**Sri Lanka Institute of Information Technology**



**Fundamentals of Data Mining (IT3051)**

**Mini Project – Final Report**

|  |  |
| --- | --- |
| **Registration Number** | **Name** |
| IT20300746 | Samarakoon S.M.N.D. |
| IT20060008 | Wickramaratne D.D.W |
| IT19067216 | Siribaddana S.G. |
| IT20046798 | Alwis L.W.S.J. |
| IT20198718 | Halis M.K.M. |

**Table of Contents**

[Introduction 3](#_Toc36650)

[Problem Statements and Solutions 4](#_Toc36651)

[Technologies 4](#_Toc36652)

[Dataset Introduction 5](#_Toc36653)

[Regression 11](#_Toc36654)

[1. Data Preprocessing and Transformation 11](#_Toc36655)

[2. Implementation of Model 17](#_Toc36656)

[3. Prediction and Visualization 19](#_Toc36657)

[Deployment and Integration 27](#_Toc36662)

[Challenges 27](#_Toc36663)

[Testing 28](#_Toc36664)

[Deliverables 34](#_Toc36665)

[Work Distribution 41](#_Toc36668)

[References 42](#_Toc36669)

# 

# Introduction

Horse racing is an equestrian performance sport, typically involving two or more horses ridden by Jockeys over a set distance for competition. It is one of the most ancient of all sports, as its basic premise is to identify which of two or more horses is the fastest over a set course or distance. It has been mostly unchanged since at least classical antiquity.

Horse races vary widely in format, and many countries have developed their own particular traditions around the sport. Variations include restricting races to particular breeds, running over different distances, running on different track surfaces, and running in different gaits. In some races, horses are assigned different weights to carry to reflect differences in ability, a process known as handicapping.

While horses are sometimes raced purely for sport, a major part of horse racing's interest and economic importance is in the gambling associated with it, an activity that in 2019 generated a worldwide market worth around US$115 billion.

This web app is going to help us with predicting the speed of a certain horse depending on the physical and physiological factors that affect its performance in the racetrack.

# 

# Problem Statements and Solutions

* **Problems**

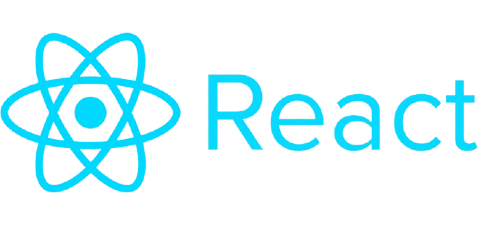
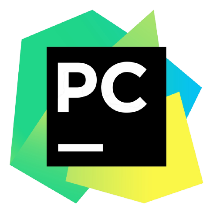
Betting on horses is called Gambling and it’s an act that’s widely considered to be a matter of chance. It can be exceptionally hard to predict the winning horse for a novice.

* **Solutions**

Our solution was to use a multiple linear regression model to predict the speed of a horse, once several variables are provided.

# Technologies

* **Front end** – Streamlit Library , Python
* **Back end** - Python
* **IDE** - VS code



# 

# Dataset Introduction

The data was obtained from various free sources and is presented in CSV format. Personally-identifiable information, such as horse and jockey names, has not been included. However these should have no relevance to the purpose of this dataset, which is purely for experimental use.

There are two files: Runs.csv , Races.csv

For our purpose we’ve decided to combine those 2 datasets and drop several columns as we deemed them to be unusable.

Feature engineering was done by Excel.

