Final Report — UCI Heart Disease Classification Project

This project was carried out collaboratively by three team members, each contributing to specific parts of the pipeline:

- **Baptise Fillie-Santin** *Responsible for dataset acquisition and project report.* He identified and prepared the UCI Cleveland dataset and documented the overall workflow.
- **Gaspard Martouzet** *Responsible for enhancing model evaluation and export functionality.* He improved the evaluation metrics, automated the export of results, and ensured reproducibility through model saving and reporting.
- **Simon Jennequin** *Responsible for EDA and preprocessing.* He performed exploratory data analysis, generated statistical summaries and visualizations, and implemented preprocessing steps such as imputation and scaling.
- You can find the complete project and source code on our GitHub repository: https://github.com/baptoufs/Machine_Learning_Project.git

Phase 1 : Data Exploration and Preprocessing

Dataset Selection

- **Source:** UCI Machine Learning Repository — Heart Disease dataset.
- **Problem Type:** Classification (predicting presence of heart disease).
- **Features:**
    - **Numerical:** age, trestbps (resting blood pressure), chol (serum cholesterol), thalach (max heart rate), oldpeak (ST depression).
    - **Categorical/Binary:** sex, cp (chest pain type), fbs (fasting blood sugar), restecg (ECG results), exang (exercise-induced angina), slope, ca (major vessels), thal (thalassemia).
    - **Target:** heart disease presence (0–4 in raw dataset, simplified to binary for modeling).
- **Real-world relevance:** Early detection of cardiovascular disease is critical for preventive medicine and clinical decision-making.

Exploratory Data Analysis (EDA)

- **Summary statistics:**
    - Average age: 54.4 years (range 29–77).
    - 68% male.
    - Average cholesterol: 246.7 mg/dL.
    - Average resting blood pressure: 131.7 mmHg.
    - Max heart rate: mean 149.6 bpm.
- **Visualizations:**

- **Target distribution (Figure: Target Distribution):** Class imbalance — majority class 0 (~165 cases), minority classes 1–4 (20–55 cases each).
- **Correlation heatmap (Figure: Correlations):** cp (chest pain type) strongly correlated with target; thalach negatively correlated with exang; oldpeak positively correlated with exang.

Preprocessing

- **Missing values:** ca (4 missing), thal (2 missing) → imputed with mode.
- **Scaling:** StandardScaler applied to continuous features (age, trestbps, chol, thalach, oldpeak).
- **Encoding:** Binary features already encoded; categorical features (cp, slope, thal) one-hot encoded for linear models.
- **Rationale:** Scaling improves KNN and Logistic Regression; imputation preserves dataset size; encoding avoids artificial ordinal relationships.

Phase 2: Model Implementation and Evaluation

Models Implemented

1. **K-Nearest Neighbors (KNN)**
   - Hyperparameters: n_neighbors=5, weights=uniform.
   - Confusion Matrix (Figure: KNN Confusion Matrix): TN=26, FP=7, FN=0, TP=28.
   - ROC Curve (Figure: KNN ROC): AUC=0.923.
   - Report (`KNN_report.txt`): Precision=0.80, Recall=1.00, F1=0.89, Accuracy=0.89.
2. **Logistic Regression**
   - Hyperparameters: C=0.01.
   - Confusion Matrix (Figure: Logistic Regression Confusion Matrix): TN=29, FP=4, FN=3, TP=25.
   - ROC Curve (Figure: Logistic Regression ROC): AUC=0.961.
   - Report (`LogisticRegression_report.txt`): Precision=0.86, Recall=0.89, F1=0.88, Accuracy=0.89.
3. **Random Forest**
   - Hyperparameters: max_depth=5, n_estimators=200.
   - Confusion Matrix (Figure: Random Forest Confusion Matrix): TN=29, FP=4, FN=2, TP=26.
   - ROC Curve (Figure: Random Forest ROC): AUC=0.957.
   - Report (`RandomForest_report.txt`): Precision=0.87, Recall=0.93, F1=0.90, Accuracy=0.90.

Comparative Results

Model	F1	Precision	Recall	AUC	Best Params
Logistic Regression	0.877	0.862	0.893	0.961	C=0.01
Random Forest	0.897	0.867	0.929	0.957	max_depth=5, n_estimators=200

KNN	0.889	0.800	1.000	0.923	n_neighbors=5, weights=uniform

Interpretation:

- KNN achieves perfect recall (no false negatives) but lower precision.
- Logistic Regression provides the highest AUC and interpretability.
- Random Forest achieves the best overall F1-score and balance between precision and recall.

Phase 3: Insights, Recommendations, and Future Work

Insights

- **Key predictors:** chest pain type (cp), exercise-induced angina (exang), ST depression (oldpeak), slope, and max heart rate (thalach).
- **Model trade-offs:**
  - KNN is sensitive to scaling and maximizes recall.
  - Logistic Regression is interpretable and well-calibrated.
  - Random Forest captures non-linear relationships and provides robust performance.

Recommendations

- **Best model for deployment:** Random Forest (highest F1 and balanced metrics).
- **Clinical use case:** Logistic Regression may be preferred when interpretability is critical (e.g., explaining risk factors to clinicians).
- **Threshold tuning:** Adjust probability thresholds to optimize recall vs precision depending on clinical priorities (e.g., minimize false negatives in screening).

Limitations and Future Work

- **Dataset imbalance:** Majority class dominates; could bias models.
- **Population bias:** 68% male — results may not generalize across genders.
- **Future improvements:**
  - Apply resampling techniques (SMOTE) to balance classes.
  - Explore multi-class classification (target values 0–4).
  - Use SHAP values for feature importance and explainability.
  - Validate models on external datasets for robustness.