**Project Cover Sheet**

Student IDs : 53545289, 40101063

Student Names : WANG HAOYING, Malo Ferriol

Student Emails : [haoyiwang3-c@my.cityu.edu.hk](mailto:haoyiwang3-c@my.cityu.edu.hk), [mferriol2-c@my.cityu.edu.hk](mailto:mferriol2-c@my.cityu.edu.hk)

**Title:**

Prediction of football match results based on Python parallel machine learning and classification of clubs based on Spark machine learning

**Highlights:**

* Choose Premier League dataset so the analysis is meaningful for real world problem
* Use 3 different models (LR, RNC, XGBC) for Python paralleling
* Design and implement K-means clustering solution in Scala
* Comparison of stand-alone (with and without HDFS) and AWS EMR clustering

**List of Deliverables**

|  |  |
| --- | --- |
| **File Name** | **Description** |
| Final report.pdf | Main Project Report |
| BigdataPrez.pptx | Presentation slides |
| <https://drive.google.com/open?id=1VyqjBYoOKI8Jo_2TPZinx3nq-MGRRXSK> | Presentation video link |
| Code.py | Code of Python paralleling |
| Scala files(3) | Code of Scala |
| Figures.zip | Figures used in the project report |

------ END ------