Link -  <https://www.kaggle.com/datasets/gdaley/hkracing?select=races.csv>

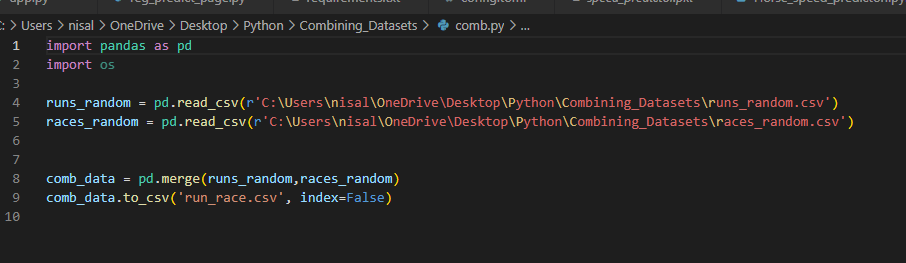
**Dataset row information**

|  |  |
| --- | --- |
| **Field** | **Description** |
| horse\_type | **String**  The type of the horse ex : Gelding, Mare , …. |
| config | **String**  Race track configuration |
| horse\_age | **String**  The age of the Horse |
| horse\_weight | **Integer**  Weight of the Horse |
| surface | **Integer**  Dirt or Turf |
| average\_speed | **float**  The average speed of the horse during the race |

