PDAN8411 POE PART 2

Makabongwe Lwethu Sibisi

ST10145439

Table of Contents

QUESTION 1 2

Dataset Selection 2

QUESTION 2 3

Algorithm Justification: 3

QUESTION 3 4

Exploratory Data Analysis 4
Model Training 6
Interpret and evaluate model 7
Write a report 8

QUESTION 4 9

QUESTION 5 9

QUESTION 6 9

Question 1

Dataset Selection

When building a classification algorithm for a medical aid scheme, the first step is to identify features with strong predictive power. These features should include demographic factors such as age and gender. lifestyle indicators like smoking history and environmental exposures. Clinical markers such as nodule size and pulmonary test results, and symptom profiles.

Next, the quality of the data should be evaluated. Ensuring that the datasets are cleanly structured with consistent numeric and categorical data types, minimal missing values, and no duplicates or formatting errors. Common issues that can compromise accuracy include class imbalance (like too many man vs female participants), outliers distorting analysis, and hidden confounders. To address these issues, apply stratified sampling, rigorous outlier detection, and dimensionality-reduction techniques. By proactively resolving these issues, we can ensure that the algorithm trains on reliable and actionable data.

I chose Nancy Al Aswad's "Lung Cancer Dataset" because it meets these criteria. The dataset contains 309 records organised for clarity, including critical risk factors such as age, gender, smoking intensity (pack-years), and environmental exposures, all of which are directly relevant to medical aid risk assessment. The Kaggle data card confirms balanced distributions across gender, smoking status, and alcohol consumption, which helps minimise bias during training. While the dataset is well-organised, I will conduct exploratory analysis (EDA) to verify feature ranges, resolve any overlooked duplicates or missing values, and confirm class distributions. This ensures the algorithm operates efficiently and transparently, meeting the medical aid scheme's need for trustworthy and equitable predictions. (Sakshi, 2023) (Chapagain, 2010) (Viyaleta, 2025)

Question 2

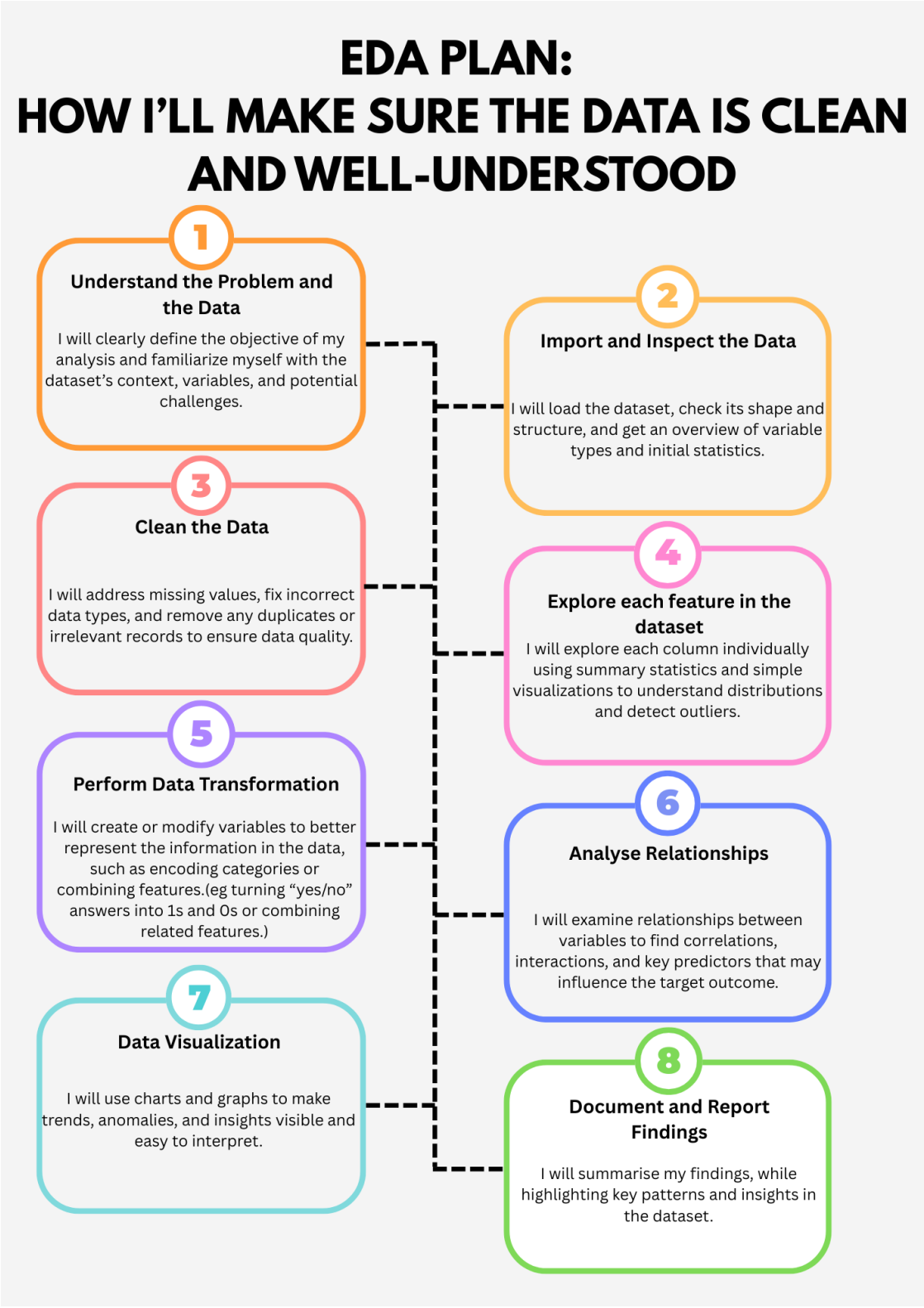
Algorithm Justification:

I choose a Decision Tree classifier because it delivers the ideal combination of speed, accuracy, and interpretability for a real-time cancer-benefits system. The medical scheme needs a system that can make fast and clear decisions, as highlighted by their mission to “help in speeding up their ability to apply their dreaded disease benefits for the customers that need it.” A Decision Tree meets this demand. Its inference process which involves traversing only a few nodes, operates almost instantly with minimal computational cost. Ms Nancys’ lung-cancer dataset’s tabular format, combining categorical predictors (gender, smoking history, binary symptom flags) and a single continuous variable (age), aligns perfectly with a tree’s split logic, eliminating complex preprocessing. Moreover, with just 309 records (and potentially fewer once we clean and filter the data), a single Decision Tree is ideal. It can learn meaningful splits without the massive data requirements of ensemble or deep-learning algorithms, and retrains almost instantly.

While alternatives like Random Forests could marginally improve precision through ensemble averaging, their multi-tree architecture introduces inference latency, slowing down real-time decisions. Similarly, Naive Bayes might predict slightly faster but sacrifices accuracy due to unrealistic assumptions about feature independence. A single decision tree avoids these pitfalls, offering a high-speed, auditable prototype that clinicians can deploy immediately and validate transparently. (Geeks for Geeks, 2025) (Geeks for Geeks, 2025)

Question 3

Exploratory Data Analysis



(Geeks for Geeks, 2024)

Feature selection Plan

To select the most relevant features for my Decision Tree model, I will start by encoding all categorical values using LabelEncoder or one-hot encoding as needed. Then, I'll perform a correlation analysis to identify how each feature relates to the target variable (lung cancer). Next, I'll apply backward elimination using p-values from a Logistic Regression model to remove features that are not statistically significant. Finally, I will confirm my selection using univariate feature selection with SelectKBest from sklearn, which ranks each feature based on its strength of relationship with the target variable (lung cancer). This ensures that my feature selection process is fully data-driven, unbiased, and optimized for building a clear and effective classification model. (Geeks for Geeks, 2024)

Model Training

Following the exploratory data analysis (EDA) and feature selection phases. The cleaned data and selected features, validated via correlation analysis, backward elimination using p-values, and SelectKBest, will form the foundation for model training. As outlined in the feature selection phase, categorical variables will be encoded using scikit-learn's LabelEncoder to ensure numerical compatibility.

The dataset will be split using an 80/20 stratified train-test split to preserve the distribution of the target variable (lung cancer occurrence), which is critical for addressing potential class imbalances. Feature scaling will be omitted as decision trees are scale-invariant (Olamendy, 2023)

Model training will proceed in two key stages:

Baseline Establishment:

An initial DecisionTreeClassifier (scikit-learn) will be trained with default parameters to benchmark performance.

