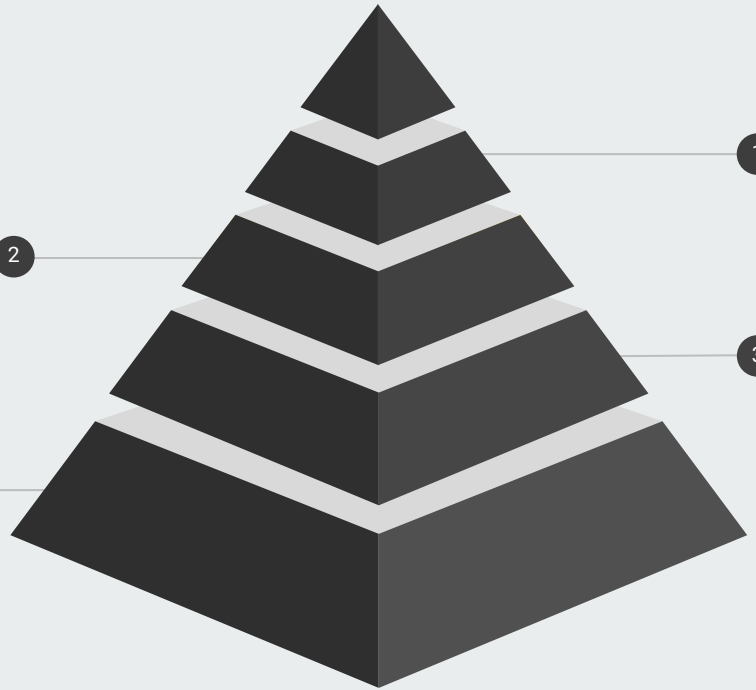
– Data Mining –

Comparative Analysis of Imputation Techniques in Australian Rainfall Data

Proposal Presentation by

Leona Hasani
Leona Hoxha
Nanmanat Disayakamonpan

Agenda



1. Project Introduction
2. Implementation Plan and Literature Resources
3. Purpose of the Project
4. Learning Outcomes

Project Introduction

- **Addressing Common Daily Challenges: What are potential issues that may arise from missing values in a dataset?**
- **Why is it crucial to address missing values in a dataset with a more thorough approach?**
- **What made us pick this dataset?**
- **What does each row in the dataset represent, and what exactly are we trying to find or predict, also known as the target variable?**

Some of the variables/columns in our dataset:

- MaxTemp/MinTemp
- WindSpeed
- Humidity
- Evaporation
- Sunshine
- RainTomorrow

Implementation Plan and Literature Resources

IMPLEMENTATION PLAN

- **Programming Tool:** Python
- Employ various ***imputation techniques*** for missing values (Mean/Mode, KNN, EM, MICE, Hot Deck)
- Use various ***classification models***:
  Logistic Regression, Decision Trees, Random Forest, SVM, ANN, KNN, Naive Bayes and so on
- Apply ***feature selection techniques, hyperparameter tuning, cross-validation***
- Evaluate ***the accuracy scores*** of the classification models

LITERATURE RESOURCES

- Systematic Review of Using Machine Learning in Imputing Missing Values
- A Comparison of Strategies for Missing Values in Data on Machine Learning Classification Algorithms
- A survey on missing data in machine learning
- Impact of imputation of missing values on classification error for discrete data

Purpose of the Project & Learning Outcomes

PURPOSE

- Address ***the challenges of missing values*** in the data, a common issue in real-world datasets.
- Experiment with ***diverse imputation methods***
- Enhance our understanding of ***how to handle missing data effectively.***
- Improve ***data quality and reliability*** in predictive modeling.

LEARNING OUTCOMES

- To gain knowledge in ***preprocessing and handling missing values*** in the datasets
- To understand 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 –