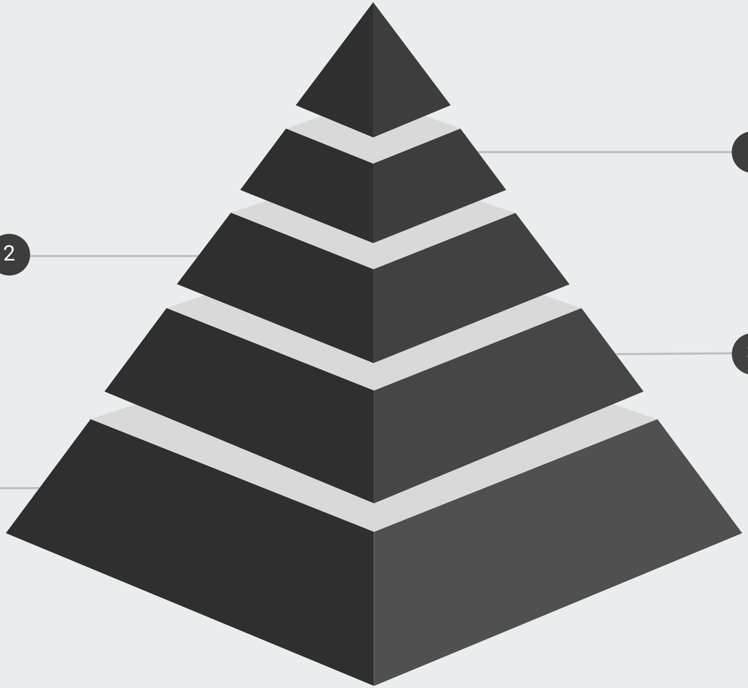
– Data Mining –

Comparative Analysis of Imputation Techniques in Australian Rainfall Data

Progress Presentation by

Leona Hasani
Leona Hoxha
Nanmanat Disayakamonpan

Agenda



1. Implementation Update
2. Analysis and Outcomes
3. New Ideas or Plans?
4. Learning Reflection

Implementation Update

Finding Element	Dataset: Weather - Rain in Australia (Link)
Features	**145k+ observations & 23 variables** (17 categorical, 6 numerical)
Target Variable	**RainTomorrow** (YES or NO)
Missing Values	**Missing at Random** (MAR)
Preprocessing Steps	1. **Dropping row(s) with missing values:** “RainToday”, “RainTomorrow” 2. **Removing unnecessary column(s):** “Date” 3. **Label encoding for categorical variables:** “‘Location”, “WindGustDir”, “‘WindDir9am”, “WindDir3pm”, “RainToday”,, “RainTomorrow” 4. **Checking whether the data is imbalanced?:** No, with 0 (NO) ~78% and 1 (YES) ~22%
Modeling *(Original Data VS Scaled Data)*	1. Logistic Regression 2. Decision Tree Classification 3. Random Forest Classification 4. Gradient Boosting Classification 5. KNeighbors Classification 6. Adaboost Classification

Implementation Update

Person in Charge	First Imputation Approach	Second Imputation Approach	Third Imputation Approach
Leona Hasani	**Mean:** Consists of replacing the missing data for a given variable by **the mean** of all known values of that variable	**Expectation Maximization (EM):** **Iterative means** of imputing one or more plausible missing data a (EM single or multiple imputations) values	**Listwise Deletion (LD):** Statistical method that handles missing data by **deleting or ignoring the entire record of missing values** in a dataset
Leona Hoxha	**Median:** Filling missing values with ***the median*** of the non-missing values in the dataset.	**Multiple Imputation (MICE):** Filling missing values **iteratively** by predicting them **from other variables in the dataset across multiple iterations.**	**Regression:** Filling missing values by predicting them using **regression models based on the observed values of other variables.**
Nanmanat D.	**Mode:** Filling missing values with ***the most frequent*** non-missing values in the dataset.	**K Nearest Neighbour (KNN):** Filling missing values by estimating based on ***the values of the nearest neighbors.***	**Linear Interpolation:** Filling missing values by assuming a ***linear relationship between adjacent data points.***

Analysis and Outcomes

Performance Metrics	Best Imputation Approach	Best Model	Result
Best Accuracy	Mean Technique	Logistic Regression (Not Scaled)	~85.6%
Best Precision	Mice Technique	Random Forest Classifier (Not Scaled)	~82.7%
Best Recall	Regression Technique	Decision Tree Classifier (Not Scaled)	~61.6%
Best F1-Score	Mice Technique	Random Forest Classifier (Not Scaled)	~70.2%
ROC & AUC	Mice Technique	Random Forest Classifier (Not Scaled)	~78.6%
Best Speed / Least Computational Cost	KNN Technique	Logistic Regression (Scaled)	~0.22 second

New ideas or Plans & Learning Reflection

NEW IDEAS OR PLANS

- We would like to delve deeper on some imputation approaches:
  - For **KNN imputation**, we would like to explore more on the process of selecting ***the optimal number of "k"*** and evaluate ***the mean accuracy through cross validation.***
  - For **interpolation**, we would like to investigate the differences between different types of interpolation methods such as ***linear interpolation VS time-series interpolation.***
- Which performance metrics we should prioritize when it comes to choosing the best imputation technique and model?
- In what ways we can further analyse the best performing techniques and possibly improve the results such hyperparameter tuning.
- We would like to focus and investigate the reason behind why a certain model and technique is performing the best in our case.

LEARNING REFLECTION

- Gained experience in ***preprocessing and handling missing values*** in datasets
- Understanding the nature of the missing data in the datasets (**MCAR, MAR, MNAR**)
- Understanding various ***imputation techniques*** *(statistical methods and machine learning models)*
- To develop ***proficiency in using tools*** for data analysis and modeling.
- To comprehend on how to ***evaluate and interpret the performance*** of different imputation strategies.

– Questions & Comments –