A.I. Homework- 4

Classical Machine Learning vs In-Context Learning with GPT on UCI Datasets

Naive Bayes vs Decision Tree vs MLP vs GPT (In-Context Learning) on three UCI datasets – Iris, Breast Cancer, Adult Income

Nishant Kumar, 2022326

Assignment Objectives & Tasks

- **Goal:** Empirically compare traditional supervised models with an LLM used only via **In-Context Learning (ICL)**.

- **Datasets:** 3 UCI datasets with different sizes and difficulty:
  - Small: Iris (150 samples, 3-class classification)
  - Medium: Breast Cancer Wisconsin (569 samples, binary classification)
  - Large: Adult Income (48k+ samples, binary classification, many categorical features)

- **Core tasks:**
  - Data preprocessing: handle missing values, encode categorical features, scale numeric features where needed.
  - Train + tune **Naive Bayes, Decision Tree, MLP** using **k-fold CV (k ≥ 5)**.
  - Design **few-shot prompts** for an LLM (GPT) and evaluate it on the **same test splits**.
  - Compare models using **accuracy, F1-score, confusion matrices**, plus qualitative aspects (interpretability, training cost, robustness, etc.).

Dataset Overview and Preprocessing

Dataset	Size	I/P type	Target
Iris	150 rows, 4 numeric features	Flow Measurements	Species (3 classes)
Breast Cancer	569 rows, 30 numeric features	Cell nucleus measurements	Benign
Adult income	48,842 rows, mix of num + cat	Demographic & work attributes	Income ≤50K vs >50K

Common steps

- Train–test split using fixed random seed for reproducibility.
- Separate features and targets, keep the same test split for all models including GPT.

Numeric features

- For Iris & Breast Cancer: standardisation using `StandardScaler`.

Categorical features (Adult)

- Handle `"?"` as missing values and drop such rows.
- One-hot encode categorical columns (e.g., workclass, education, marital-status, occupation).

Pipelines

- Used `sklearn` Pipelines so that **preprocessing is inside CV**, preventing data leakage.

Models & Experimental Setup

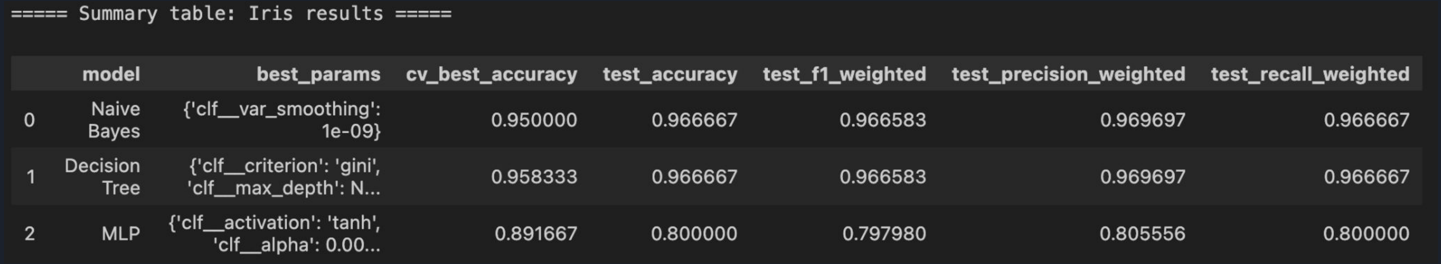
- **Classical models (scikit-learn):**
  - **Naive Bayes (GaussianNB)** – simple generative baseline, assumes feature independence.
  - **Decision Tree** – axis-aligned splits, good interpretability.
  - **MLPClassifier (Neural Network)** – one hidden layer (I tuned size), non-linear decision boundaries.

- **Hyperparameter tuning (all datasets):**
  - Used **GridSearchCV** with **k-fold cross-validation**:
    - Iris & Breast Cancer: k = 5
    - Adult: k = 5 (on large dataset; more folds would be expensive)

  - Searched over:
    - Decision Tree: `criterion`, `max_depth`, `min_samples_split`, `min_samples_leaf`
    - MLP: hidden layer size, activation (`tanh`/`relu`), learning rate, `alpha`
    - NB: `var_smoothing` for GaussianNB

- **Metrics on test split:**
  - **Accuracy**, **weighted F1-score**
  - **Confusion matrix** to see types of errors.

