**UNIVERSITY OF LONDON INTERNATIONAL PROGRAMMES**

**BSc Computer Science and Related Subjects**



**CM3070 PROJECT**

**FINAL PROJECT REPORT**

Student Name: Chai Jia Min Shermin

SIM Student Number : 10222083

Date of Submission: 25 September 2023

Supervisor Name: Yeo Zhiyu Aren

Table of Content

[CHAPTER 1: INTRODUCTION 3](#_Toc146563502)

[Project Concept 3](#_Toc146563503)

[Motivation 3](#_Toc146563504)

[Aims and Objectives 4](#_Toc146563505)

[Deliverables 4](#_Toc146563506)

[CHAPTER 2: LITERATURE REVIEW 5](#_Toc146563507)

[Research Papers 5](#_Toc146563508)

[Supervised Learning Models 5](#_Toc146563509)

[Neural Network Models 6](#_Toc146563510)

[Evaluation Metrics 9](#_Toc146563511)

[Conclusion 9](#_Toc146563512)

[CHAPTER 3: PROJECT DESIGN 10](#_Toc146563513)

[Domain and Users 10](#_Toc146563514)

[Olympic Athletes Dataset 10](#_Toc146563515)

[Design Choice 11](#_Toc146563516)

[Overall Structure 11](#_Toc146563517)

[Technologies and Methods 12](#_Toc146563518)

[Testing and Evaluation Plan 13](#_Toc146563519)

[CHAPTER 4: FEATURE PROTOTYPE (DEVELOPMENT) 14](#_Toc146563520)

[Feature Engineering 14](#_Toc146563521)

[Data Cleaning and Exploration 14](#_Toc146563522)

[Execution and Evaluation of Binary Classification 14](#_Toc146563523)

[Execution and Evaluation of Artificial Neural Network 15](#_Toc146563524)

[CHAPTER 5: EVALUATION 16](#_Toc146563525)

[Results from Binary Classification and Artificial Neural Network 16](#_Toc146563526)

[Evaluation 16](#_Toc146563527)

[Further Work 18](#_Toc146563528)

[CHAPTER 6: CONCLUSION 19](#_Toc146563529)

[CHAPTER 7: APPENDICES 20](#_Toc146563530)

[CHAPTER 8: REFERENCES 25](#_Toc146563532)

# CHAPTER 1: INTRODUCTION

Project Template: Deep Learning on a Public Dataset (CM3015 Machine Learning and Neural Networks)

Project Concept: Prediction Model for Choosing the Best Swimmers for future Olympic Games

## Project Concept

Modern Olympics are leading international sporting events which feature various summer and winter sporting events. Singapore’s first Olympic gold medalist was Joseph Schooling where he achieved a gold medal in the 100m butterfly at the 2016 Olympics. His accomplishment instilled a sense of national pride within Singapore citizens, inspiring them to partake more in swimming. Hence, it will be in the nation’s best interest to try and replicate this effect.

This project will be analyzing the relationship between the chances of an Olympic swimmer winning a medal and their physical characteristics with the use of Machine Learning. The physical characteristics that affect the Olympic swimmer the most will also be determined. The physical characteristics that I will be looking at are Age, Height, and Weight, which are all available in the dataset. Furthermore, swimming was chosen as it is an individual sport where the attainment of a medal will not be influenced by other teammates.

## Motivation

Pure hard work will not allow athletes to be good in their sports, much less guarantee a win in the sports event. Many other factors play a part in training an athlete. This includes physical characteristics, diets, and training regimens. Thus, for this project I have decided to focus on physical characteristics as the measurement of the physical characteristics can be determined immediately. All the measurement of the physical characteristics can be found in the dataset. Deep learning algorithms using neural networks can determine the ideal physical characteristics for a swimmer to have by taking the average measurement of swimmers who have attained a medal during the past Olympic Games. The physical characteristics that affect the chances of swimmers achieving a medal the most can also allow coaches to take note when recruiting athletes for the team. This can help coaches to come up with training regiments specific to each individual swimmers in the team.

## Aims and Objectives

The main aim of this project is to analyze the dataset and create a prediction model that would help accurately predict a swimmer’s potential in receiving an Olympic medal. This can help swimming coaches select the best swimmers among their team to take part in future Olympic Games to claim more medals and accomplishments for the country. In this report, I will be answering a few questions to aid my research:

* Will physical characteristics play an important role in identifying potential Olympic medalists?
* If so, among height, weight, and age, which of them will be the most crucial?

## Deliverables

2 deep learning prediction models will be delivered to predict the chances of a swimmer attaining an Olympic medal based on their physical characteristics. The first model will use the Binary Classification method since there are 2 main outcomes: Medal and No Medal. The second model will use the Artificial Neural Network to train and test the algorithm. We will then compare the results of both models and determine which method can yield the best results.