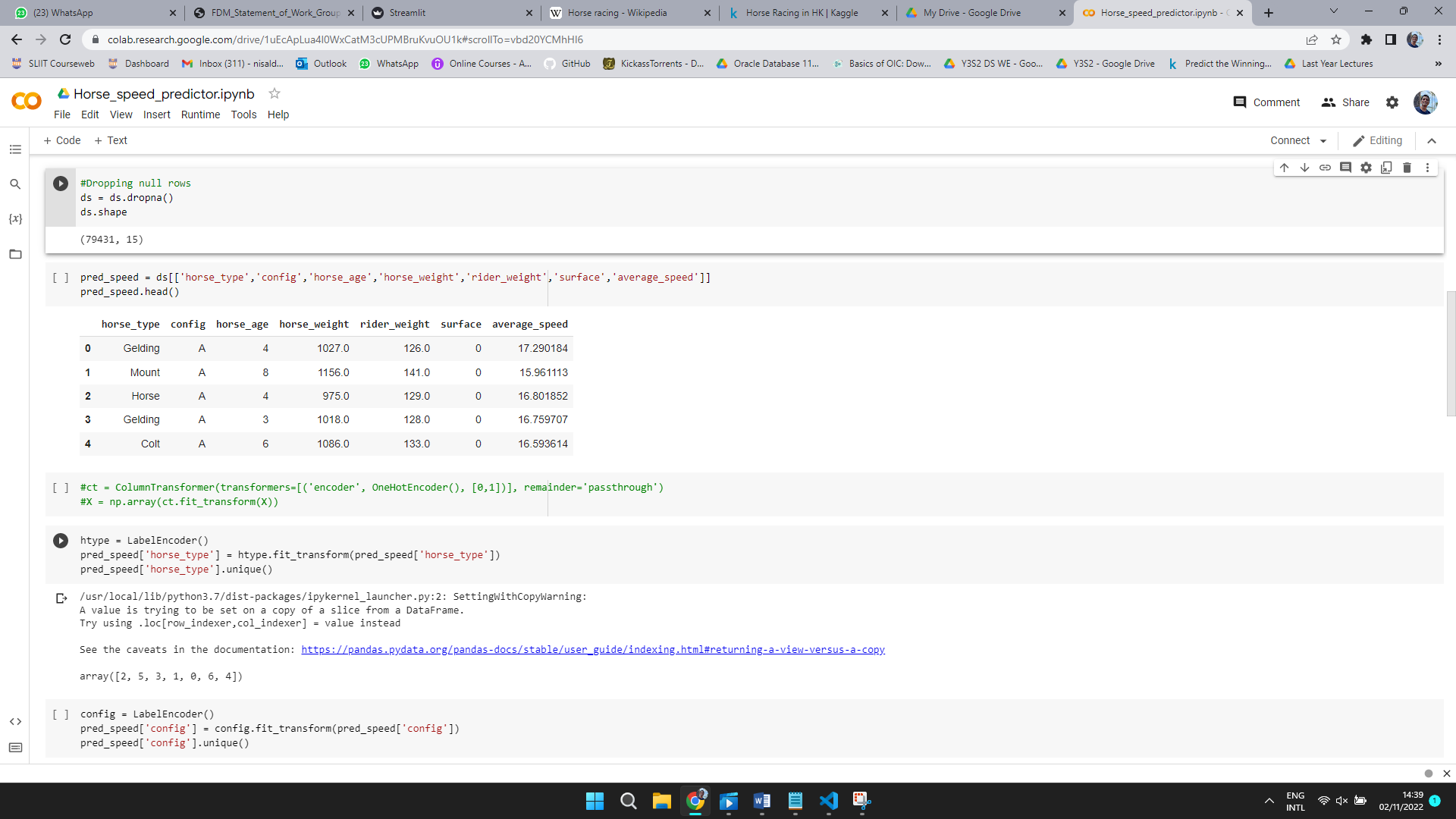