Iris Results (Small, Clean Dataset)



```
===== Summary table: Iris results =====
```

	model	best_params	cv_best_accuracy	test_accuracy	test_f1_weighted	test_precision_weighted	test_recall_weighted
0	Naive Bayes	{'clf__var_smoothing': 1e-09}	0.950000	0.966667	0.966583	0.969697	0.966667
1	Decision Tree	{'clf__criterion': 'gini', 'clf__max_depth': N...	0.958333	0.966667	0.966583	0.969697	0.966667
2	MLP	{'clf__activation': 'tanh', 'clf__alpha': 0.00...	0.891667	0.800000	0.797980	0.805556	0.800000

Best CV accuracies:

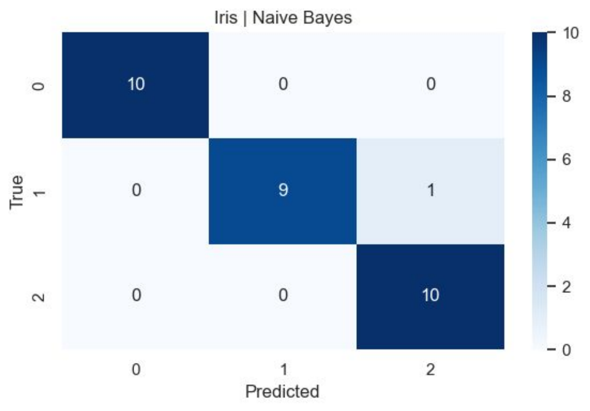
- Naive Bayes: **0.95**
- Decision Tree: **0.9583**
- MLP: **0.8917**

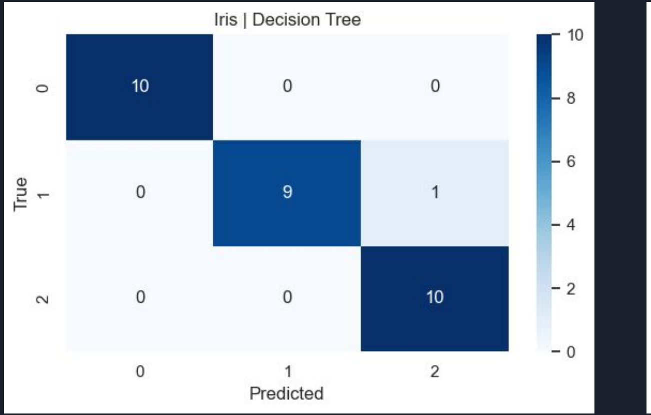
Test performance (30-sample test split):

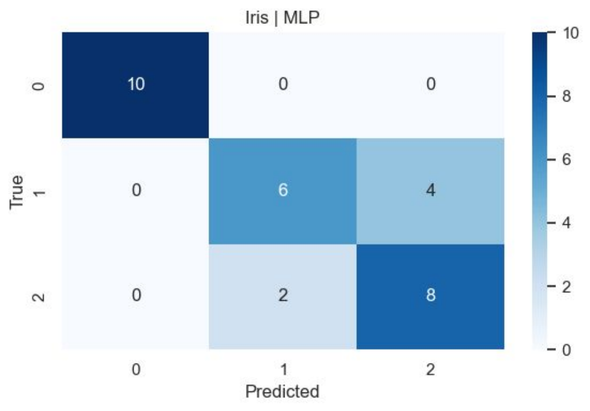
- Naive Bayes: accuracy **0.9667**, F1-weighted **0.9666**
- Decision Tree: accuracy **0.9667**, F1-weighted **0.9666**
- MLP: accuracy **0.80**, F1-weighted **0.798**

Observation: On this small, clean dataset, **simple models (NB/DT)** already achieve near-perfect performance; the tuned MLP actually underperforms them.

Iris Dataset – Confusion Matrices (Naive Bayes, Decision Tree, MLP)







