# Portfolio Specification

Jack Foley — C00274246

Supervisor: Greg Doyle

September 27, 2024

# Contents

# 1 Introduction

This is a technical specification document for the Data Science (DS) and Machine Learning (ML) module run by Greg Doyle. As a part of the module, we, the students, are required to create this document to plan the various machine learning and data science projects that we will be working on throughout the first semester. By the end of the semester we are expected to have completed all the projects defined in this document and then present them to the class. This document will be used as a reference throughout the semester to ensure that we are on track to complete the projects on time.

# 2 Portfolio Content

The portfolio will be a simple static website that will contain various information about myself, including links to my CV, contact information, as well as projects that I have completed throughout the semester. The website will be built using HTML, CSS, and JavaScript, if needed. The projects will be displayed on the website one by one.

The headings for the portfolio will be as follows:

- About me
- Employment
- Projects
- Education
- Contact Links

# 3 Available Hardware

ML can use up a lot of computational power, so it is important to have access to machines that are capable enough for running ML tasks. Below is a list of the hardware that I have access to for the duration of the semester. What piece of hardware I use will depend on the task at hand.

- Personal Laptop — Lenovo Thinkpad E16 Gen 1
  - **CPU**: Intel Core i5-13420H (8C/16T)
  - **Discrete GPU**: N/A
  - **Memory**: 24GB DDR4
- Desktop — Custom Build
  - **CPU**: AMD Ryzen 7 5800X3D (8C/16T)
  - **Discrete GPU**: Nvidia RTX 4070

  - **Memory**: 32GB DDR4
  - **Additional Note**: I have full remote access to this machine, therefore it will be used for any tasks that can use a GPU.

# 4 Ideas

## 4.1 Sea Level Prediction

- **Brief Description:** Predict future sea levels using a polynomial regression algorithm.
- **Dataset Source:** Click here
- **Technologies:**
  - Python - Used for writing ML algorithms
  - Pandas - Used for data processing
  - Numpy - Used for any mathematics
  - Matplotlib - Used for plotting data
  - Scikit-learn - Used for ML models
  - Jupyter Notebook - Used for running the code / organisation
  - ML Model: Polynomial regression
- **Detailed Description:** In a world where global warming is steadily heating up the planet, a side affect is the rise of the sea levels across the globe. It is important that we have a prediction of where the sea levels will end up in the future so that we can prepare for the worst. To do this, we can use a polynomial regression algorithm to predict future sea levels. The dataset will have to be cleaned and preprocessed before we can use it to train the model.
- **Hardware:** The hardware I will be using will be my laptop, it is sufficient for this task since it is not too computationally intensive.

## 4.2 Titanic Death Prediction

- **Brief Description:** Predict whether a passenger on the Titanic survived or not.
- **Dataset Source:** Click here
- **Technologies:**
  - Python - Used for writing ML algorithms
  - Pandas - Used for data processing
  - Numpy - Used for any mathematics

  - Matplotlib - Used for plotting data
  - Scikit-learn - Used for ML models
  - Jupyter Notebook - Used for running the code / organisation
  - ML Model: Decision Tree
- **Detailed Description:** The Titanic was a ship that sank in 1912 after hitting an iceberg. The ship was carrying 2,224 passengers and crew, of which 1,502 died. The Titanic dataset is a common task done as a beginner friendly ML exercise. The dataset contains various information about the passengers such as age, class, if they had children, etc. Using a decision tree model, we can predict whether a passenger survived or not. Dataset cleaning and preprocessing will be required before we can train the model.
- **Hardware:** The hardware I will be using will be my laptop, it is sufficient for this task since it is not too computationally intensive.

### 4.3 Inflation Prediction using CPI Data

- **Brief Description:** Inflation prediction using Consumer Price Index (CPI) data and a linear regression model.
- **Dataset Source:** Click here
- **Technologies:**
  - Python - Used for writing ML algorithms
  - Pandas - Used for data processing
  - Numpy - Used for any mathematics
  - Matplotlib - Used for plotting data
  - Scikit-learn - Used for ML models
  - Jupyter Notebook - Used for running the code / organisation
  - ML Model: Linear Regression
- **Detailed Description:** Inflation is the rate at which the general level of prices for goods and services is rising, and subsequently, purchasing power is falling. It is important to have a prediction of where inflation will end up in the future so that we can prepare for the worst. To do this, we can use a linear regression algorithm to predict future inflation. The dataset will have to be cleaned and preprocessed before we can use it to train the model since it has some missing data and/or outliers.
- **Hardware:** The hardware I will be using will be my laptop, it is sufficient for this task since it is not computationally intensive.

### 4.4 OCR on Chars74k Dataset to Predict Characters

- **Brief Description:** Perform Optical Character Recognition (OCR) on the Chars74k dataset using Deep Learning and CNNs.
- **Dataset Source:** Click here
- **Technologies:**
  - Python - Used for writing ML algorithms
  - Pandas - Used for data processing
  - Numpy - Used for any mathematics
  - Matplotlib - Used for plotting data
  - Scikit-learn - Used for ML models
  - Jupyter Notebook - Used for running the code / organisation
  - TensorFlow - Used for deep learning
  - Keras - Used for building the CNN
  - ML Model: CNN
- **Detailed Description:** Optical Character Recognition (OCR) is the recognition of printed or written text characters by a computer. The Chars74k dataset is a dataset that contains 74,000 images of handwritten characters. Using a Convolutional Neural Network (CNN), we can predict what character is in the image.
- **Hardware:** The hardware I will be using will be my desktop, it is sufficient for this task since it is computationally intensive and requires a GPU.

## 5 Deployment

Since the portfolio is a website, it will have to be deployed somewhere. The first option is a Vercel deployment, which is a platform that allows you to deploy static websites for free using the Hobbyist plan. The second option is a GitHub Pages deployment, which is also a platform that allows you to deploy static websites for free. Any of the actual ML projects will be on Google Colab, which will allow users to run the code without having to install any dependencies.

## 6 GDPR

It is important that the datasets used in the projects are GDPR compliant. This means that the data must be anonymised and/or pseudonymised. This will be done in the preprocessing stage of the projects.