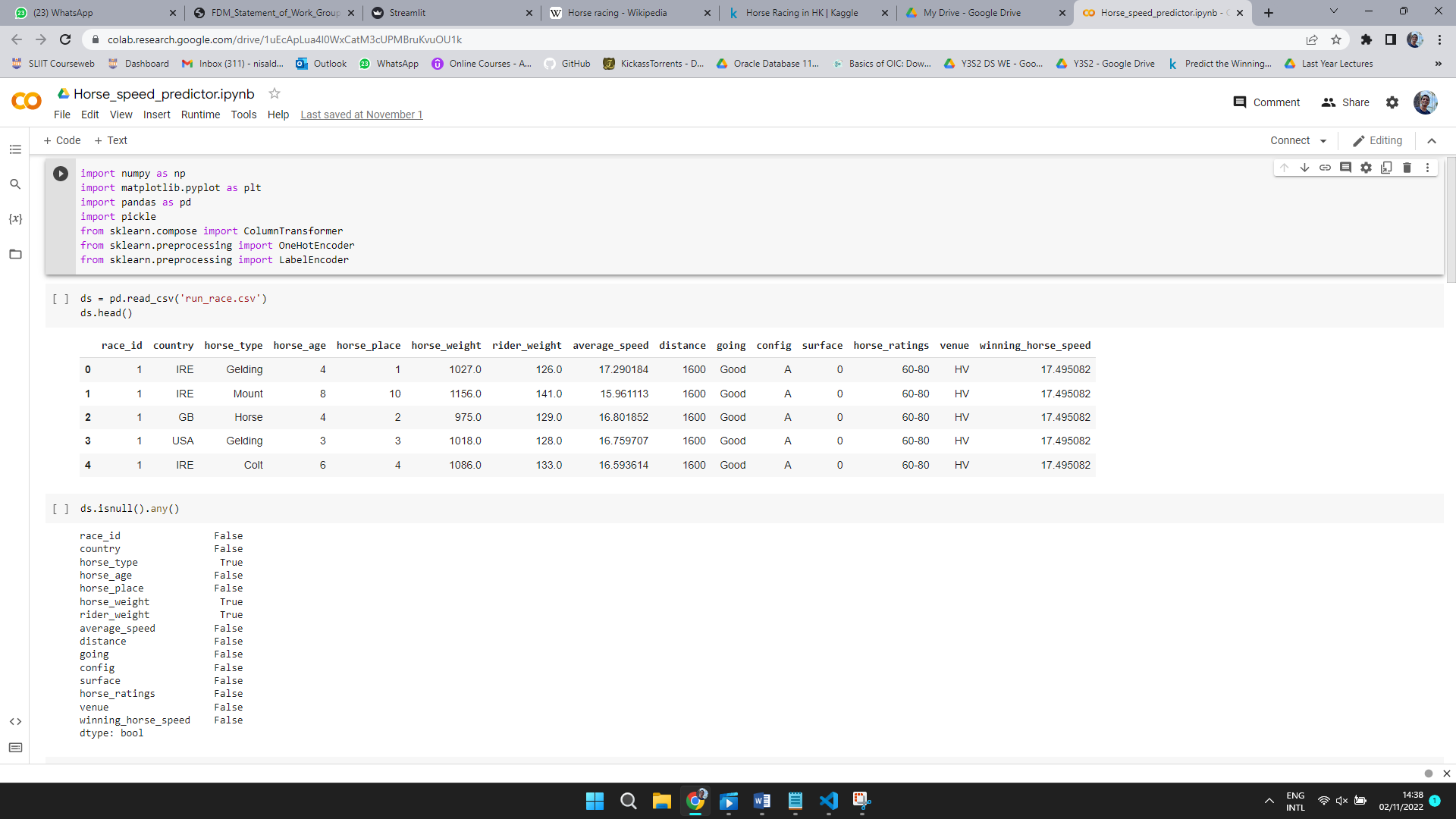
# Regression

