian_vather_regression_ninja"

4. **Create a notebook (kernel) on Kaggle** for you to work on.

5. Develop, train and validate your classification model. **Use good practices to version control your code** on github.

6. **Ensure your notebook can produce a valid submission directly to the competition**. Submit this output to Kaggle to be placed on the Challenge Leaderboard (you can do this multiple times). This is a quick guide on how to read data into your kernel, and here you can learn how to output a file.

7. Your Final Github Repository is to be placed in a word doc **Zipped** and **Submitted** on Athena under the **Predict** tab.

Task 1) Rules

- You are required to make a submission on Kaggle.

- You are free to share your ideas with your pod and fellow students, however, **you are not allowed to share your code, solutions, or submissions with other individual**.

- Your Predict **mark will be based upon your score on the Leaderboard**.
  - **Macro F1-score** will be used to score the accuracy of your predictions to the leaderboard
  - Achieving a **Macro F1 of 0.68 or more will score you a passing grade of at least 50%**
  - Achieving a **Macro F1 of 0.78 or more will earn you a grade of 100%**
  - Achieving a **Macro F1 of 0.68 - 0.78 will be scored on a linear scale between 50% and 100%** (e.g. an F1 Macro score of 0.73 will earn you 75%)

- The leaderboard will officially close on a **specified date** and **time**, conveyed to you through an email and the forum.
  No submission will be accepted after the submission deadline specified.

- You will be **required to prove how you obtained a given submission result**. Students who cannot will receive a mark of 0 for the predict.

EXPLORE || DIGITAL SKILLS

Task 2) Document your Findings

As a Data Scientist, you need to **continually communicate** your work and findings to various audiences.

Within this Predict, you will be required to **explain your work to fellow data scientists**.

You will do this in the following ways:

1. As you **develop your solution notebook for Task 1**, you will need to **ensure that your work is fully documented and is reproducible** by a technical individual. To do this, it needs to:

    ○ **Have logical structure**, with an introduction, body and conclusion.

    ○ **Contain essential steps** within your model development process, such as an Exploratory Data Analysis (EDA), data preprocessing, modelling, performance evaluation, and model analysis.

    ○ Be supported with **appropriate visuals and metrics**.

    ○ Separate these sections and **explain your work using Markdown** cells.

    ○ Contain well written code which is sufficiently commented and **meets best coding practices**.

Task 3) Communicate your Findings

2. Give a **Video Conference Presentation** to your EDSA peers and Supervisors.

   ○ **Provide insight into the influential factors and opinions surrounding climate change** to government or a relevant business **using a slide deck** (and/or your Streamlit app).

   ○ Your presentation should **explain your approach, and findings** throughout the entire process, in language that can be understood by non technical people.

   ○ **Make extensive use of graphs and visuals** to provide a data-driven argument. We are giving you a little more freedom this time- we want to see that you are able to connect your insights to business value

   ○ At the end of your presentation, a **panel of supervisors** and audience members will be given an opportunity to **ask technical and non-technical questions** related to your presented work.

   ○ Presentations to last **5-7 minutes, and can be done via video submission, or in a live presentation**

Task 3) Bonus

1. Version control using Github and Git bash

2. The following train will take you through the process. [click here](#)

3. You will submit your github link in a zipped word/ txt file under the Predict tab

Task 4) Bonus

In the previous sprint you created an API using Flask to allow external parties to access your model and make predictions. This time we will be creating something more **user friendly and interactive**.
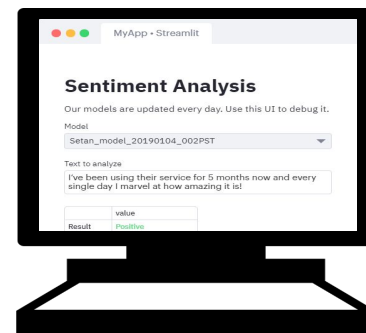
Within the Classification Predict, **we will help you build your own app using [Streamlit's](#) open-source app framework**. You will then go on to host this app within an AWS EC2 instance, making your model and analysis available to your friends, workplace, future clients, and essentially the world at large!
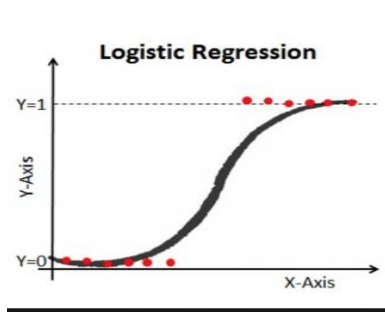
MODEL → **STREAMLIT** → **DEPLOY**

Task 4) Bonus

1. We have created a template repo on GitHub [here](#) to help you setup your app.

2. In each team, one member needs to **Fork the repo***. As a team, you should then **clone the new repo**, and begin collaborating on its development.

3. Send the URL of your team's new GitHub repo to your supervisor.

4. While the template provides a skeleton of what is required to create your app, you must **modify the code** to enable the app to **serve your model/s developed in Task 1** of the predict.

5. The repo contains instructions on how to setup the app on a local computer, as well as on an EC2 instance. Please follow these instructions carefully.

6. **We will view and run the app from the EC2 instance** to ensure that it functions as expected. This task is a great way to show us your creativity and to let your ideas flow into something you're creating - Don't hold back :)

* Forking a repo creates an exact copy of it. The copied repo, or fork, exists separately from the original i.e. changes to the fork don't affect the original repo.