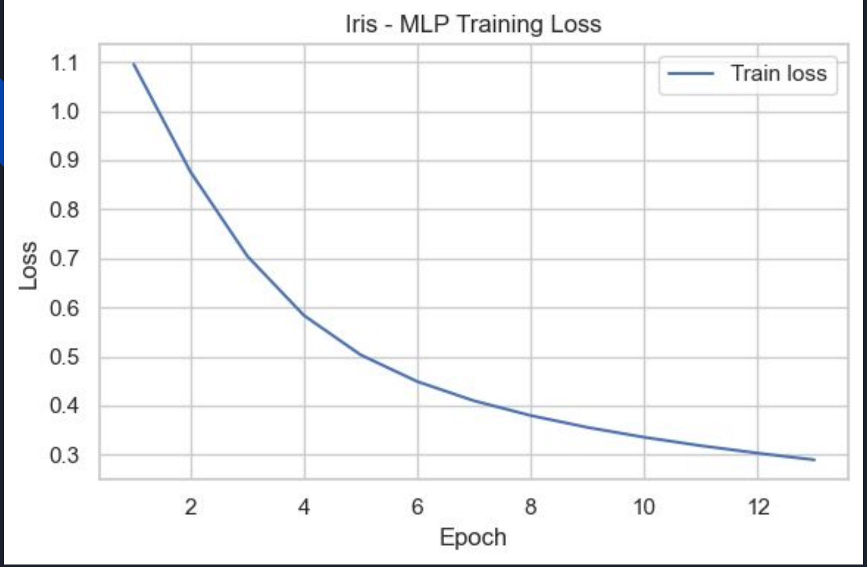
Naive Bayes & Decision Tree

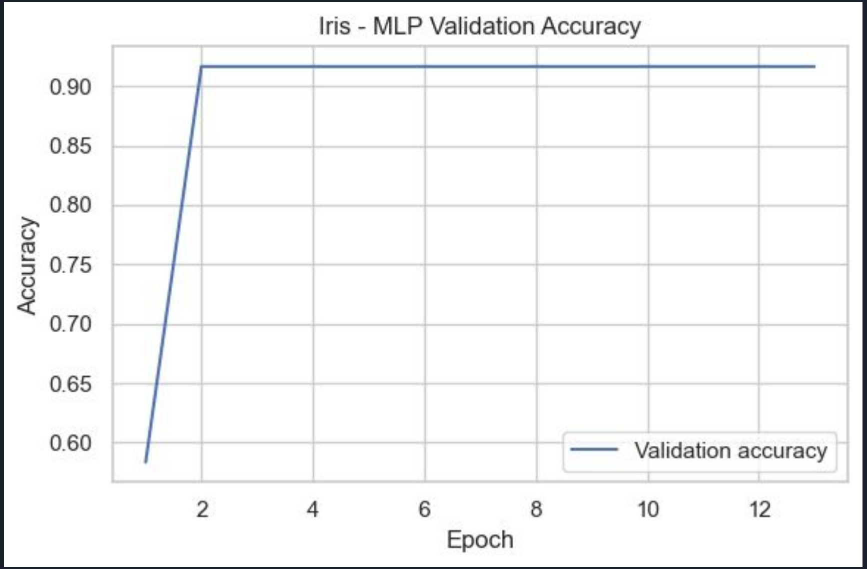
- Both achieve **96.7% test accuracy** and **F1 ≈ 0.97**.
- Only **1 sample of class 1** is misclassified as class 2; the other 29/30 points are correct.

MLP (Neural Network)

- Test accuracy drops to **80% (F1 ≈ 0.80)**.
- Misclassifies part of class 1 and class 2 (see off-diagonal counts).

On this small, clean and almost linearly separable dataset, **simple models (NB/DT)** already perform almost perfectly, while the more complex MLP **does not add benefit and even underperforms**.



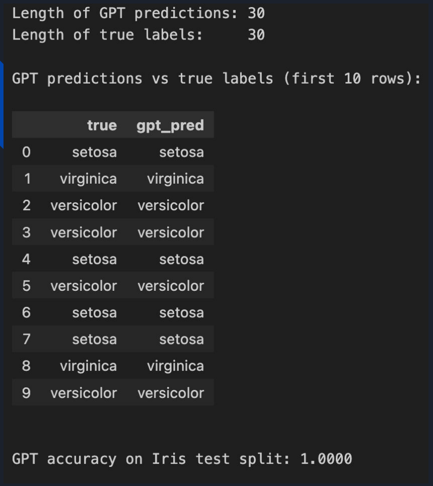


Training loss steadily decreases from ≈1.1 to ≈0.29 over 13 epochs → the network is successfully minimising its objective.

Validation accuracy jumps quickly from ≈0.58 to ≈0.92 by epoch 2 and then stays flat → the model learns a good decision boundary very early and then stabilises.

No clear sign of overfitting on the validation curve, but the **final test accuracy is only 80%**, lower than Naive Bayes / Decision Tree (≈96.7%).

