Departement Informatika
Department of Informatics

Assignment 01

Data Preparation and Supervised Learning

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

University of Pretoria

Neetu Ramsaroop, PhD

□ **Length (5 - 10). Excluding the cover page and reference page**

□ **Type: PDF file only. The page must include the student number and the student's proper name**

Submission Date: 05 September 2025

Submission Method: Online only

□ **Scenario:** You are working as a junior data analyst for a South African bank. Your task is to explore and prepare historical marketing campaign data and to assist in building a machine learning model to predict whether a client will subscribe to a term deposit.

□ Learning Objectives:

By completing this assignment, you will:

- Apply real-world data preparation techniques
- Perform exploratory data analysis and feature engineering
- Encode and preprocess data for modeling
- Train and evaluate a classification model
- Generate and interpret model evaluation metrics

□ Assignment Tasks:

Part 1 – Data Preparation

1.1 Data Cleaning

- Load the dataset into the Jupyter Lab environment.
- Identify and handle missing or "unknown" values in the dataset:
  - Search for missing values (NaN) or placeholder values such as 'unknown' in the dataset.
  - Describe how many such values exist in each affected column.
  - Explain your strategy for handling them (e.g., removal, replacement, or imputation) and justify your choice based on data quality and relevance.
  - Provide the code used and summarize the dataset after cleaning.
- Identify and categorise the features in the bank dataset as either numerical or categorical.
  - Provide the code used to perform this task, and include a table in your report listing each feature, its type (numerical or categorical), and a brief note or description.

1.2 Feature Engineering: New Features

- Create at least two new features from the existing dataset to enhance model performance:
    - Engineer at least two meaningful new features using the existing data (e.g., contacted_before, age_group, or other relevant transformations).
    - Clearly describe the logic and purpose behind each new feature:
    - What problem does it solve or insight does it provide?
    - Why might it help in predicting term deposit subscriptions?
    - Include the code used to create these features.
    - Add the new features to your DataFrame and display a sample of the updated data.
- Document and explain your feature creation process.

1.3 Basic Statistics

- Perform basic statistical analysis and explore correlations in the dataset.

a) Generate Summary Statistics

- Use appropriate methods to compute summary statistics (mean, median, std, min, max, etc.) for numerical features.
- Present the output in your notebook and briefly interpret any notable patterns or outliers.

b) Analyse Feature Correlations

- Use a correlation matrix to explore relationships between numerical variables.
- Identify and discuss any strong positive or negative correlations that may impact feature selection or model performance.
- Provide interpretations in context, which variables appear to influence others, and how might this guide modeling decisions?

1.4 Data Visualisation

- Create visualisations to explore data distributions and relationships with the target variable.

a) Generate at least four relevant plots to help understand the dataset. You may include:

- Histograms
- Bar charts
- Boxplots
- Heatmaps

b) Use visualisations to explore relationships between features and the target variable.

- Identify which features appear to have the strongest relationship with subscription outcomes.

- Describe at least two insights that could influence feature selection or model building.

Note: *Include all plots in your* ***notebook and PDF report*** *with appropriate titles, labels, and brief captions explaining what each plot reveals.*

1.5 Data Preprocessing: Additional Data Transformations

- Apply any necessary transformations to clean or enhance the data.
- Drop irrelevant or leakage-prone features like duration
- Convert newly engineered features if needed
- Ensure consistent data types and formatting.
- Briefly describe and justify each transformation step.
- Ensure the final dataset is clean, numeric, and ready for model input.

1.6 Data Encoding

- Encode features to prepare the dataset for machine learning models.

a) Apply One-Hot Encoding

- Identify all nominal categorical variables.
- Apply one-hot encoding to convert these into a numeric format.
- Show the resulting column names and verify that the transformation is correct.

b) Ensure the Dataset is Model-Ready

- Confirm that all features are numeric after encoding.
- Print the final dataset structure and data types to verify readiness for modeling.

Part 2 – Supervised Learning

2.1 Train a classification model to predict y (term deposit subscription: yes/no).

2.2 You may use Logistic Regression and Random Forest, or any classification algorithms of your choice. **In the experiments, compare the performance of the two algorithms**.

