Online Revision Session for Lab 1

→ Health Analytics Lab Notes :

- Split 60:40 between Notebook & Report

↳ Non-technical

- Characteristics
- Meaning of the report
- Explain what the output/results means
- What is valuable within the notebook

& Technical

- Dataset :
  - → Date & time features
  - → Time series data
  - → cross sectional data
- Table of content on jupyter Notebook
- Problem statement (1-2 sentences)
- EDA
  - ↳ overview
  - ↳ Shape of data
  - ↳ Data type
  - ↳ Descriptive stats (df.info & df.des

understand the distribution

↳ Visualisations

- Univariate Analysis :

  Histogram or bar graph

  Box plots

- Bivariate Analysis

  Scatter plot for continous vs. categorical variables

- Correlation Analysis

  Heat Map or pair plot to visualize correlation

- Data Pre-posessing :

⊗ Encoding Categorical Variables

⊗ Feature Engineering

⊗ Try cat-boost in feature selectio

Gradient feature also code.

- Modelling :

① Selection of model → select appropriate algorithm. For health analytics = Logistic Regression, Random forest, Gradient boost, Neural Network

②

③ ROC_AUC plot, PR_AUC plot
↳ Good submissions
↳ Used for im-balanced data

④ Hyper parameter Tuning

GridSearchCV or RandomizedSearch CV to find best hyperparameters for your model.

⑤ Metrics :
Don't only use one metric to evaluate performance

- Precision
- Recall
- F_1-score
- ROC-AUC
- Confusion matrix
- Regression - MSE, RMSE, MAE, R-Squared

- Conclusion

↳ Summerize findings
↳ Discuss implications of model results in context of health analytics