KSAT-Quest Regression Report

Team: DataMiners

1. Objective

To predict runoff values using regression models based on a diverse dataset compiled from multiple sources.

---

2. Data Preparation

- **Source**: An Excel workbook with a main sheet (“All data”) and multiple reference sheets (`Ref #1` to `Ref #44`).
- **Approach**:
    - Each sheet was read into a DataFrame.
    - Unique columns across all sheets were compiled.
    - A **merged DataFrame** was created with all columns, filling missing values where necessary.
- **Data Cleaning**:
    - Unnamed columns and rows with all `NaN` values were removed.
    - Missing values were analyzed by percentage and quantity.

---

3. Feature Engineering

- Unused or mostly-null columns were dropped.
- Useless or Identifying features were removed
- Remaining features were preprocessed:
    - Categorical encoding
    - Numerical scaling

---

4. Modeling

- Regression models tested:
    - **Random Forest Regressor**

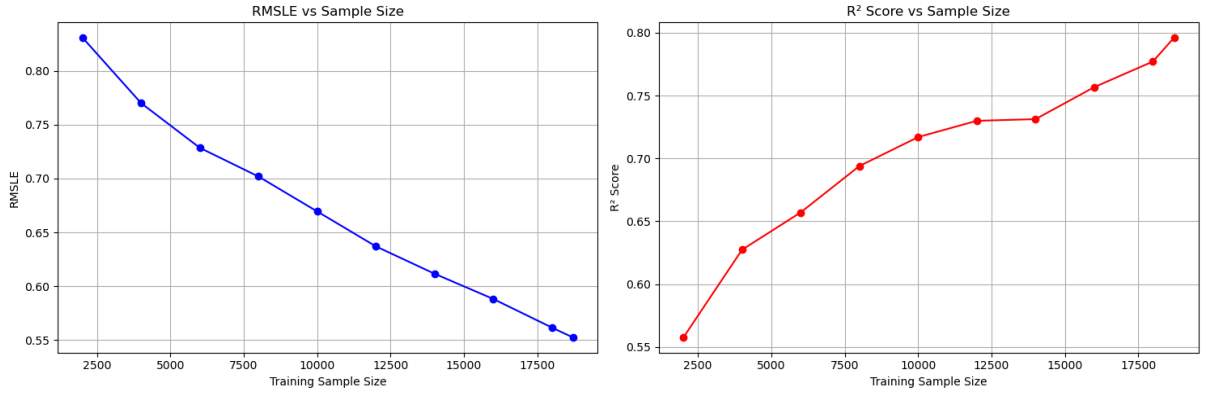
- Model performance was evaluated using metrics like:

    - Root Mean Squared Logarithmic Error (RMSLE)

    - R^2 Score

---

5. Results



We can see that our R2 is going up as we increase the sample size and it hasn't quite plateaued yet, so if we had more data we could probably make the model even better.