The prediction models will train the existing Olympic records, and I will test the prediction models on the Tokyo 2022 Swimming Event. This will allow us to analyze the accuracy of how the physical characteristics affect the performance of the swimmers attaining a medal. This is done by comparing the physical characteristics of swimmers who attained a medal against the swimmers who did not manage to attain a medal in the Tokyo 2022 Swimming Event.

# CHAPTER 2: LITERATURE REVIEW

## Research Papers

I have found several related research papers elaborating on the relationship between an athlete and their medal outcome. For the first research paper, algorithms using deep neural network and feature extractions are used to predict the outcome of sports performance and the factors that influences them. This helps to establish the relationship between performance and their training level [1]. The second research paper mentions how the psychological and emotional changes affect an athlete’s sports performance. The relationship between them was predicted using behaviour recognition algorithm, where their anxiety level, self-confidence and performance were measured. Those factors were then used to predict their performance for each game and see if their psychological and emotional changes affect their ability to achieve a medal. Even though that is not the scope of my project, it is similar in determining whether an attribute affect the chances of an athlete attaining a medal [2]. The last research paper is the most similar to my project scope where the sport performance on specific abilities by female handball players were analyzed. Linear regression, decision tree and neural networks were used to predict their performance in each specific ability [3]. Thus, we will be looking at various similar approaches of Machine Learning and Deep Learning by evaluating their advantages and disadvantages before choosing the best method to design the final prototype in this project.

## Supervised Learning Models

Supervised Learning allows an algorithm to make predictions based on the input data in a labelled dataset. Since the Olympic Athletes Dataset has labelled data, where it is clearly stated if an athlete has obtained a medal, we will be using supervised learning method to train the algorithm for it to accurately predict the outcomes.

Classification

Classification is a type of supervised machine learning method that uses the input data to predict the class label. Since there are only 2 distinct classes of outcome (Medal or No Medal), Binary Classification will be used. There are several classification models, but we will be looking at the Random Forest Classifier and the Decision Trees Classifier.

Random Forest Classifier

The Random Forest Classifier uses random feature selection techniques to build groups of decision tree models based on the new datasets gotten from the original data. These newly built datasets are randomly selected and merged based on the original data. Each dataset will correspond to a new decision tree. The test data will be passed along each decision tree and the predictions will be noted down. The majority voting will then decide the class label of the test data. In this scope, each decision tree will be exposed to different features (physical characteristics) and a different sample from the dataset. The decision tree will then predict the medal outcome of the athlete based on the information they have. The random forest classifier will then count the number of predictions for each class before choosing the most popular prediction [4]. Random Forest Classifiers are commonly used in the healthcare and finance sector as the accuracy are important in these sectors. However, the random forest classifier can be computationally expensive and require more time to train and test the dataset.

Decision Tree Classifier

The Decision Tree Classifier is a hierarchical form of instances and attributes that is used for both classification and prediction problems [5]. It has a tree structure that contains internal nodes, branches, and leaf nodes. The internal nodes represent a test on an attribute (physical characteristics), the branch represent the possible outcome of the test, and the leaf node represent the class output (medal or no medal) [6]. The root node will ask a question on the athlete physical characteristics and the branches (yes or no) will be the answer to that question. Each answer will lead the decision tree classifier to another node. Once the last node is reached, it will give the outcome of the prediction (medal or no medal). Decision tree is easy to interpret, and it requires very little computational time [5] but it is prone to overfitting [6]. Overfitting is when a model gives accurate predictions for the training data but fails to produce a generalized pattern that could occur in the new data [7]. Since we have a relatively small dataset and low processing power, using the decision tree classifier will be sufficient for this project.

## Neural Network Models

A neural network is a method in Artificial Intelligence that teaches computers to process data in a way that is inspired by the structure of the human brain [8]. It is a type of machine learning process, called Deep Learning, that uses interconnected nodes or neurons in a layered structure [8]. Neural networks take in and trains themselves to recognize the patterns presented in the data before predicting the output for a set of new data. Neural networks are made up of layers of neurons, which are the core processing unit of the network. There are 3 main layers in the neural network: Input Layer, Hidden Layer and Output Layer. The input layer receives the input, the output layer predicts the final output, and the hidden layer performs most of the computations required by the network. Neurons in each layer are connected to neurons in the next layer through channels. A numerical value is assigned to each channel, known as weights. The input is multiplied to the corresponding weights and their sum is send as input to the neurons in the hidden layer. Each neuron is then associated with a numerical value, known as the bias, which is then added to the input sum. The input sum will then be passed through a threshold function called the Activation Function. The result of the activation function determines if the neuron will be activated. Only an activated neuron will be able to transmit data to the neurons in the next layer. The neuron with the highest value at the output layer will then be determined through its probability, which will be the prediction of the neural network. We will be looking at Convolution Neural Network and Artificial Neural Network.

