**Project Report on**

**“Selection of Athletes for Olympics based on Gender Specific Anthropometric data using Python”**

**GROUP MEMBERS**

Chaitanya Mundle

Abhimanyu Abhinav

Srinath Varadharajan

**PROFESSOR**

Dr. Xizhou Feng

102A, Poole Agricultural Center, Clemson University

CPSC 6300, Applied Data Science

Department of Computer Science

**Academic Integrity:**“As members of the Clemson University community, we have inherited Thomas Green Clemson’s vision of this institution as a “high seminary of learning.” Fundamental to this vision is a mutual commitment to truthfulness, honor, and responsibility, without which we cannot earn the trust and respect of others.  Furthermore, we recognize that academic dishonesty detracts from the value of a Clemson degree.  Therefore, we shall not tolerate lying, cheating, or stealing in any form.”

With reference to the above statement, we provide guarantee that this project work is of our own and within the guidelines of the Academic integrity policy of Clemson University and take full responsibility for our work in its entirety.

**Chaitanya Mundle Abhimanyu Abhinav Srinath Varadharajan**

**ABSTRACT:**

Gender Classification has been an integral part of the selection process in the Olympics. It has been in the news for various reasons in recent years. The main reason for this is to ensure fair play in sporting events, considering the differences in physical attributes in males and females, which can affect performance. During the 2009 Berlin Olympics, South African sprinter Caster Semenya was booked by the Olympic committee for winning the 800m under suspicious manner. The arguments put forth by the committee was that she had abnormally high “masculine” traits which gave her an advantage over other female participants. The main purpose of our project is to input the athletes’ anthropometric data (height, weight and age) and use our predictive modelling techniques to predict the gender of an athlete given his/her physical attributes. If the predicted gender of that athlete contradicts the actual gender value then, we can infer that that athlete is an outlier in his/her gender category and there is strong probability of that athlete being a favorite or disadvantaged, which can have a huge impact on selection.

**INITIAL MODELLING AND ANALYSIS EFFORTS:**

We started off with a team meeting to discuss ideas. Abhimanyu Abhinav obtained the athlete\_events.csv and noc\_regions.csv from Kaggle and explained the dataset to provide us an initial understanding of the input data. Given the problem statement, he started with the initial modelling and analysis efforts. Some of the assumptions made are:  
1. The assumption that for a event, females have weaker physical attributes compared to males is derived from historical data of attributes of athletes who participated in similar events in the past.

2. The results of our model may or may not accurately reflect the ability of the athlete in specific categories. The model does not take every event as its basis of measurement.

3. This model only considered the binary sexes (Male and Female) for its analysis, thereby excluding transgenders. It also excluded athletes drug intake which might artificially increase physical features.

4. The results of the model are applicable only for the modern Summer Olympic Games.

The most challenging part in this project was the lack of medal data in the input. With the help of medal winner’s data, we could easily build a framework using our model to predict the athlete’s performance based on their gender value and select athletes that match the winner’s attributes. But the lack of medal data made the model open for future research. It opened way for many wrong assumptions, and he had to change the model accordingly to evade this disadvantage. After many brainstorming and trial and error sessions, he decided on using two modelling techniques to train our data to predict the accurate result, the Decision Tree classifier method and artificial neural networks. Depending on the accuracy of both the methods, the most suitable one will be selected.

**DATA PRE-PROCESSING**

Srinath Varadharajan was responsible for data pre-processing and visualization. He was responsible for selecting the appropriate libraries required for both the models and after validation, importing the libraries and the dataset into the notebook. Some of the libraries used in our modelling are pandas, seaborn, matplotlib, numpy, sklearn and torch is used for neural networks.

He did basic statistical analysis of the data and got a basic idea on how to proceed with making the data more presentable. The tuples that are duplicated inside both the datasets are noted and removed for data cleaning and shorter computation time. Further, this procedure is repeated after both the datasets are merged and the data is visualized in order to make sure that there are no outliers.

Figure : Data Cleaning

A screenshot of a cell phone

Description automatically generated

Figure : Distribution of data after data cleaning

A screenshot of a social media post

Description automatically generated

Thus, he obtained a proper distribution of our anthropometric data without any outliers. This concludes the step of cleaning our data.