Hyperparameter Optimization:

Using GridSearchCV with 5-fold cross-validation, critical parameters will be tuned to balance accuracy and interpretability:

- max_depth (controls tree complexity to prevent overfitting)
- min_samples_split (sets minimum samples required for node splitting)
- criterion (gini or entropy for split quality evaluation)
- class_weight (adjusts for class imbalances if detected in EDA)

The optimal hyperparameters identified will be used to retrain the final model on the full training set.

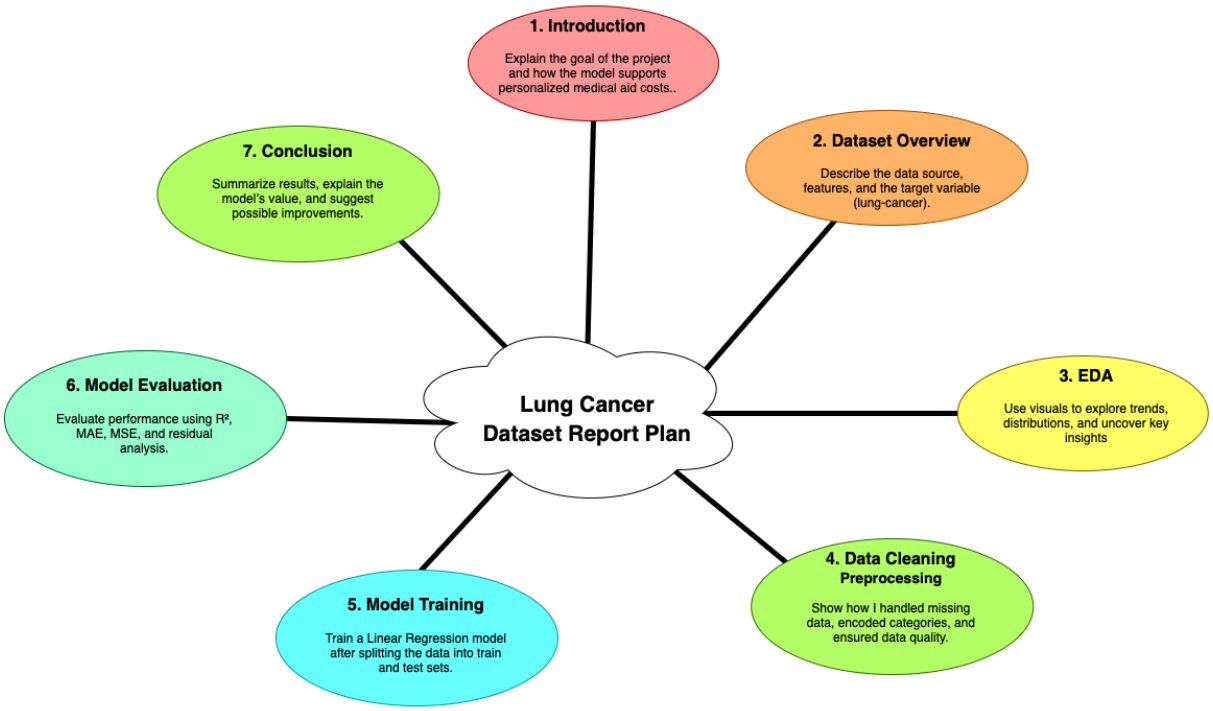
This structured training approach ensures the classifier aligns with the medical scheme's dual objectives: operational speed (leveraging decision trees' inherent efficiency) and transparency (maintaining interpretability for benefit approval audits). (Hoffman, 2020)

Interpret and evaluate model

Step	Description	Tool/Method
1. Evaluate Accuracy	Measure how well the model performs on the test set	accuracy_score from sklearn.metrics
2. Check Confusion Matrix	Understand true/false positives and negatives	confusion_matrix, ConfusionMatrixDisplay
3. Assess Precision & Recall	Evaluate model's ability to correctly identify cancer vs. false alarms	precision_score, recall_score
4. F1 Score	Balance precision and recall into one metric	f1_score
5. ROC Curve & AUC	Visualize performance across all thresholds; good for binary classification	roc_curve, roc_auc_score
6. Feature Importance	Identify which features contributed most to predictions	.feature_importances_ attribute in DecisionTree
7. Model Interpretability	Visualize tree for transparency and auditability	plot_tree from sklearn.tree
8. Cross-Validation Scores	Ensure consistent performance across different data splits	cross_val_score (with k-fold = 5)