Convolution Neural Network

Convolution Neural Network is a deep learning algorithm that can take the input data, assign its importance for the different aspects in the data and differentiate the input from one another. Hidden layers are used to build a strong convolution neural network model. This also includes the convolution layer, the pooling layer, and the fully connected layer [9]. Activation functions such as ReLu and Softmax are also commonly used. Dropout layers are also used in convolution neural network. The convolution and pooling layer help to activate features from the input and simplifies the output respectively, which helps to reduce the number of parameters that the networks need to learn [10]. The fully connected layer, also known as the Dense Layer, takes the output and classifies the input data into the target classes depending on their features. Activation Functions then help the model to learn the complex mapping between the inputs [11].

Convolution neural networks tend to have efficient image processing even though it may require high computational processes. It also requires a larger dataset to work with to produce a higher accuracy rate. This is because they learn to recognize the patterns in the image by analyzing many images. If the dataset is too small, it may cause the problem of overfitting as it will be specialized to the training dataset, but it performs badly on a new dataset [12].

However, convolution neural networks are mainly used for pattern recognition in images by using matrix multiplication to identify patterns within an image. Since the Olympic Athlete dataset does not require image recognition, convolutional neural networks will not be the best method to use in this project.

Artificial Neural Network

Artificial Neural Network are meant to mimic the behaviour of a brain. The brain can learn things about its environment using an array of neurons, or nodes, connected in a network. Artificial Neural Network contain units where are arranged in a series of layers. Units are interconnected from one layer to another. Each of the connections has weights that determine the influence of one unit on another unit [13]. The network can either be in single layer or multiple layers. The layers are meant to act as filters that differentiate the input that are fed into the system. Artificial Neural Network contains 3 main layers: the Input Layer, the Hidden Layers, and the Output Layer.

The input layer accepts input in different format which the neural network needs to analyze. The input is then passed on to the hidden layers that transform the input into data before passing it into the output layer [13]. In the hidden layer, each neuron receives input from the previous layer neurons, computes the weighted sum, and sends it to the neurons in the next layer. Different weights are added to each input before passing them to the next layer which allow the connections to be weighted, optimizing the model performance[13]. As the data transfer from one unit to another in the different layers, the model will learn more about the data which allow the model to train and produce an accurate output.

Artificial Neural Network can work with incomplete knowledge since the data can produce output even with incomplete information [5]. This might affect the performance of the model, but it also depends on the importance of the incomplete information. However, the network needs to be trained first before it can operate on the model. This requires high processing time especially if a large dataset is involved as well [14].

Artificial Neural Network works best on tabular or text data due to its ability to learn from a sample data. It is also able to capture non-linear relationship between the inputs and generalized them, allowing it to predict on unseen data[3]. In this project, we will be using Artificial Neural Network to train and test our model.

## Evaluation Metrics

Evaluation metrics helps to evaluate the model and determine the best results. The most common evaluation metrics are listed below. However, we will only be looking at a few evaluation metrics in this project.

1. Accuracy = *TP + FN / TP + FP + TN + FN*
2. Precision = *TP / TP + FP*
3. Recall = *TP / TP + FN*

Classification Accuracy and Confusion Matrix can also be used to check how good the predictions are for the Training and Testing Set. Classification Accuracy is a metric that summarizes the performance of a classification model as the number of correct predictions divided by the total number of predictions. However, classification accuracy alone can be misleading if there is an unequal number of observations in each class so calculating a Confusion Matrix can show the ways in which the classification model is confused when it makes predictions. Confusion Matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class.

## Conclusion

There are many different methods suggested in the different articles regarding the prediction of sports performance, which will help to aid this project. Some key considerations to take note of in this project is classifying the dataset using Binary Classification, building the Artificial Neural Network with more hidden layers to optimize the performance of the model, and evaluate the results using the different evaluation metrics deemed appropriate.

# CHAPTER 3: PROJECT DESIGN

## Domain and Users

This project aims to aid swimming coaches in selecting the best athletes to represent the team for future Olympic Swimming Events. The main domain of data will be the current and past records of swimmers that participated in the Olympic Swimming Events. The demographic of users will be focused on swimming coaches, but this project can be used to aid any other sports that other coaches may want to take advantage of, if there are enough data for this project to train and test. It can also be used in other sporting events, not limited to the Olympic Games. Coaches may use the Machine Learning model to input the athlete’s physical characteristics and the model will be able to predict the chances of the athlete winning a medal. Even though it is not accurate as there are many other factors affecting the performance of an athlete, this model can serve as a gauge for the coaches.