2.3 Evaluate your model using the following metrics:

- Accuracy
- Precision
- Recall
- F1 Score
- ROC-AUC score and curve plot
- Computational time

2.4 Briefly discuss the model's performance and how data preparation impacted results.

Submission Instructions

Each student must submit the following:

- Report (PDF or Word). Your report must include:
  - Introduction and problem overview
  - Summary of each data preparation step
  - Visualisations with captions
  - Model choice and justification
  - Evaluation metrics and interpretation
  - Key findings and recommendations for the bank
- Preprocessed Dataset (CSV)
  - Submit a .csv file of your final, preprocessed dataset (after encoding and cleaning but before modeling).
  - File name: yourname_preprocessed_bank_data.csv
- Jupyter Notebook or Python Script
  - Submit your complete notebook or .py file showing all steps.
  - Make sure your code is clean, well-commented, and follows the task sequence.

Rubric

	Score	**Comment**
Part 1: Data Preparation (70)		
1.1 Data Cleaning (15) - Missing/Unknown Values (5) - Feature Type Identification (5) - Code Quality and Summary (5)	5 5 5	○ Missing and 'unknown' values correctly identified, summarized, and handled with clear justification ○ All features accurately categorized with supporting code and descriptive table ○ Clean, efficient code; clear summary of dataset post-cleaning
1.2 Feature Engineering (10) - Feature Logic & Relevance (5) - Implementation & Documentation (5)	5 5	○ Two or more meaningful features created, well-justified and relevant to prediction task ○ Code correctly implements features; sample output and explanation provided
1.3 Basic Statistics (10) - Summary Statistics (5) - Correlation Analysis (5)	5 5	○ Used correct methods (e.g., .describe()); clear interpretation of key stats/outliers ○ Accurate heatmap and correlation matrix with insightful analysis
1.4 Data Visualisation (15) - Variety and Relevance (5) - Insight into Target Variable (5) - Assess skewness (3)	5 5 3	○ 4+ plots included with a strong variety and relevance to analysis ○ Visuals demonstrate relationships with the target and aid feature selection ○ Data distribution must first be examined to assess skewness. Relevant methods explained.

- Presentation Quality (2)	2	○ Titles, labels, captions used appropriately for clarity
1.5 Data Preprocessing (10 points) - Transformations Applied (5) - Dataset Readiness (5)	 5 5	 ○ Relevant transformations (e.g., dropping leakage features) correctly applied and justified ○ Final dataset is clearly numeric, clean, and model-ready
1.6 Data Encoding (10) - One-Hot Encoding Applied (5) - Model Readiness Confirmed (5)	 5 5	 ○ All nominal categorical features encoded correctly; output verified ○ Dataset structure clearly shown, all numeric features
Part 2: Supervised Learning (30)		
2.1–2.2 Model Training and Selection (10) - Model Training (5) - Model Choice Justification (5)	 5 5	 ○ Suitable classification algorithm implemented correctly ○ Algorithm choice well-justified based on data
2.3 Model Evaluation (15) - Evaluation Metrics (5) - ROC Curve (5) - Performance Interpretation (5)	 5 5 5	 ○ All metrics (Accuracy, Precision, Recall, F1, ROC-AUC, computational time) computed and interpreted ○ ROC curve correctly plotted and explained ○ Evaluation clearly linked to model choice and data prep
2.4 Discussion of Results (5) - Impact of Data Preparation	 5	 ○ Clear connection made between data cleaning/feature engineering and model performance
Total	**100**	

Suggested Report Structure (10 Pages)

Cover Page (1 page)

Title: e.g., Data Preparation and Classification Model for Term Deposit Subscription Prediction

Name, Student ID, Course, Date

Assignment 1

1. Introduction (0.5 – 1 page)

- Brief overview of the dataset and the goal
- Purpose of the assignment
- Outline of the report structure

2. Data Preparation (3 – 5 pages total)

2.1 Data Cleaning (1 – 1.5 pages)

- Description of missing/unknown values
- Strategy and justification
- Code summary and cleaned dataset overview
- Feature categorization table (numerical vs. categorical)

2.2 Feature Engineering (0.5 – 1 page)

- Description of new features and their logic
- Sample data before/after transformation

