Machine Learning Specification

C00270395

Ronan Green

Contents

[Introduction 3](#_Toc177656353)

[Ideas 3](#_Toc177656354)

[Summer Olympic Athletes 3](#_Toc177656355)

[Data Application 3](#_Toc177656356)

[Spotify Top Streamed Songs 4](#_Toc177656357)

[Data Application 4](#_Toc177656358)

[Top Rated IMBD TV Shows 4](#_Toc177656359)

[Data Application 4](#_Toc177656360)

[Sleep Patterns and Lifestyles 4](#_Toc177656361)

[Data Application 4](#_Toc177656362)

[Technologies 5](#_Toc177656363)

[Languages 5](#_Toc177656364)

[Python 5](#_Toc177656365)

[Libraries 5](#_Toc177656366)

[NumPy: 5](#_Toc177656367)

[Pandas: 5](#_Toc177656368)

[Matplotlib: 5](#_Toc177656369)

[Seaborn: 6](#_Toc177656370)

[Scikit-learn: 6](#_Toc177656371)

[Tensorflow: 6](#_Toc177656372)

[Pytorch: 6](#_Toc177656373)

[SciPy 7](#_Toc177656374)

[Tools 7](#_Toc177656375)

[Jupyter Notebook: 7](#_Toc177656376)

[Google Colab: 7](#_Toc177656377)

[RapidMiner: 7](#_Toc177656378)

[Weka 3: 7](#_Toc177656379)

[Anacoda: 7](#_Toc177656380)

[MLFlow: 8](#_Toc177656381)

[Appendix 8](#_Toc177656382)

[Dataset Links 8](#_Toc177656383)

[Summer Olympics 8](#_Toc177656384)

[Spotify Streamed Songs 8](#_Toc177656385)

[Top Tv Shows 8](#_Toc177656386)

[Sleep Health and Lifestyle 8](#_Toc177656387)

[Technology Resources 8](#_Toc177656388)

[Python 8](#_Toc177656389)

[NumPy 8](#_Toc177656390)

[Pandas 8](#_Toc177656391)

[Matplotlib 8](#_Toc177656392)

[Seaborn 9](#_Toc177656393)

[Scikit-learn 9](#_Toc177656394)

[Tensorflow/Keras 9](#_Toc177656395)

[Pytorch 9](#_Toc177656396)

[SciPy 9](#_Toc177656397)

[Jupyter Notebook 9](#_Toc177656398)

[Google Colab 9](#_Toc177656399)

[RapidMiner 9](#_Toc177656400)

[Weka 3 9](#_Toc177656401)

[Anacoda 9](#_Toc177656402)

[MLFlow 10](#_Toc177656403)

# Introduction

This specification report includes ideas I propose form my Machine Learning project. Included in this report are the ideas along with various possible technologies I would use and why I would use them. The appendix will have links to sites used as reference for the datasets and information on the technologies.

# Ideas

## Summer Olympic Athletes

This project includes a comprehensive list of athletes who have participated in the summer Olympics since in inaugural year on Athens in 1896 to the 2024 Paris Olympics. The data it includes ranges from the athletes name, sex, country, year city, sport, event and medal won. This has more than 130,000 unique sets of data.

### Data Application

* What this data set could be used for is predicting which country will win the most medals.
* Which country produce the most efficient athletes for certain sports.
* Countries that have been doing or worse in recent years and where they could end up in years to come.
* We can also see which country has more proficient males or more proficient females.

## Spotify Top Streamed Songs

This project idea includes the details of the most streamed songs of 2023. It includes details such as artist, song name, bpm of the song, the key the song is mainly in, the number of playlist it is in the day, month and year of release, number of artist on the song and ranking in charts. This dataset has 943 rows.

### Data Application

* using this data we could see what bpm makes the most popular songs.
* What key listeners find most appealing.
* See what song name format is the most popular
* upcoming and dying artists and possibly the transition of artists between different bpm and keys
* What time of year yields the most successful releases.

## Top Rated IMBD TV Shows

This project includes the data from the top 250 TV shows of all time according to IMBD. It give rankings names, release dates, run time, rating tag and rating. This has 250 rows of data.

