**Predicting MLS Results**

Alexandre Moreira Batista

Armando Defendi Rossi

Stephen Chiong

Douglas College

**Predicting MLS results**



**Executive Summary:**

This project aims to provide sports fans with accurate predictions for Major League Soccer (MLS) games using robust machine learning models. By utilizing web scraping, database management, machine learning analysis, and an Android application, this project will deliver valuable predictions to users without the need for them to pay exorbitant fees to sports betting companies. This project will cover all aspects of the Data Analytics program at Douglas College. The primary beneficiaries will be soccer fans with access to accurate game predictions at no cost.

**Table of Contents**

[1. Proposal 4](#_Toc126570130)

[A. Statement of the Problem 4](#_Toc126570131)

[B. Significance of the Study 4](#_Toc126570132)

[C. Technologies 4](#_Toc126570133)

[D. Project Implementation and timelines 5](#_Toc126570134)

[E. Potential challenges for this project: 5](#_Toc126570135)

[2. Report 01 6](#_Toc126570136)

[A. Project Implementation 6](#_Toc126570137)

[i. Complete Steps 6](#_Toc126570138)

[a) Development and testing of database architecture to receive data. 6](#_Toc126570139)

[b) GitHub configuration for Project 7](#_Toc126570140)

[c) Development of Python scripts for web scraping 8](#_Toc126570141)

[d) Stage 01 - Development of the Android App 9](#_Toc126570142)

[e) Web scrapping for Machine Learning 11](#_Toc126570143)

[ii. Pending implementations 13](#_Toc126570144)

[B. Proposed Reviews 13](#_Toc126570145)

[3. References 15](#_Toc126570146)

# Proposal

## Statement of the Problem

The sports betting industry has seen significant growth recently, with large companies and start-ups competing for customers and revenue. While many companies offer predictions for game results, access to this valuable information often comes at a cost. This project aims to provide a cost-effective solution for all sports fans to access accurate predictions for MLS games.

## Significance of the Study

This project aims to deliver predictions that improve the odds ratio when determining the winner of a game. By utilizing techniques such as web scraping, database management, machine learning, and analytics, this project will also identify features that impact the accuracy of predictions. Additionally, this project will deliver an Android application, providing users with easy access to past and future game predictions.

## Technologies

The project will make use of the following technologies:

* Operational System: Windows in a Virtual Machine
* Programming Language: Python for web scraping and data prediction, Android Studio for mobile application development
* Database: Microsoft SQL Server
* Front-end and Backend: Android Studio

## Project Implementation and timelines

* Development and testing of database architecture to receive data - 1 week
* Development and testing of Python scripts for web scraping - 2 weeks
* Research on modelling and machine learning uses in soccer results prediction - 2 weeks
* Preparation and setup of a virtual machine in the cloud (Azure) to run the database and modelling - 1 week
* Data wrangling in Python - 1 week
* Data modelling (machine learning to predict and check prediction accuracy) - 2 weeks
* Development of the Android App - 4 weeks
* Testing the entire ecosystem and fixing any bugs - 1 week.

## Potential challenges for this project:

* Data acquisition: One of the main challenges of this project will be acquiring enough data for the machine learning models to be trained on. The data needs to be of high quality and accuracy to ensure the predictions are as accurate as possible. This process will require web scraping techniques and robust data wrangling techniques to obtain, validate and process the data.
* Model Selection and Training: It will be challenging to select the appropriate machine learning model(s) to use and train it/them to make accurate predictions. It will require careful consideration of the type of data, the goal of the predictions, the size of the dataset and the computational resources available in the Virtual Machine.
* Data privacy and legal issues: Web scraping for data can raise privacy and legal concerns, such as potential data breaches, data misuse or copyright infringement. The project must comply with all applicable laws and regulations regarding data collection and use.
* Handling missing data and unexpected results: Not all data may be available for every game, and some results may be unexpected, even with accurate predictions. The project will need a strategy to handle missing data and unexpected results to minimize the impact on the predictions.
* Testing and validation: The project will require robust testing and validation of the web scraping, data wrangling, modelling, and android application stages to ensure the predictions are accurate and the application is user-friendly and bug-free. This can be a time-consuming task that requires many resources and testing hours.

# Report 01

## Project Implementation

### Complete Steps

#### Development and testing of database architecture to receive data.

After some testing on the Azure SQL (figure 01), the team decided to move forward using MySQL, and the database structure is represented in figure 02. The MYSQL solution is hosted online and can be accessed using Python scripts (web scraping) and Android APP.