**DATA VISUALIZATION**

With our cleaned data, he needed to explore more into the various insights that the data might provide. The easy and quick way to understand the patterns is to graphical represent and visualize our data. So, for the purpose of gaining more insights, he proceeded with visualizing our gender data using the matplotlib library.

He conducted a gender specific analysis on swimming category of the athlete\_events dataset. Upon combining, he visualized the distribution of male and female swimming participants from the advent of the modern Olympics in 1896 until the recent one in 2016. The results shown were indeed satisfactory and both graphs showed a drastic increase in the participation of males and females on subsequent games.

Using the combine attribute,

Figure : Analysis for Male Population

A screenshot of a cell phone

Description automatically generated

Figure : Analysis for Female Population

A screenshot of a cell phone

Description automatically generated

Figure : Python Code for Visualizing Male Data

A screenshot of a cell phone

Description automatically generated

Figure : Male Data Visualization Graph

A screenshot of a cell phone

Description automatically generated

Figure : Python Code for Visualizing Female Data

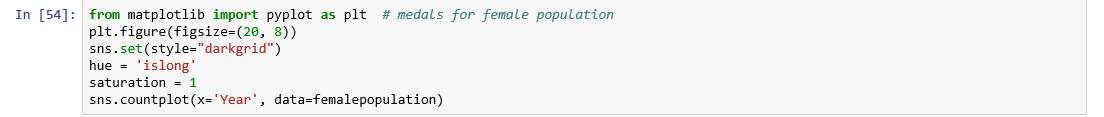


Figure : Female Data Visualization Graph

A screenshot of a cell phone

Description automatically generated

When he visualized our data, he made a conviction that this dataset can be better tested initially through “Decision Tree” model as the only attributes used were height, weight, season for make and female nodes. Here our assumption was that if an athlete is a male then we label his gender as true i.e. 1 otherwise 0. Also, for season, summer is labeled as 1 and winter as 0.

Based on his conviction, we started with the Decision Tree model (method 1).

**METHOD 1:**

**DECISION TREE CLASSIFIER:**

Chaitanya Mundle and Abhimanyu Abhinav were responsible for the data processing of Decision tree classifier. The method involves breaking each attribute of the dataset into smaller subsets and select the best one. In this case, they broke down Season into Summer & Winter and Sex into Male & Female. Abhimanyu eliminated all missing values and built this classifier in order to fit it into the training data, thus giving the dataset on building a framework on its own. With this framework, he would be able to predict the outcome of the test data. After the classification, he create the variables for our train and test data using the sklearn library for machine learning.

Figure : Data Classification and Variable Declaration in Decision Tree

A screenshot of a cell phone

Description automatically generated

After classification and variable declaration, Chaitanya trained the data by fitting the classifier to the train variables and predict the test variables using the framework of the trained data.

Figure : Fitting and Testing of Train and Test Data

A close up of a logo

Description automatically generated

He calculated the percentage of accuracy of our model. Using numpy, he created an array for our test set by calculating the average of the results. The model showed an accuracy percentage of 80.09%.

Figure : Accuracy Percentage Calculation for Decision Tree Classifier

A screenshot of a cell phone

Description automatically generated

To sum up, with the decision tree method, they found out that there is a good 80% chance that the model will accurately predict whether the athlete is male or female, given his/her anthropometric data.

**METHOD 2:**

**ARTIFICIAL NEURAL NETWORKS:**

Artificial neural networks is a model that processes information which resembles the function of the neurons in the brain. Chaitanya Mundle was responsible for framing and processing this method. In our project, he used the PyTorch library to run this model. For the purpose of simplicity, he used only 30% of the entire dataset i.e. 60k tuples for the input.

The model counts the number of self transforms or iterations for each observation in the dataset and returns a value until the last iteration after which the model goes on to count the next item. With the number of tuples and iterators, he then declared the variables for the train and test data by fitting this model with the variables.

He then defined the whole model using epoch , which is the total number of times the data is provided to neural network in forward and backward fashion which can prove helpful instead of iteration technique as iteration process is slow because it decomposes the dataset into batches.