2.3 Basic Statistics (0.5 – 1 page)

- Summary statistics with interpretations
- Correlation analysis with heatmap

2.4 Data Visualization (0.5 – 1 page)

- 4 visualizations (histogram, bar chart, boxplot, heatmap)
- Short interpretation of insights from each
- For all numerical variables, the *data distribution* must first be examined to **assess skewness**. Plot the distribution of variables exhibiting significant skewness, and normalize them using appropriate **transformation techniques such as logarithmic, square root, or Box-Cox transformations** to reduce skewness. Plot the normalized distribution variables, reducing skewness. After normalization, all numerical variables must be scaled (e.g., using Min-Max) to **bring them into a comparable range**. Finally, generate a plot to visualize the distribution of all normalized and scaled variables in a single graph(s), facilitating comparison across features and categories (Figure 1). Discuss the impact of this process in real life for a Bank.

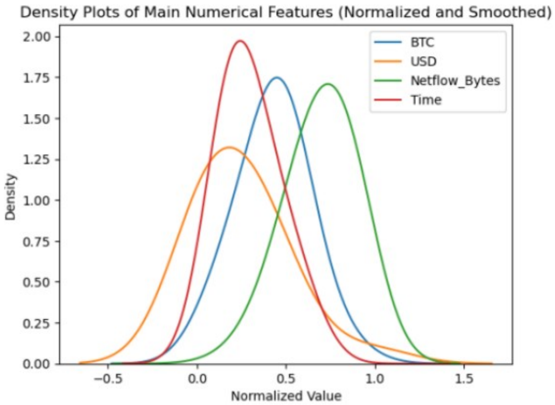

Figure 1.Normalized Features

2.5 Preprocessing and Encoding (0.5 – 1 page)

- Transformations applied

- Encoding steps and final dataset readiness

3. Supervised Learning (1 – 2 pages)

3.1 Model Training and Choice (0.5 page)

- Algorithm selected (e.g., Logistic Regression vs. Random Forest)
- Reason for selection

3.2 Model Evaluation (1 page)

- Metrics results: Accuracy, Precision, Recall, F1 Score, computational time, and ROC-AUC
- ROC Curve plot with brief analysis

3.3 Interpretation (0.5 page)

- Discussion on how data preparation influenced results
- Final reflections on model performance

4. Conclusion (0.5 page)

- Recap of key findings
- Possible improvements or next steps (e.g., hyperparameter tuning, feature selection)

Appendices (Optional, not counted toward 10 pages)

- Full code listings (if not inline in report)
- Extra plots or detailed tables

Tips to Stay Within 10 Pages:

- Use figures and tables efficiently (combine multiple visuals into one if needed).
- Keep code snippets short in the main report; link or refer to full code in the appendix.
- Use concise bullet points or tables for explanations where appropriate.

Referenced Information. To be able to complete this assignment successfully, the student should reference the following sources, in the following order of priority:

Due Dates & Times. **ASSIGNMENTS SUBMITTED LATE WILL NOT BE MARKED. NO EXCEPTIONS, NO EXCUSES.**

Submission Instructions. This is an ***INDIVIDUAL ASSIGNMENT***. All assignments will be submitted to Turnitin to assess the originality or the similarity of assignment content. If duplication is found between different sets of project documentation, then the person involved will be penalized.

While AI tools like ChatGPT can assist you, it is crucial that your submissions reflect your own understanding and analysis. Copying or heavily relying on AI-generated content undermines your learning experience and academic integrity. The use of AI tools as a supplement to your own thinking will be penalized.

Assignment Layout and Appearance

□ No handwritten work will be accepted.

□ The manuscript should contain a title page (should be captivating), your name, student number, the course name and instructor's name. The manuscript should only contain an introduction, body, and conclusion.

□ All graphs, images or diagrams should be clearly visible.

- Content on the graphs, images or diagrams should be clearly legible.
- Graphs, images or diagrams should be clearly labeled and entitled.

□ Assignment Referencing.

- References are required for this assignment.
- If a student makes use of any sources, it will be required for students to reference the sources used.
- If a student makes use of direct quotations from any source, then it will be required that the student make use of Harvard referencing and appropriate citation mechanisms.

Penalties. Penalties will also apply to issues relating to page length, spelling, punctuation, grammar, and figure manipulations.