## Olympic Athletes Dataset

In the Olympic Athletes Dataset, there are information about each athlete and the event that they have competed in. The team that the athlete represents and the Olympic committee in charge of them are available in the dataset. The Olympic game information such as the year, season and city that the Olympic game was held in are also available in the dataset. The dataset also contains the sports and event that each athlete took part in. Most importantly, the gender, age, height, weight and medal outcome of each athlete are also in the dataset, which we will be using to train and test our Machine Learning Models. However, since we do not need so much information, we will be cleaning up the dataset in the first part of the model. This allows us to remove columns that we do not need, where we will end up with only the team, gender, age, medal and the physical characteristics of the athlete. However, we notice a large discrepancy between the Medal and No Medal class, so we tried to narrow down the dataset furthermore. We decided to extract only male swimmers that take part in swimming events representing the United States, China, Singapore and South Korea. This leaves us with a size of 1051 records in this dataset, containing 527 records of Medal class and 524 records of No Medal class.

## Design Choice

Data will be taken from Kaggle.com, where all the past records of the details of each athlete and their respective Olympic Events has been consolidated into a dataset. The design choices cam be broken into different phrases: Extracting Data, Pre-processing Data, Implementing Classifiers and Machine Learning Models, and Evaluation of Techniques and Results. We will first extract the data we want from the original dataset before we pre-processed it for the machine learning models to train and test the data. Pre-processing includes cleaning up the data, selecting the columns we want, and removing duplicated and empty records from the dataset. This helps to give a more accurate dataset for the models to train. We will then split the model up into training and testing dataset with 80% of the data used for training and 20% of the data used for testing. We will use Binary Classification as the first Machine Learning method to calculate the classification accuracy and confusion matrix for each of the physical characteristics. For the second Machine Learning method, we will be using Artificial Neural Networks where we will use layers to train and test the data. We will then compare both methods and see which method yield the better result.

## Overall Structure

Google Colaboratory (Colab), a cloud based Jupyter Notebook environment, will be used to build this project. We will be using the Python language in this project, and Colab will then run the code through the browser. It does not require any setup to be done, which allows for easy execution of the models.

As mentioned, the data will be taken from Kaggle.com where it will be preprocessed first before we can use it to run the model. Binary Classification and Artificial Neural Networks are the two models that will be used in this project. The data will be run by both models by training the data before testing them.

Evaluation metrics will be used to gauge the performance of the model. The Binary Classification will use metrics such as Classification Accuracy and Confusion Matrix while the Artificial Neural Network will use evaluation metrics such as Accuracy, Precision and Recall Rate. By using the evaluation metrics, we can then determine the best optimal model for this project.

## Technologies and Methods

Python: It has the required libraries that this project needs to run both the Binary Classification and the Artificial Neural Network models such as SKLearn, Matplotlib, TensorFlow and Keras. We can use those libraries to train our model and display the appropriate visualization for those results as well. Other libraries that are deemed appropriate can also be easily imported and use in Python due to the large number of libraries available.

Binary Classification: The data for Age, Height and Weight will be run separately using the same model. We will first split the data into training and testing data before creating a Decision Tree Classifier. We will use the decision tree to train the data and calculate the classification accuracy based on the result before plotting a confusion matrix to get a visual representation of the classification. We then test the decision tree classifier on the test data for the same physical characteristics and calculate the classification accuracy before plotting the confusion matrix. This will be done for all the physical characteristics separately. The accuracy, precision and recall rate for the physical characteristics will then be appended into a table. We can then easily compare the results from the training dataset and the testing dataset. The execution of each segment will be explained in the next chapter. However, when the results for the classification accuracy for the univariate binary tree was analyzed, it was always below 70%. To improve the score, a multivariate decision tree was developed using all 3 variables as predictor. We will then append all the values that we have calculated thus far into a table and compare its result.

Artificial Neural Network Model: We will create the object using a class of Keras named Sequential. We will then create 2 hidden layers, 1 for input and 1 for output, using the Dense class which have 2 inputs: Units and Activation. Since this project focuses on a binary classification problem, we will only need to allocate one neuron to output the result, while using sigmoid activation function in the output layer. We can then compile our Artificial Neural Network using 3 inputs: Optimizer, Loss, and Metrics. We will finally train and fit our ANN by inputting 4 inputs: X\_train, Y\_train, batch sizes and epochs. We can then test the Artificial Neural Network on our test data and allow it to predict the results. The execution of the model will be explained in the next chapter.

## Testing and Evaluation Plan

For Binary Classification, the evaluation metrics used are Confusion Matrix and Classification Accuracy.

Confusion Matrix: There are 4 main outcomes namely: True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). The outcomes will be displayed in the confusion matrix.

Classification Accuracy: It calculates the accuracy of the model by comparing the actual test result and its predicted value. It takes the total number of true values and divides it by the total number of observations. A high Classification Accuracy value indicates a better result of the model.