## 1. Data Pre-processing and Transformation

* Combining CSV files

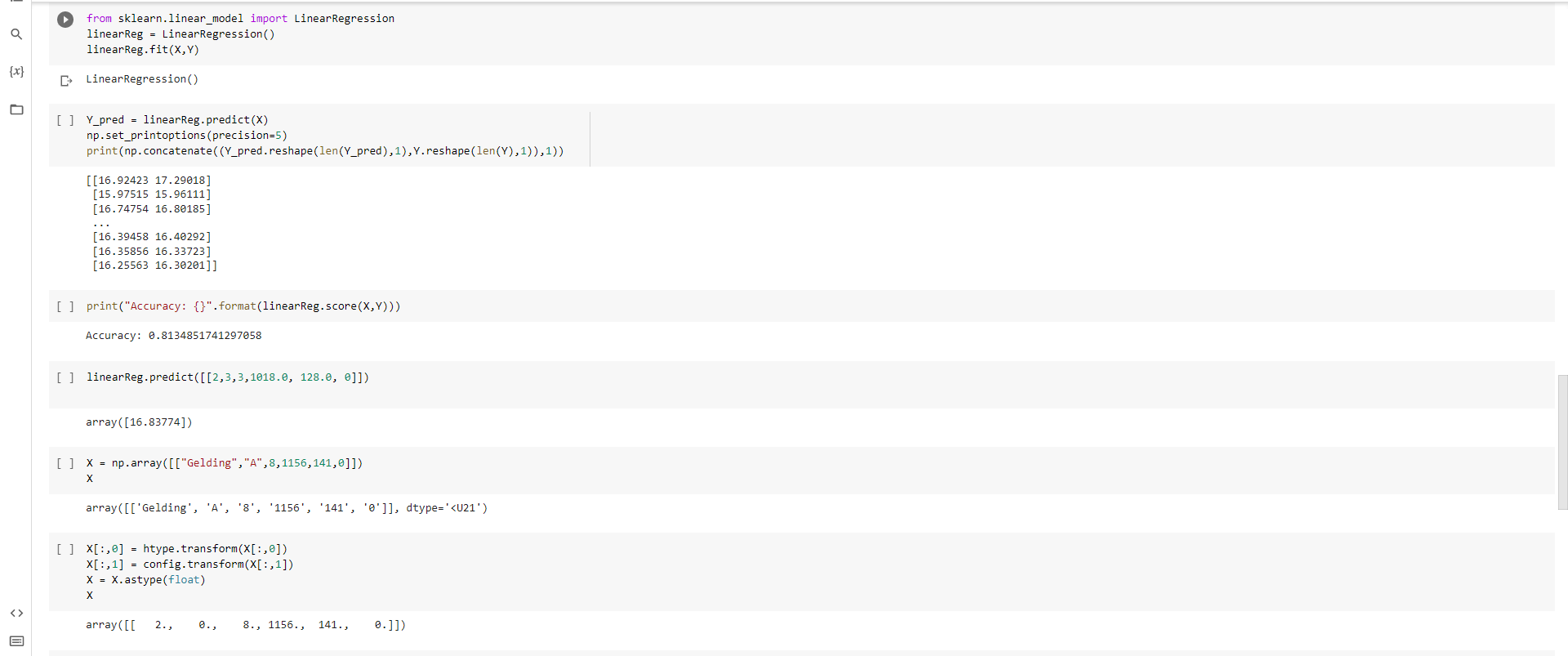


* Pre-processing before implementing the model



## 2. Implementation of Model

* Multiple Linear Regression

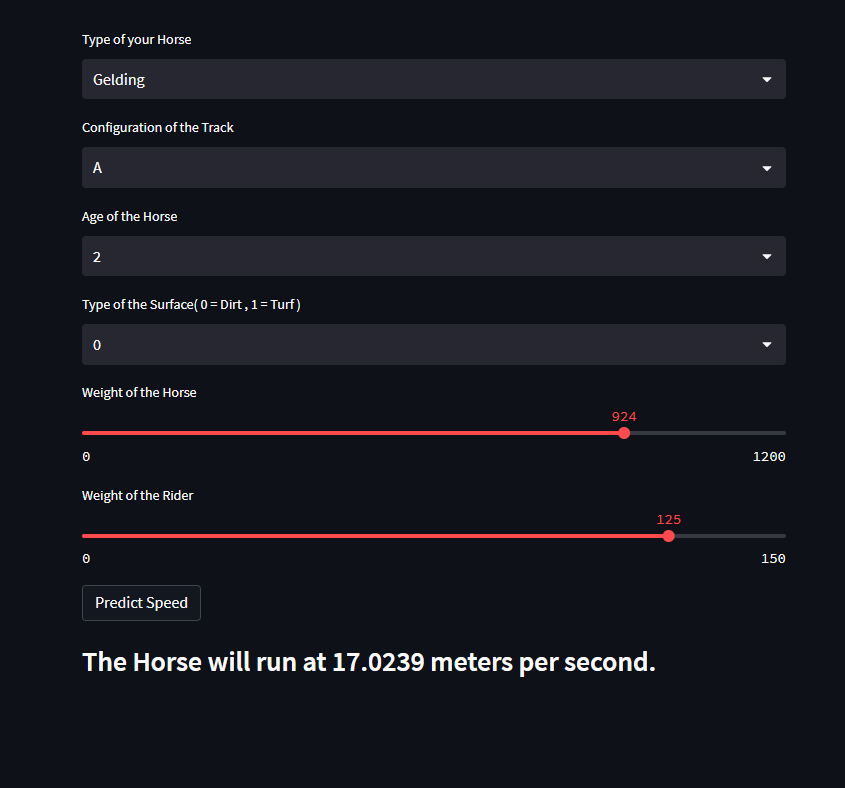


**Note**

* This particular model was chosen after testing the accuracy of several models which were made using different variables.

## 3. Prediction

Once the Necessary variables are entered the app will print the potential average speed of the Horse.



# Deployment and Integration

* Streamlit was used to Intregate the Frontend and Backend
* The project was pushed to GitHub and it was connected to Streamlit.io

GitHub link : <https://github.com/NisalDaksith/Horse_Race_Speed_Predictor_Webapp>

* Deployed project: <https://nisaldaksith-horse-race-speed-predictor-webapp-app-uru4vx.streamlitapp.com/>

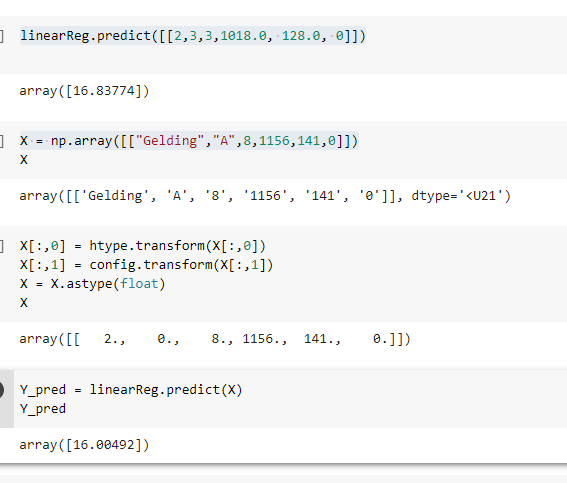
# Challenges

* Since the original 2 datasets were not combined it was hard to get suitable X and Y values for the model. We managed to overcome that by Feature Engineering with MS Excel.
* We originally wanted to use Flask but due to time constraints we switched to Streamlit.