Figure : Variable Declaration and Training Data in Neural Networks

A screenshot of a cell phone

Description automatically generated

He trained the model using “negative log likelihood” using 8 epochs and LR=0.001.

Figure : Epoch Iteration

A screenshot of a cell phone

Description automatically generated

Now for the particular epoch count and the tested data, he calculated the accuracy percentage in a similar way Abhimanyu did for the decision tree classifier. The result he obtained was an accuracy percent of 78.45% with an average loss of 0.4605.

Figure : Accuracy Percentage Calculation for Neural Network Model

A screenshot of a cell phone

Description automatically generated

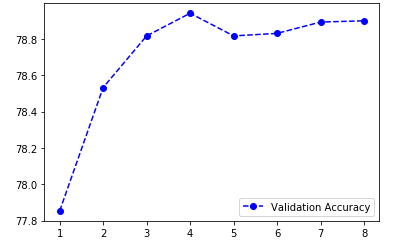
He wanted to investigate more into the insights and inferences for the accuracy percentage and loss, so he visualized the results in a graphical manner for further analysis.

Figure : Code for Visualizing Validation Loss and Accuracy

A screenshot of a cell phone

Description automatically generated

Figure : Validation Accuracy and Validation Loss with epoch count

A picture containing screenshot

Description automatically generated

**INFERENCES**

Some of the inferences we made upon collecting our results for both the models are as follows:

1. The decision tree classifier is deemed to be more effective than the neural network model since it has a higher accuracy percentage compared to neural networks. This might be due to the simplicity and easy natural flow of the decision tree classifier, given the low number of attributes to train.
2. The sharp decrease in the validation loss graph is justified by the sharp decline in the validation loss graph. This mainly depends on the optimal selection of the number of epochs selected for trial and the iterations run.

**SCOPE FOR FUTURE RESEARCH**

Our project has been designed to provide scope for future research. With our anthropometry data, we can include other intrinsic factors such as agility, speed, stamina which are not gender specific. The inclusion of ethnicity and nationality can also add more insights to the output data. This might require a more complex classification model such as neural networks, which might give better accuracy compared to the decision tree model. We can also try on reducing the number of assumptions as low as possible, which might produce a very effective model. This requires a more detailed approach on machine learning, which we are very much excited to explore. We can also use the model to identify athletes of matching gender values but being disadvantaged due to their attributes.

**CONCLUSION**

The study material of our data science course was very useful for understanding and implementing the Decision tree classifier and the various research papers on Gender & Oympics proved to be vital for our analysis. The importance of our model is evident throughout not only in the Olympics, but also for the field of sports in general. Our model might be a huge prescreen factor for the national Olympic selection team, which might select the best athletes based on their anthropometric data. In the case of Semenya, with this model, she would have been an outlier in the favorites subset for her category and would have been advised against to participate for the event to prevent disqualification. Our model might be the starting step to measures taken for gender equality and proper team selection in sports, which is the need of the hour.

**REFERENCES**

Donnelly, P., & Donnelly, M. K. (2013). *The London 2012 Olympics: A gender equality audit*. Toronto: Centre for Sport Policy Studies.

Capranica, L., Piacentini, M. F., Halson, S., Myburgh, K. H., Ogasawara, E., & Millard-Stafford, M. (2013). The gender gap in sport performance: equity influences equality. *International Journal of Sports Physiology and Performance*, *8*(1), 99-103.

Norton, K., Olds, T., Olive, S., & Craig, N. (1996). Anthropometry and sports performance. *Anthropometrica*, 287-364.

https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results https://medium.com/@jamesdacombe/an-introduction-to-artificial-neural networks-with-example ad459bb6941b

https://clemson.instructure.com/courses/89573/files

https://www.olympic.org/women-in-sport/background/statistics

https://torokhtiy.com/blogs/warm-body-cold-mind/weight-category-and-height

https://www.nytimes.com/2012/06/18/sports/olympics/the-line-between-male-and-female-athletes-how-to-decide.html

http://www.312analytics.com/decision-trees-vs-neural-networks/