(IBM, 2025)

Write a report



Question 4

Please find Reponses within the github repo. The following YouTube videos were used as reference for the code.

https://github.com/VCDN-2025/pdan8411-part-2-just-makab.git

4a. (Keith, 2020)

https://youtu.be/78ut-S-QOEQ?si=LjyM8S3I30C2fIom

4b. (Genius, 2020)

https://youtu.be/HYcXgN9HaTM?si=RRUXe3dTDinbMZ1E

4c. (Turp, 2022)

https://youtu.be/wxS5P7yDHRA?si=ccxZEMAxFSs5ZUNJ

Question 5

https://github.com/VCDN-2025/pdan8411-part-2-just-makab.git

Question 6

Please refer to the Report Document.

Bibliography

Sakshi, B., 2023. *Exploring Classification Algorithms: Guide to Select the Right Model for Your Data*. [Online]
Available at: https://medium.com/@sakshi.babbar/exploring-classification-algorithms-guide-to-select-the-right-model-for-your-data-73b08b187a01
[Accessed 26 May 2024].

Chapagain, M., 2010. *Which machine learning classifier to choose, in general? [closed]*. [Online]
Available at: https://stackoverflow.com/questions/2595176/which-machine-learning-classifier-to-choose-in-general
[Accessed 26 May 2025].

Viyaleta, A., 2025. *Choosing Classification Model Evaluation Criteria*. [Online]
Available at: https://towardsdatascience.com/choosing-classification-model-evaluation-criteria-1e1c0f6f13ce/
[Accessed 26 May 2025].

Geeks for Geeks, 2025. *Random Forest Algorithm in Machine Learning*. [Online]
Available at: https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/
[Accessed 26 May 2025].

Geeks for Geeks, 2025. *Naive Bayes Classifiers*. [Online]
Available at: https://www.geeksforgeeks.org/naive-bayes-classifiers/
[Accessed 26 May 2025].

Geeks for Geeks, 2024. *Steps for Mastering Exploratory Data Analysis | EDA Steps*. [Online]
Available at: https://www.geeksforgeeks.org/steps-for-mastering-exploratory-data-analysis-eda-steps/#step-2-import-and-inspect-the-data
[Accessed 27 May 2025].

Geeks for Geeks, 2024. *Feature Selection in Python with Scikit-Learn*. [Online]
Available at: https://www.geeksforgeeks.org/feature-selection-in-python-with-scikit-learn/
[Accessed 27 May 2025].

Olamendy, J. C., 2023. *Decision Trees: Titans of simplicity & power*. [Online]
Available at: https://medium.com/@juanc.olamendy/decision-trees-titans-of-simplicity-power-188ddb32dfb2
[Accessed 27 May 2025].

Hoffman, K., 2020. *Decision Tree Hyperparameters Explained*. [Online]
Available at: https://ken-hoffman.medium.com/decision-tree-hyperparameters-explained-49158ee1268e
[Accessed 17 May 2025].

IBM, 2025. *What is a decision tree?*. [Online]
Available at: https://www.ibm.com/think/topics/decision-trees
[Accessed 27 May 2025].

Keith, M., 2020. *Python: univariate statistics*. [Online]
Available at: https://www.youtube.com/watch?v=78ut-S-QOEQ&list=PLe9UEU4oeAuV7RtCbL76hca5ELO_IELk4
[Accessed 27 May 2025].

Genius, K. T., 2020. *Logistic Regression in Python Step by Step in 10 minutes*. [Online]
Available at: https://youtu.be/HYcXgN9HaTM?si=FevSLwOxtKbcv5EL
[Accessed 27 May 2025].

Turp, M., 2022. *How to Implement Decision Trees in Python (Train, Test, Evaluate, Explain)*. [Online]
Available at: https://youtu.be/wxS5P7yDHRA?si=ccxZEMAxFSs5ZUNJ
[Accessed 28 May 2025].