For Artificial Neural Networks, the evaluation metrics used are Accuracy, Precision, Recall and Area Under Curve (AUC).

Accuracy: It calculates the accuracy of the model by comparing the true values and its predicted value. It takes the total number of true values and divides it by the total number of observations. A high Accuracy value indicates a better result of the model.

Accuracy = (TP + TN) / (TP + FP + FN + TN)

Precision: It calculates the precision rate of the model by comparing the number of true positives and its predicted value. It takes the number of true positives and divides it by the total number of positives predicted. A high Precision value indicates a better result of the model.

Precision = TP / (TP + FP)

Recall: It calculates the recall rate of the model by comparing the number of true positive and the total number of positive outcomes. It takes the number of true positive and divides it by the total number of actual positive outcomes. A high Recall value indicates a better result of the model.

Recall = TP / (TP + FN)

AUC: It calculates the area under the Receiver Operating Characteristics (ROC) curve and provides an aggregate measure of performance for the model. It measures the ability of the model to differentiate between the classes. When the value of AUC is nearer to 1 than 0, it indicates a better result of the model. When AUC = 1, the model can differentiate between all the positive and negative classes correctly.

# CHAPTER 4: FEATURE PROTOTYPE (DEVELOPMENT)

## Feature Engineering

Data Cleaning and Exploration, Execution of Algorithm and Evaluation are important features that are necessary to evaluate the performance of the models accurately. Data Cleaning and Exploration allows the dataset to be cleaned by removing any non-applicable values such as missing data or duplicated data. Execution of Algorithm splits the data into the training data and the testing data before training them using the model and evaluating the performance.

## Data Cleaning and Exploration

Cleaning up of data allow us to have a more accurate dataset based on what the project is supposed to achieve. In this case, we remove all the columns that we do not need by extracting the specific columns we want (Figure 1). Next, we need to replace all the different medal types such as Bronze, Silver and Gold to Y class and all other NaN rows to N class (Figure 2). This helps us to differentiate which athlete won a medal. We then drop all the empty rows with no input and change the object type of Medal to ‘Category’ (Figure 3a and 3b). However, the difference in value between the 2 classes was too huge so we need to even out the Y and N classes. This was done by extracting data of male swimmers from the United State, South Korea, China, and Singapore (Figure 4a and 4b).

## Execution and Evaluation of Binary Classification

We need to split the data up to the training data and the testing data according to the different physical characteristics (Figure 5). Age will be the first physical characteristic that the model will be training and testing on. After splitting the dataset, we will create the Decision Tree Classifier to fit the training data (Figure 6). We then use the decision tree to predict the result of the training data before calculating the accuracy, precision, and recall rate of the prediction (Figure 7). A confusion matrix is also plotted for easy visualization of the result. We then use the same decision tree to test it on the testing data before calculating the accuracy, precision, and recall rate of those predictions (Figure 8). A confusion matrix will also be plotted. We will store the accuracy, precision, and recall rate of both the training and testing predictions in a list. We then repeat the whole training and testing model on the Height and Weight dataset. However, when the score was analyzed, the accuracy rate for the univariate binary tree was almost always below 70%. Hence, a multivariate decision tree using all 3 physical characteristics as predictor was developed to try and improve the score. The multivariate decision tree had a better score compared to the univariate binary tree which shows that much more data points lie within the category of correct predictions (Figure 9). Based on the score table, Height (Test) has the highest accuracy, precision, and recall rate of 0.74, 0.74, and 0.72 respectively. This indicates that physical characteristics are important predictors of medal classes since they affect the chances of athletes attaining a medal. With Height being the most crucial physical characteristics out of the 3, Weight comes in next while Age comes in last based on the score in the table (Figure 10).

## Execution and Evaluation of Artificial Neural Network

We need to split the data up to the training data and the testing data according to the different physical characteristics (Figure 5). Age will be the first physical characteristic that the model will be training and testing on. We need to use Label Encoder to convert the categorical values of the Medal Class into numerical form (Figure 11). Weights of the class is also calculated to use it to fit it into the output model later (Figure 12). For this Artificial Neural Network (ANN), the optimizer and loss function used will be RMSprop and Binary Cross-entropy respectively. The evaluation metrics used are Accuracy, Precision, Recall and AUC (Figure 13). We will start initializing the ANN model by using Sequential from Keras. We can then start adding in hidden layers for it to train the data. In this case, we have added 5 layers for the data to be trained before the last output layer. Since this is a binary classification problem, only one neuron is allocated to output this result and the activation function should always be Sigmoid (Figure 14). The model will then be complied before having the model run through the dataset (Figure 15). The number of epochs define the number of times the model will run. The results of the prediction will then be printed as an output (Figure 16). The model will then be run two more times with the Height and Weight dataset. Based on the results, Weight yields the best results with an accuracy, precision, recall and AUC of 0.65, 0.67, 0.56 and 0.69 respectively. This is followed by Height then Age. However, the results obtained are less accurate as compared to the Binary Classification Model as the Binary Classification model has a higher accuracy rate for Height and which is why I will be using the Binary Classification model to prove my aims and objectives for this project.

