Artificial Intelligence – Week 3 (Homework Assignment)

Instructor: Teerapong Panboonyuen

Course Repository:
https://github.com/kaopanboonyuen/SC310005_ArtificialIntelligence_2025s1

🎯 Objective

This week, you will practice supervised machine learning techniques using the provided cancer_med_dataset. Your task is to improve upon the models developed in class by exploring at least **five different machine learning algorithms or techniques**, fine-tuning hyperparameters, or testing new models.

Optionally, you may experiment with modern AI approaches like Convolutional Neural Networks (CNNs) or other deep learning models, but your final models must **outperform the accuracy baseline** achieved in class.

📚 Dataset Description

The dataset contains patient medical data related to cancer diagnosis with features such as:

- Age
- Gender
- BMI
- Smoking habits
- Genetic Risk
- Physical Activity
- Alcohol Intake
- Cancer History
- Diagnosis (target variable)
- Symptom (categorical textual feature)

📝 Assignment Instructions

You must:

- Load the dataset (`cancer_med_dataset.csv`) into a pandas DataFrame.
- Experiment with at least **five different supervised learning techniques** (e.g., Decision Tree, Random Forest, Logistic Regression, Support Vector Machines, k-NN, Gradient Boosting, Neural Networks).

- Fine-tune hyperparameters for each model to improve performance (e.g., tree depth, number of estimators, regularization strength, learning rate).
- Optionally, try modern AI models, such as CNNs or other deep architectures, but ensure they outperform the class baseline.
- Convert categorical features properly (e.g., one-hot encoding for symptoms and gender as taught).
- Split your data into train/test sets with a fixed random seed — **use `random_state=42` everywhere** to ensure reproducibility.
- Document your model training, evaluation metrics (accuracy, precision, recall, F1), and insights.
- Present a summary table comparing all models.
- Visualize results using confusion matrices, ROC curves, and AUC scores.
- Provide clear explanations of how to interpret these metrics and visualizations.

🔥 Examples of ML Techniques You May Explore

- Decision Tree Classifier
- Random Forest Classifier
- Logistic Regression
- Support Vector Machine (SVM)
- k-Nearest Neighbors (k-NN)

- Gradient Boosting Machines (XGBoost, LightGBM)
- Multi-layer Perceptron (Neural Networks)
- Convolutional Neural Networks (CNN) — optional and advanced

---

💡 Deliverables

Submit a Jupyter notebook or Google Colab file that includes:

- Code implementing the ML models and training process.
- Comments explaining your approach and reasoning.
- Clear outputs displaying evaluation metrics and comparisons.
- Visualizations (confusion matrices, ROC curves, etc.).
- A summary section with a results comparison table.
- (Optional) A section detailing any advanced techniques or deep learning models used.

🎯 Evaluation Criteria

- Correctness and completeness of your ML pipeline.
- Effectiveness in improving accuracy beyond the class baseline.
- Proper data preprocessing and encoding techniques.
- Use of `random_state=42` or equivalent seed to ensure reproducibility.

- Quality of explanations and code readability.
- Creativity and use of advanced techniques for bonus credit.
- Clear and insightful visualizations of results.

🚀 Getting Started: Dataset Access

You can download the dataset here:

`https://github.com/kaopanboonyuen/SC310005_ArtificialIntelligence_2025s1/raw/main/dataset/cancer_med_dataset.csv`
(Provided via class repository or link)

📅 Submission Deadline

The final submission date will be announced and agreed upon during class. Please ensure you note this date and plan your work accordingly.