### Data Application

* This data can be used to see what length of show is the most popular in what years.
* If the runtime of shows is increasing or decreasing due to popularity,
* How many episodes is optimal for popularity.
* What rating tag is the most featured and how these ratings tags correlate to the ratings itself.

## Sleep Patterns and Lifestyles

This dataset includes a wide array of information pertaining to a persons gender, age, occupation, sleep duration, quality of sleep, physical activity level, stress level, BMI category, blood pressure, heart rate, daily steps, sleep disorder. This has 374 rows of data.

### Data Application

* This data can be used to predict the length and quality of sleep for each age group.
* What jobs end up having a positive/negative impact on sleep length.
* How much physical activity and sleep plays into BMI, blood pressure and BPM
* How stress level relate to sleep length and quality.

# Technologies

## Languages

### Python

Python has become the de facto language for machine learning. This id due to many factors such as its simplicity, versatility and its wide range of libraries and frameworks.

## Libraries

Python gives access to many useful libraries for machine learning. These libraries include the following:

### NumPy:

NumPy is a python library that enhances arrays making them faster than regular python arrays. This library also provides functions for working with linear algebra, fourier transform and matrices. NumPy also supports a wide arrange of hardware and computing platforms. Due to it being a python library the syntax is high level accessible and productive for all level programmers.

### Pandas:

Pandas is a python library used for working with datasets. It can analyse the data and has functions for cleaning exploring and manipulating data. Panda is able to get a lot of information form datasets and gives you answers about the correlation between data, average values, and making this data readable and relevant. Pandas is suited for working with spreadsheets or SQL tables. Pandas is built on top of the NumPy library meaning there is cross over in the structures used. The data used can be used in plotting functions for Matplotlib, statical analysis in SciPy and ML algorithms in Scikit-learn. It can also handle missing data, data visualization and can perform split-apply-combine operations on data sets.

### Matplotlib:

Matplotlib is a python library that can be used for creating static, animated or interactive visualizations. It can generate a wide range of graphs such as line, scatter, bar, histograms, pie charts and more. Along with supplying the tools for displaying graphs it allows customization of these graphs. These graphs are of professional quality. It can work seamlessly with NumPy allowing data from arrays to be plotted easily. It also works with Seaborn, Pandas and basemap.

### Seaborn:

Seaborn is a python library that builds on top of Matplotlib and closely with Pandas to make statistical graphics in python. Seaborn uses plotting functions to operate on data frames and arrays containing whole datasets. The dataset-orientated API lets the focus be on different elements of the plot and not the details on drawing the plot. Seaborn provides various plots such as relational plots, categorical plots, distribution plots, regression plots, matrix plots, multi-plot grids.

### Scikit-learn:

Scikit-learn is a python library that provides diverse algorithms for classification, regression, clustering, and dimensionality reduction. The library is build using other languages like SciPy and NumPy and closely connected with Pandas and Seaborn. It provides simple and efficient tools for data mining and data analysis.

### Tensorflow:

TensorFlow is a python library for numerical computation using data flow graphs. It provides functions for building and training deep learning models as it facilitates the creations of computational graphs and efficient execution on a wide range of hardware platforms. A few features TensorFlow has are Antidifferentiation, Eager execution, Distribute, Losses, Metrics, TF.nn, Optimizers. TensorFlows’s APIs use Keras to allow users to make their own machine-learning models and also helps load the data to train the model.

#### Keras:

Keras is an API that runs on top TensorFlow. It is a high-level API used for training and building neural networks. Its allows you to, with minimal code, to build, train and deploy deep learning models. It is known for its user friendly interface that allows all level of coders to have viable access to the API. Keras has good Extensibility and Customizability providing help for creating custom layers, loss functions and preprocessing task. Allows for complex architectures and provides subclassing to write models from scratch. With it providing sequential along with functional APIs it allows ease of use working with single input and output models as well as multiple input and output models.

### Pytorch:

Pytorch is a python library that provides many convenient tools to help build neural networks and train them efficiently. PyTorch is built using tensor which is similar to NumPy. It uses the autograd library for automatic differentiation, this computes gradients of tensor. The graphs are made during run-time allowing for dynamic changes to graphs to be made.