# CHAPTER 5: EVALUATION

## Results from Binary Classification and Artificial Neural Network

Key: BC = Binary Classification, ANN = Artificial Neural Network

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Algorithm Used** | **Dataset** | **Accuracy Rate** | **Precision Rate** | **Recall Rate** |
| BC | Age Train | 0.59 | 0.60 | 0.54 |
| ANN | Age Train | 0.54 | 0.77 | 0.11 |
| BC | Age Test | 0.57 | 0.58 | 0.50 |
| ANN | Age Test | 0.55 | 0.54 | 0.49 |
| BC | Height Train | 0.68 | 0.67 | 0.72 |
| ANN | Height Train | 0.65 | 0.73 | 0.47 |
| BC | Height Test | 0.74 | 0.74 | 0.72 |
| ANN | Height Test | 0.62 | 0.77 | 0.32 |
| BC | Weight Train | 0.68 | 0.70 | 0.66 |
| ANN | Weight Train | 0.64 | 0.69 | 0.51 |
| BC | Weight Test | 0.65 | 0.67 | 0.57 |
| ANN | Weight Test | 0.65 | 0.67 | 0.56 |

The results obtained from Binary Classification had a high accuracy rate throughout the different tests as compared to the Artificial Neural Network.

## Evaluation

The main aim of this project is to create a prediction model that can accurately predict a swimmer’s potential in receiving an Olympic medal based on their physical characteristics. We used Binary Classification and Artificial Neural Networks to analyze the likelihood of the Olympic swimmers attaining a medal. A binary classification model was chosen as it allows us to directly categories the outcome into 2: Medal or No Medal. The artificial neural network model was chosen as it allows us to analyze any non-linear and complex relations between the physical characteristics and the chances of them attaining a model.

Different performance metrics were used to gauge the effectiveness of both models. The binary classification model used classification accuracy and confusion matrix, while the artificial neural network model used accuracy, precision, and recall rate to assess the models. Both methods were trained to predict the medal outcome and which physical characteristics of the Olympics swimmers affected the chances of them attaining a medal the most. The Binary Classification model can also predict the medal outcome based on the physical characteristics given in the Tokyo 2022 Event, achieving the aims and objectives of this project.

Using the Binary Classification Model and the Artificial Neural Network Model, we were able to use those models to predict the medal outcome of each athlete based on their physical characteristics. This can be concluded as the model have a high rate of accuracy which shows that the model manages to correctly predict the medal outcome of the athletes based on the value of each physical characteristics given to them. Furthermore, we were able to analyze the results and conclude the importance of each physical characteristics when predicting the medal outcome. This allows future coaches to make use of this model and predict if their athlete can get a medal in the swimming competition based on their physical characteristics. However, there are other factors affecting an athlete’s medal outcome, so it is important to consider the other factors as well.

There are limitations to both methods as there are other factors such as training regimens and diets which will affect a swimmer’s ability to attain a medal that are not considered in this project. Such factors can be used in further development of the project in future work. Coaches will then be able to use such a model as the accuracy for each model increase when more factors affecting the Olympic swimmers are taken into considerations.

This project display evidence of originality as such model for predicting the chances of an Olympic swimmer based on their physical characteristics has not been researched yet. Even though the concept of sports performance and their achievement has been the topic of different research, the focus on the relationship between Olympic swimmers and their physical characteristics such as Height, Weight and Age has not been explored before.

## Further Work

We used the Binary Classification models to predict the outcome of the medal based on the physical characteristics given to each model in a Tokyo 2022 Swimming Event. The results are presented below:

Tokyo Predictions for Binary Classification

A table with numbers and names

Description automatically generated

*Figure 17: Tokyo Medal Outcome Prediction based on Binary Classification*

Accuracy Rate: 0.64

Precision Rate: 0.75  
Recall Rate: 0.5

This proves that coaches can input their athlete’s physical characteristics into the Binary Classification model and the model will be able to predict the medal outcome of that athlete since the model has an accuracy rate of 0.64.

The Artificial Neural Network model did not manage to predict the medal outcome for new data as I was not able to implement that segment due to my poor knowledge in Artificial Neural Network. This can be done as a future work if I had more time and resources to do so. However, the Binary Classification model was able to predict the medal outcome for the athlete, which allows me to achieve the aims and objectives of this project.

# CHAPTER 6: CONCLUSION

This project aims to predict the medal outcome of an Olympic swimmer based on their physical characteristics using Binary Classification and Artificial Neural Network. Through this project, our questions asked in the introduction can be answered.

