**Documentation and Reporting Template for NCAA Women's Basketball Analysis**

**1. Introduction**

* Clearly state your research question:

"Can we predict a player's NCAA division based on their demographic and physical attributes (e.g., height, year/class, state or country of origin, position)?"

**2. Data Preparation**

* Sources: NCAA Women's Basketball Rosters dataset provided by the University of Maryland.
* Actions Taken:
  + Merged and unified multiple data sources.
  + Reduced categories (e.g., states, positions).
  + Cleaned and normalized textual data.

**3. Exploratory Data Analysis (EDA)**

* Summary Statistics: Provided initial insights into distributions.
* Visualizations:
  + Distribution plots for height and categorical variables.
  + Correlation heatmaps.
  + Boxplots and countplots to understand feature relationships.
* Key Insights:
  + Height, geographic location, position, and year/class significantly differ across NCAA divisions.

**4. Data Cleansing**

* Outliers: Detected via boxplots and IQR; minimal outliers observed.
* Missing Values: Addressed via imputation, minimal missingness found.

**5. One-Hot Encoding**

* Converted categorical features into binary variables, ensuring compatibility with ML algorithms.

**6. Feature Engineering & Selection**

* Enriched Data: Created binary features indicating height above average.
* Standardization: Scaled numeric features for better algorithmic performance.
* Feature Selection:
  + Conducted feature selection using logistic regression with L1 penalty.
  + Identified significant features: division indicators, height, international status, positions, and geographic indicators.

**7. Model Selection and Fine Tuning**

* Model Chosen: Random Forest Classifier.
* Cross-validation: Employed to ensure robustness.
* Hyperparameter Optimization: GridSearchCV utilized to optimize hyperparameters, resulting in:
  + max\_depth: 20, min\_samples\_split: 10, n\_estimators: 150

**8. Model Evaluation**

* Accuracy: Achieved 82%
* Precision: 75% (True Class)
* Recall: 78% (True Class)
* F1-score: 77%
* ROC-AUC: Evaluated for robust understanding of model performance.
* Confusion Matrix: Provided detailed insight into prediction errors.

**Key Findings:**

* + Physical attributes (height), demographic details (year/class), geographic origin, and player positions strongly predict NCAA division classification.