ChatGPT Review: CRISP-DM

Based on the initial review of your Colab notebook, here's a phase-by-phase critique of your work, followed by additional general comments:

Phase-by-Phase Critique

1. Business Understanding

- **Strengths:**
  - The explanation of CRISP-DM and the business objective is clear and well-structured.
  - You provided a clear link between predicting income levels and its value for various institutions.
- **Suggestions for Improvement:**
  - Consider adding more detail about the potential business implications of misclassifications (e.g., false positives vs. false negatives) to deepen the connection between the model's performance and business outcomes.

2. Data Understanding

- **Strengths:**
  - The initial steps of importing libraries and loading the dataset are well-documented.
  - The setup for data exploration and initial statistics about the dataset are mentioned.
- **Suggestions for Improvement:**
  - Ensure you have a detailed EDA, including visualizations of distributions, relationships between features, and correlations.
  - Mention any notable findings from the exploratory data analysis explicitly. This helps to clearly identify issues like missing values, skewness, or other patterns before moving to the next phase.

3. Data Preparation

- **Strengths:**
  - The code for handling missing values, encoding, and scaling appears to be in place.
  - Feature engineering techniques seem to be covered.
- **Suggestions for Improvement:**
  - Clearly state which encoding techniques were used for each categorical variable (e.g., label encoding, one-hot encoding) and why.
  - Consider describing any specific outliers removed or transformations applied (e.g., log transformation), along with a brief rationale for each step.

4. Modelling

- **Strengths:**
  - A variety of models have been implemented, including Logistic Regression, Decision Tree, Random Forest, SVM, and XGBoost.
  - Hyperparameter tuning is evident, suggesting a good understanding of model optimization.
- **Suggestions for Improvement:**
  - Add a summary table comparing model performance (accuracy, precision, recall, F1-score) for easier visualization.
  - Consider adding comments to describe why certain models performed better or worse than others based on dataset characteristics.

5. Evaluation

- **Strengths:**
  - Metrics like accuracy, precision, recall, F1-score, and AUC-ROC are used to evaluate models, showing a comprehensive approach.
- **Suggestions for Improvement:**
  - Add more detail to the evaluation results, such as the impact of false positives/negatives on the business objective.
  - Consider additional evaluation metrics like confusion matrix visualizations or precision-recall curves for models that may handle imbalanced classes differently.

6. Deployment

- **Note:** Since deployment is not required for this assignment, you did well to focus on other phases.
- **Suggestions for Improvement:**
  - You could include a brief note about potential deployment strategies as a placeholder for completeness, even though it's not necessary to implement them.

General Comments

- **Clarity & Organization:** The notebook is generally well-structured with clear phase titles, making it easy to follow the CRISP-DM methodology.
- **Code Readability:** Consider adding more comments within the code cells to explain specific transformations or modelling decisions.
- **Visualizations:** More data visualizations (e.g., histograms, bar charts, correlation matrices) would enhance the Data Understanding and Evaluation phases by providing more insights into feature distributions and model performance.
- **Documentation:** Add a final summary that recaps the model performance, key findings, and possible next steps for further improvement or deployment (if applicable).