# 

**Testing**

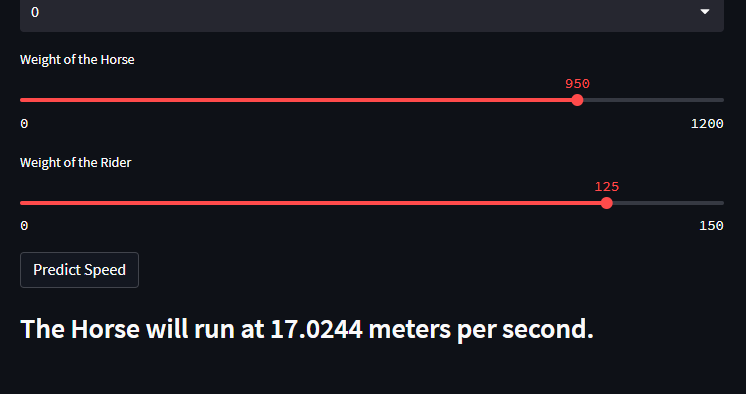
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Test Id** | 01 | | |  |
| **Test case** | Testing the selected model working properl | | | y |
| **Prerequisite** | Dataset files were preprocessed | | |  |
| **Action** | **Inputs** | **Expected output** | **Actual Output** | **Test result** |
| Entering inputs using a code | X = np.array([["Gelding",  "A",8,1156,  141,0]]) | Array  ([16.00492]) | Array  ([16.00492]) | Pass |



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Test Id** | 02 | | |  |
| **Test case** | Testing the data input of the web app | | |  |
| **Prerequisite** | Web app needs to be launched | | |  |
| **Action** | **Inputs** | **Expected output** | **Actual Output** | **Test result** |
| Testing the web app | Inputs were entered successfully | None | None  (No errors) | Pass |



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Test Id** | 03 | | |  |
| **Test case** | Testing the outcome | | |  |
| **Prerequisite** | Inputs need to be given | | |  |
| **Action** | **Inputs** | **Expected output** | **Actual Output** | **Test result** |
| Testing the web app predict button | Shown in the picture | prediction | prediction | Pass |



# Deliverables

The user can predict the speed of the race as long as the necessary variables are provided. However the user must acquire the derails of the Horse and the Racetrack prior to using this app.

# Work Distribution

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Registration Number** | **Role** | **Responsibility** |
| S.M.N.D. Samarakoon | IT20300746 | Implementing the Machine Learning related Tasks | Data Pre-processing, Model Designing and model Building |
| Wickramaratne D.D.W. | IT20060008 | UI/UX development and ML research | Extensive research on UI/UX and ML models |
| Siribaddana S.G. | IT19067216 | ML related Research and Model Building | Identify the model outcome and planning the data visualization |
| Alwis L.W.S.J | IT20046798 | Data mining, Research and Documentation | Data Mining, Documentation and evaluating the Progress |
| Halis M.K.M. | IT20198718 | UI/UX development and ML research | Extensive research on UI/UX and ML models |

# References

* Dataset Description : <https://www.kaggle.com/datasets/gdaley/hkracing?select=races.csv>
* Web App : [https://nisaldaksith-horse-race-speed-predictor-webapp-app-uru4vx.streamlitapp.com/](https://nisaldaksith-horse-race-speed-predictor-webapp-app-uru4vx.streamlitapp.com/%20)
* GitHub : <https://github.com/NisalDaksith/Horse_Race_Speed_Predictor_Webapp>
* Video :