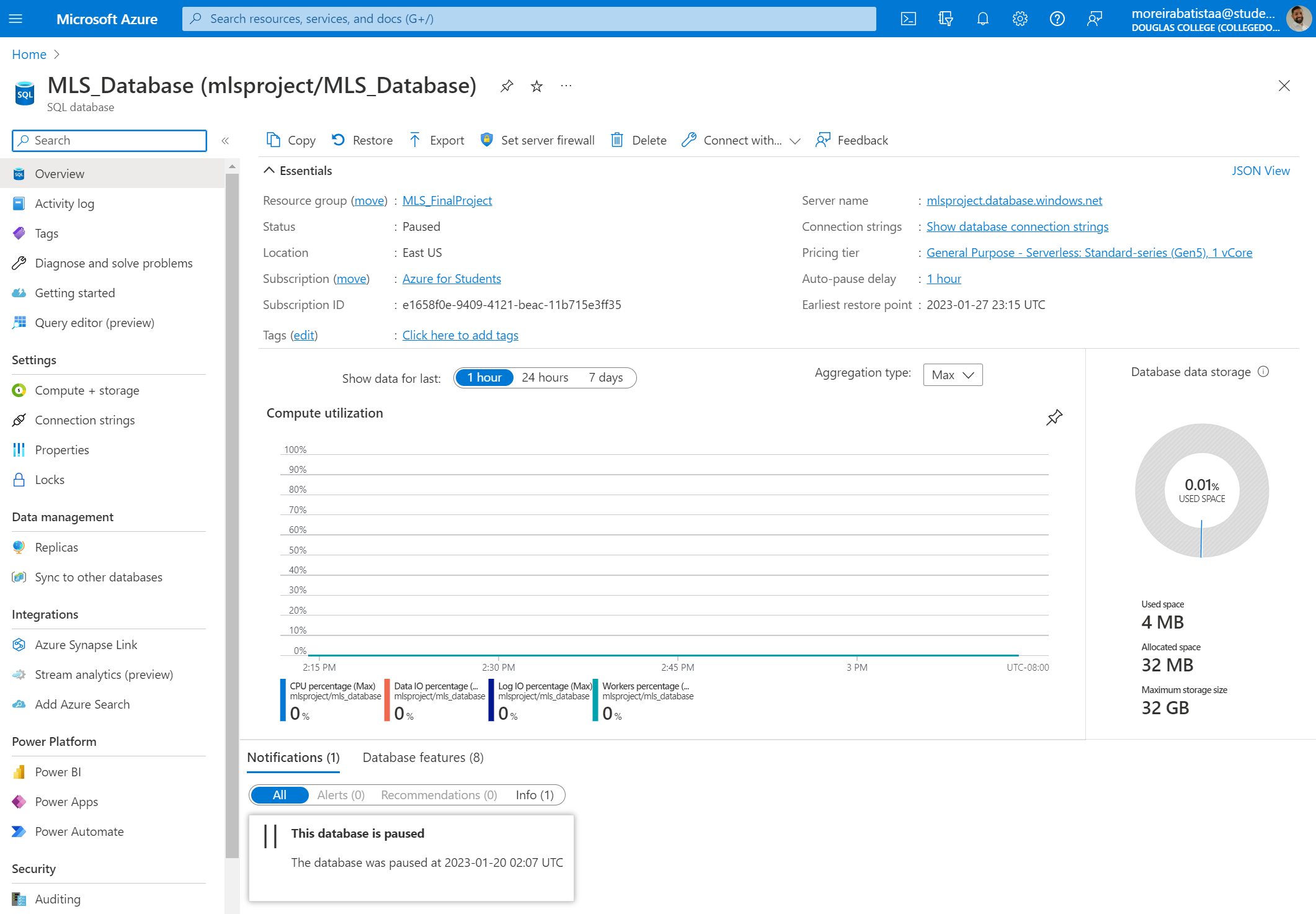


Figure 1 - Azure SQL Server

**Graphical user interface, text, application

Description automatically generated**

Figure 2 - MySQL structure

The first tables were created, and the connection tested.

Database configuration: Alexandre on 16-20/Jan/2023  
Database creation of tables/fields: Armando on 16-20/Jan/2023

#### GitHub configuration for Project

A repository was created to manage the versioning and save all the scripts and codes online as shown on figure 03.

GitHub creation: Armando on 16-20/Jan/2023

Graphical user interface, text, website

Description automatically generated

Figure 3 - GitHub directory

#### Development of Python scripts for web scraping

The scripts were developed using some libraries from Python based on the page: [*https://www.mlssoccer.com/schedule/scores#competition=mls-regular-season*](https://www.mlssoccer.com/schedule/scores#competition=mls-regular-season). It is a dynamic page, and the usual libraries, such as Request and BeautifulSoap, do not work with this kind of page.

Graphical user interface, text, application, email

Description automatically generated

Figure 4 - Python script

The scripts bring data from 2015 to actual year and insert it into the schedules table. We got 3,573 entries for that period.

Graphical user interface, application

Description automatically generated

Figure 5 - Schedule table details

Python scripts: Armando on 23-26/Jan/2023

#### Stage 01 - Development of the Android App

The first visuals were developed, including the Login, Register and User administration page layout.

Graphical user interface, text, application

Description automatically generated Graphical user interface, text, application, email

Description automatically generated  
*Figure 6 - Login page* *Figure 7 - Registration page*

Graphical user interface, application

Description automatically generated Graphical user interface, text, application

Description automatically generated  
*Figure 8 - User Active page* *Figure 9 - User Inactive page*

Graphical user interface, text, application

Description automatically generated  
*Figure 10 – GitHub repository*

App Development: Armando on 23/Jan to 05/Feb/2023

#### Web scrapping for Machine Learning

In order to improve the accuracy of the machine learning model's forecasts, it was determined that additional data was needed beyond the results of each game. To acquire this information, the development of Python scripts to scrape data from the FB Ref website was initiated.*Graphical user interface, application

Description automatically generated*  
*Figure 11 – Python web scrapping for ML*

Graphical user interface, text, application, table

Description automatically generated  
*Figure 12 – Python web scrapping for ML*

Web scraping Development: Alexandre on 23/Jan to 05/Feb/2023

### Pending implementations

As of this stage of the project, progress is being made in accordance with the proposed schedule.

## Proposed Reviews

The original database solution was adjusted to accommodate the infrastructure requirements of Azure, including the implementation of a firewall with a restricted list of IP addresses. However, it was later determined that this solution was not ideal and a change to a MySQL database was necessary. The choice to switch to MySQL was based on its accessibility and the ability to easily connect to it from various sources, including the Python scripts, Android app, and Workbench, using only the credentials provided. Additionally, this database solution was found to be more cost-effective as it is free to use.

# References

Major Soccer League. Schedule & Scores: <https://www.mlssoccer.com/schedule>

FB Ref – Soccer References Databases: <https://fbref.com/en/comps/22/history/Major-League-Soccer-Seasons>