以下是加上問號後的 `REPORT.md` 模板：

HW2 Report

Table of Contents

Introduction

- **Objective**：本次作業的目標是基於醫院入院資料構建一個準確的分類模型，以預測病人的生存率。目標變量為二元變量（`has_died`）。
- **Dataset**：數據集包括訓練集和測試集，訓練集有 44,939 筆樣本，測試集有 19,260 筆樣本，包含人口統計和醫學檢查數據。
- **Evaluation Metric**：模型性能將依據 Macro F1-Score 進行評估，並輔以其他評估指標。

Data Preprocessing

1. **Dataset Analysis**
    - 每個特徵的關鍵統計量是什麼（例如平均值、中位數、標準差）？
    - 使用了哪些視覺化方法？如盒鬚圖、直方圖和相關矩陣。
    - **Observations**：有無初步的洞察或異常值？

2. **Data Cleaning**
    - 如何處理缺失值、異常值和其他異常數據?
    - 清理數據時做了哪些假設和使用了什麼具體閾值?
3. **Data Transformation**
    - **Feature Encoding**：進行了哪些類別特徵的編碼（如 one-hot 編碼）？
    - **Normalization/Standardization**：是否進行了標準化或正規化？原因為何?
4. **Data Imputation**
    - 使用了哪些方法來填補缺失值（如平均值插補、KNN 插補）？
    - 為什麼選擇這些方法？對數據分佈有何影響?
5. **Handling Data Imbalance**
    - 使用了哪些方法來處理數據不平衡（如 SMOTE、欠採樣）？
    - 平衡技術對目標分佈有什麼影響?
6. **Visualization Analysis**
    - 使用摘要統計和分佈圖分析數據模式的結果如何?
    - 哪些視覺化方法被用來展示數據，如盒鬚圖、直方圖和相關矩陣?

Classification Methods

1. **Model Selection**
    - 使用了哪些機器學習算法（如 Logistic Regression, Random Forest, XGBoost）？為何選擇它們?
    - **References**：使用了哪些算法參考文獻?
2. **Implementation Details**
    - 模型訓練中使用了哪些關鍵參數和具體的配置?
    - 如何確保模型訓練的可重現性?
3. **Reproducibility**
    - 重現模型訓練的指令或腳本是什麼?
    - 使用的軟件包和庫的版本有哪些?

Cross-Validation & Evaluation

1. **KFold Cross-Validation**

    - 使用了多少折進行交叉驗證（如 5 折）？
    - **Metrics**：在各折中計算了哪些指標的平均值，如 AUROC 和 Macro F1-Score?
    - 在交叉驗證過程中遇到了哪些挑戰或收穫?
2. **Evaluation Metrics**
    - **F1-Score**：為什麼 F1-Score 在此任務中重要?
    - **AUROC**：AUROC 指標的意義是什麼？為什麼與此任務相關?
3. **Results Snapshot**
    - AUROC 和 Macro F1-Score 的交叉驗證結果如何?

Explainable Experiment

1. **Feature Importance**
    - 使用了什麼方法來計算特徵重要性（如 SHAP 值）？
    - 列出的最重要的前 20 個特徵是什麼？每個特徵有何描述?
2. **Analysis & Insights**
    - 從這些重要特徵中獲得了哪些洞察?
    - 這些特徵對模型性能有何影響?

Results & Analysis

1. **Performance on Testing Set**
    - Kaggle 提交的最終 Macro F1-Score 是多少?
    - 模型的表現與基準表現相比如何?
2. **Comparison with Baseline**
    - 與基準分數相比，結果如何?
    - 提交後進行了哪些性能改進?
3. **Discussion of Errors**
    - 常見的錯誤分類有哪些？原因為何?
    - 未來迭代有什麼可能的改進?

Conclusion

- 專案的關鍵收穫有哪些?

- 此方法的優點和缺點是什麼?
- 未來的工作及改進方向有哪些?

References

- 使用了哪些參考文獻，包括論文、軟件包文件和 Kaggle 資源鏈接?

Additional Notes:

- 根據需要在附錄中添加截圖或額外的視覺化資料。
- 所有圖表和表格是否已編號並加上標題，便於引用?

這個模板包含了每個部分的提問，幫助撰寫時引導出更詳細的分析和回應。