**NAME: Druthisha. B**

**Computer science Artificial Intelligence**

**Reg no: 24UG00309**

**WEEK 1 – DATA COLLECTION & PREPARATION**

**OBJECTIVES:**

• Gather and clean FIFA World Cup data fit for machine learning.

• Create a bespoke web scraper for specialized sports data.

• Clean and engineer features for improved prediction accuracy.

**DATA SOURCES:**

• FIFA official archives for match outcomes.

• Kaggle datasets for past player and team statistics.

• Transfer markt for player experience and transfer.

• Live FIFA rankings and players' caps custom scraper.

**WEB SCRAPER INFORMATION:**

A Requests + BeautifulSoup targeted HTML tables custom scraper was applied to team stats, goals, and match results.

Dynamic pages were managed through Selenium.

Data were parsed into organized CSV files.

**DATA CLEANING PROCESS:**

• Duplicates removed and missing values managed (mean/mode imputation).

• Dates and team names standardized.

• Transformed categorical features for modeling compatibility.

**FEATURE ENGINEERING:**

Feature | Description

team\_avg\_age | Team player average age

fifa\_ranking | Current FIFA ranking

goal\_diff | Goals scored - goals allowed

win\_rate | Win rate of last 10 matches

experience\_score | Squad player average caps

**TOOLS USED:**

Python, Pandas, NumPy, BeautifulSoup4, Selenium, Scikit-learn

**RESULTS:**

A cleaned dataset 'cleaned\_fifa\_worldcup.csv' with 20+ engineered features developed.

FIGURE 1 – Web Scraping Workflow (Placeholder)

TABLE 1 – Feature Description (Placeholder)

**CHALLENGES:**

• Lacking historical information.

• Content on websites rendered with JavaScript.

• Normalizing datasets from multiple sources.

**CONCLUSION:**

Week 1 set a solid and clean data foundation for building the model