* Will physical characteristics play an important role in identifying potential Olympic medalists?
* If so, among height, weight, and age, which of them will be the most crucial?

Physical characteristics does play an important role in identifying potential Olympic medalists as both models were able to identify if an Olympic swimmer will be able to attain a medal in the competition based on their physical characteristics. Based on the Binary Classification Model, the Height of an Olympic swimmer plays a more crucial role in predicting the medal outcome, followed by their Weight then Age. This can be seen by comparing the accuracy rate of the test dataset shown in Figure 10.

The model performance for both the Binary Classification and Artificial Neural Network has high accuracy rates that indicate the success of both models, allowing the model to predict the medal outcome of the Olympic swimmer based on their physical characteristics. Even though the Artificial Neural Network model is less accurate than the Binary Classification model, it can conclude that the Weight plays a more important role in determining the medal outcome of an athlete, followed by the Height then Age. Both models concludes that the Age plays the least factor in determining the medal outcome of the athlete. The Binary Classification model is also able to perform well with new and unseen data, which result in high accuracy, precision, and recall rate.

One of the limitations could be the lack of variables provided in the dataset. The dataset provided contains variables such as Sex, Age, Height, and Weight which are factors out of the athletes’ control. Thus, the model could only be trained using physical characteristics as predictors. However, external factors such as training hours, facilities provided, nutrition and lifestyle all play a vital role in athletes’ performance.

# CHAPTER 7: APPENDICES



Figure 1: Extracting columns that we want from the original dataset

A screenshot of a computer

Description automatically generated

Figure 2: Replacing the different types of medals (Bronze, Silver and Gold) to ‘Y’

A screenshot of a computer

Description automatically generated

Figure 3a: Removing all the rows with empty cells

A screenshot of a computer program

Description automatically generated

Figure 3b: Changing the object type of Medal to ‘Category’

A screenshot of a computer program

Description automatically generated

Figure 4a: Further extracting rows with the desired parameters (Male Swimmers that won Medals)

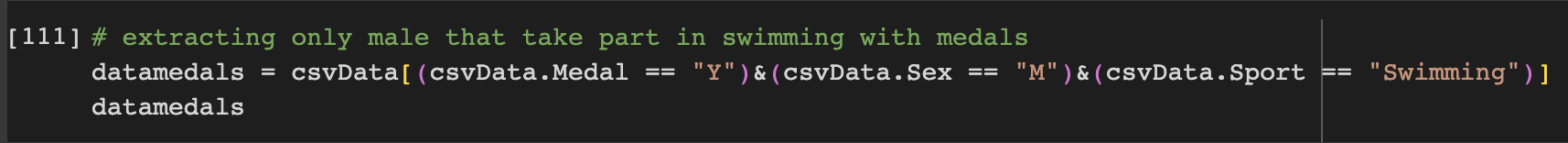


Figure 4b: Further extracting rows with the desired parameters (Male Swimmers from US, Singapore, South Korea and China)

A black screen with text

Description automatically generated

Figure 5: Splitting the data into train and test dataset for each physical characteristic

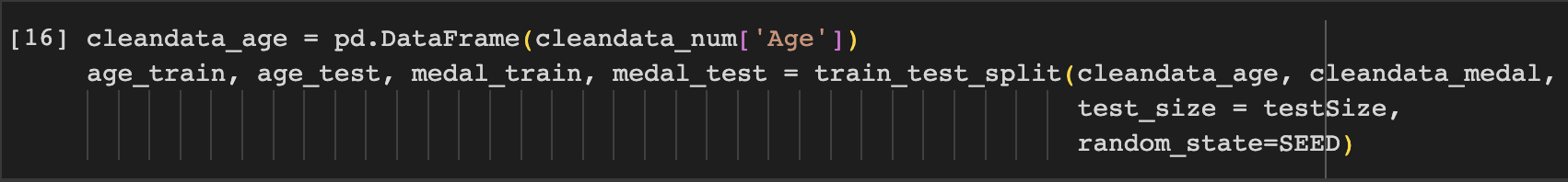


Figure 6: Creating a Decision Tree Classifier to fit the training data

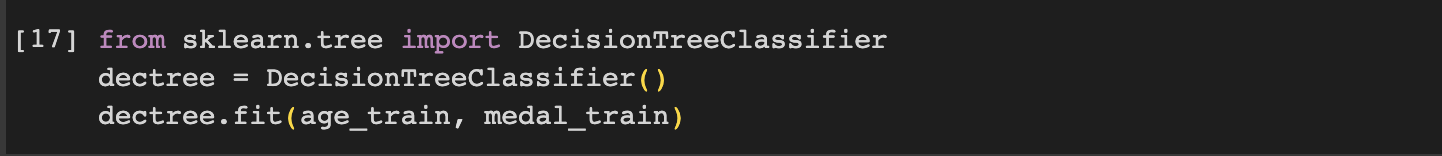


Figure 7: Using the decision tree to predict the medal outcome, plot the confusion matrix and calculate the accuracy, precision and recall rate on the training data