### SciPy

SciPy is a python library that is used for scientific computation and uses NumPy. It provides utility functions for optimisation, stats and signal processing. Scify is very similar to NumPy but has added functions and is optimised. The vast amount of algorithms it provides ranges from optimisation and integration to algebraic and differential equations along with many more.

## Tools

### Jupyter Notebook:

Jupyter notebook is an web application used to create and share documents that contain live code, visualizations and text. It is commonly used with data science and machine learning.

### Google Colab:

Google Colab is like google docs but for python code. It is a cloud based service that allows the user to write and run code in a jupyter notebook environment. Google colab allows the user to take advantage of powerful CPUs and GPUs without having the hardware yourself. It can be used to write and execute code, develop models, and collaborate with other devs.

### RapidMiner:

Rapidminer is a comprehensive data science platform with visual workflow design and full automation. It is one of the most popular data science tools. RapidMiner is used for data extraction, data mining, deep learning, machine learning, and predictive analysis.

### Weka 3:

Weka Is used to provide a comprehensive suits of tools for data analysis and predictive modelling. It helps users to analyse large datasets and applies various machine learning algorithms for task such as clustering, classification, regression, association rule data mining and data processing.

### Anacoda:

Anaconda is a distribution of the python and r programming languages for scientific computing that aims to simplify package management and deployment. Packages in Anaconda are managed by the packet management system Conda.

### MLFlow:

MLFlow provides an array of tools that are aimed to simplify the ML workflow. The functionalities of MLflow are rooted in several components such as Tracking, Model Registry, MLflow deployments for LLMs, Evaluate, Prompt Engineering UI, Recipes and Projects.

# Appendix

## Dataset Links

### Summer Olympics

<https://www.kaggle.com/datasets/stefanydeoliveira/summer-olympics-medals-1896-2024>

### Spotify Streamed Songs

<https://www.kaggle.com/datasets/nelgiriyewithana/top-spotify-songs-2023>

### Top Tv Shows

<https://www.kaggle.com/datasets/khushipitroda/imdb-top-250-tv-shows>

### Sleep Health and Lifestyle

<https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset>

## Technology Resources

### Python

<https://pythonbasics.org/why-python-for-machine-learning/>

### NumPy

<https://numpy.org/>

<https://www.w3schools.com/python/numpy/numpy_intro.asp>

### Pandas

<https://www.geeksforgeeks.org/introduction-to-pandas-in-python/>

<https://www.w3schools.com/python/pandas/pandas_intro.asp>

### Matplotlib

<https://matplotlib.org/>

<https://www.geeksforgeeks.org/python-introduction-matplotlib/>

### Seaborn

<https://seaborn.pydata.org/tutorial/introduction>

<https://www.geeksforgeeks.org/introduction-to-seaborn-python/>

### Scikit-learn

<https://datagy.io/python-scikit-learn-introduction/>

<https://www.geeksforgeeks.org/learning-model-building-scikit-learn-python-machine-learning-library/#features-of-scikitlearn>

### Tensorflow/Keras

<https://www.geeksforgeeks.org/introduction-to-tensorflow/>

<https://www.geeksforgeeks.org/what-is-keras/>

### Pytorch

<https://www.geeksforgeeks.org/getting-started-with-pytorch/>

<https://builtin.com/machine-learning/pytorch>

### SciPy

<https://scipy.org/>

<https://www.w3schools.com/python/scipy/scipy_intro.php>

### Jupyter Notebook

<https://realpython.com/jupyter-notebook-introduction/>

### Google Colab

<https://bytexd.com/what-is-google-colab-a-beginner-guide/>

### RapidMiner

<https://mindmajix.com/rapidminer-tutorial>

<https://www.analyticsvidhya.com/blog/2021/10/intro-to-rapidminer-a-no-code-development-platform-for-data-mining-with-case-study/>

### Weka 3

<https://www.geeksforgeeks.org/introduction-to-weka-key-features-and-applications/>

### Anacoda

<https://en.wikipedia.org/wiki/Anaconda_(Python_distribution)>

### MLFlow

<https://mlflow.org/docs/latest/introduction/index.html>