A screen shot of a computer program

Description automatically generated

Figure 8: Using the decision tree to predict the medal outcome, plot the confusion matrix and calculate the accuracy, precision and recall rate on the testing data

A screen shot of a computer code

Description automatically generated

Figure 9: Accuracy, Precision and Recall rate for all the 3 physical characteristics in the training and testing dataset with a multi-variate tree score that used all 3 physical characteristicsA screenshot of a computer

Description automatically generated

Figure 10: : Accuracy, Precision and Recall rate for all the 3 physical characteristics in the training and testing dataset A screenshot of a graph

Description automatically generated

Figure 11: Converting the categorical value of the Medal Class using Label Encoder

A computer screen shot of white text

Description automatically generated

Figure 12: Computing the weights for the class

A black screen with text

Description automatically generated

Figure 13: Parameters for the ANN algorithmsA screen shot of a computer program

Description automatically generated

Figure 14: Initializing ANN using the different layers

A computer screen shot of a program code

Description automatically generated

Figure 15: Compiling and Fitting the model

A screen shot of a computer program

Description automatically generated

Figure 16: Results of the prediction based on the accuracy, precision, recall and AUC score

A screen shot of a computer program

Description automatically generated

GitHub Repo Link: https://github.com/SherminChai/FYP.git

# CHAPTER 8: REFERENCES

1. Qi, Z. (2022). Sports Achievement Prediction and Influencing Factors Analysis Combined with Deep Learning Model. *Scientific Programming*, *2022*, 1–8. Available at: <https://doi.org/10.1155/2022/3547703> (Accessed: 16 September 2023)
2. Zhang, S., & Feng-Jun, S. (2022). Feature extraction of athlete’s Post-Match psychological and emotional changes based on deep learning. *Computational Intelligence and Neuroscience*, *2022*, 1–9. Available at: <https://doi.org/10.1155/2022/2995205> (Accessed: 11 July 2023)
3. Oytun, M. *et al.* (2020) *Performance Prediction and Evaluation in Female Handball Players Using Machine Learning Models*. Available at: https://ieeexplore.ieee.org/abstract/document/9122509/ (Accessed: 11 July 2023).
4. Shafi, A. (2023). *Random Forest Classification with Scikit-Learn*. Available at: <https://www.datacamp.com/tutorial/random-forests-classifier-python> (Accessed: 16 September 2023).
5. Mijwil, M.M. (2018) *Artificial Neural Networks Advantages and Disadvantages*. Available at: https://www.linkedin.com/pulse/artificial-neural-networks-advantages-disadvantages-maad-m-mijwel (Accessed: 10 July 2023).
6. *Advantages and disadvantages of different classification models* (2020) *GeeksforGeeks*. Available at: https://www.geeksforgeeks.org/advantages-and-disadvantages-of-different-classification-models/ (Accessed: 11 July 2023).
7. *What is Overfitting* (2023) *Amazon*. Available at: https://aws.amazon.com/what-is/overfitting/ (Accessed: 11 July 2023).
8. *What is Neural Network* (2023) *Amazon*. Available at: https://aws.amazon.com/what-is/neural-network/ (Accessed: 11 July 2023).
9. Shahid, M. (2019) *Convolutional Neural Network*, *Medium*. Available at: https://towardsdatascience.com/covolutional-neural-network-cb0883dd6529 (Accessed: 11 July 2023).
10. *What is a convolutional neural network?* (2023) *What Is a Convolutional Neural Network?* Available at: https://www.mathworks.com/discovery/convolutional-neural-network-matlab.html (Accessed: 11 July 2023).
11. Arc, A. (2018) *Convolutional Neural Network*, *Medium*. Available at: https://towardsdatascience.com/convolutional-neural-network-17fb77e76c05 (Accessed: 11 July 2023).
12. *Advantages and disadvantages of Convolutional Neural Network (CNN)* (2023) *Advantages and Disadvantages of Convolutional Neural Network*. Available at: https://aspiringyouths.com/advantages-disadvantages/convolutional-neural-network-cnn/ (Accessed: 11 July 2023).
13. *Artificial Neural Networks and its applications* (2023) *GeeksforGeeks*. Available at: https://www.geeksforgeeks.org/artificial-neural-networks-and-its-applications/ (Accessed: 11 July 2023).
14. Ginni (2021) *What are the advantages and disadvantages of Artificial Neural Networks*. Available at: https://www.tutorialspoint.com/what-are-the-advantages-and-disadvantages-of-artificial-neural-networks (Accessed: 11 July 2023).