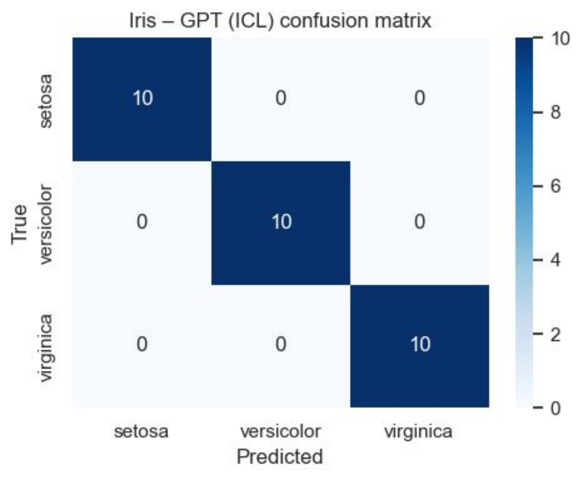
This suggests that on such a **small, low-dimensional dataset**, a simple MLP with these hyper-parameters does **not generalise as well as the simpler models**, and performance is quite sensitive to the exact train/validation/test split.

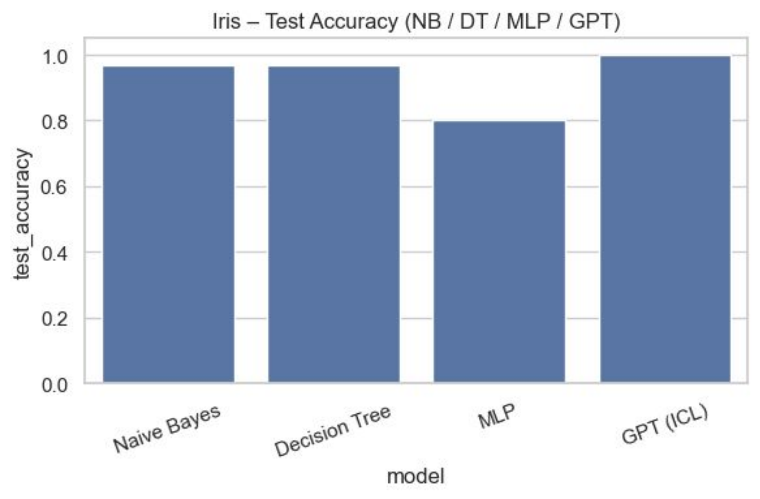


```
Length of GPT predictions: 30
Length of true labels:     30

GPT predictions vs true labels (first 10 rows):
```

	true	gpt_pred
0	setosa	setosa
1	virginica	virginica
2	versicolor	versicolor
3	versicolor	versicolor
4	setosa	setosa
5	versicolor	versicolor
6	setosa	setosa
7	setosa	setosa
8	virginica	virginica
9	versicolor	versicolor

```
GPT accuracy on Iris test split: 1.0000
```





GPT with in-context learning (ICL) reaches **100% accuracy** on the Iris test split, slightly outperforming the classical models.

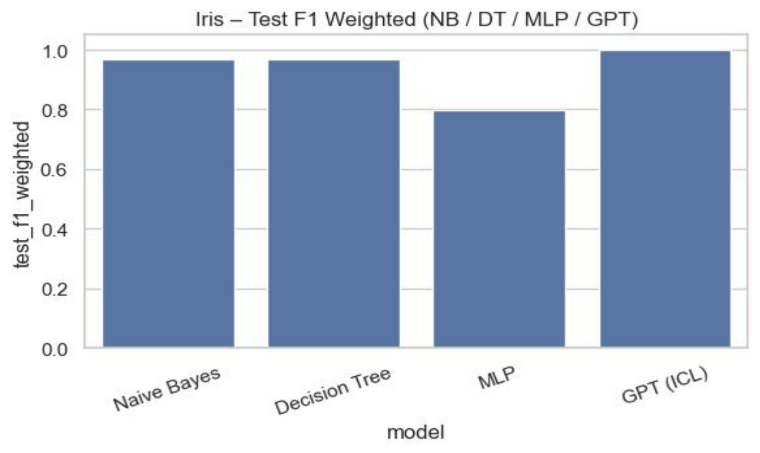
Prompt setup: GPT was given a few labeled Iris examples (feature → species) and asked to predict labels for **30 unseen test samples** using only in-context learning (no parameter training).

Perfect predictions on Iris: Confusion matrix is purely diagonal – GPT correctly classifies **all 30/30 test points (accuracy = 1.00, F1 = 1.00)**.

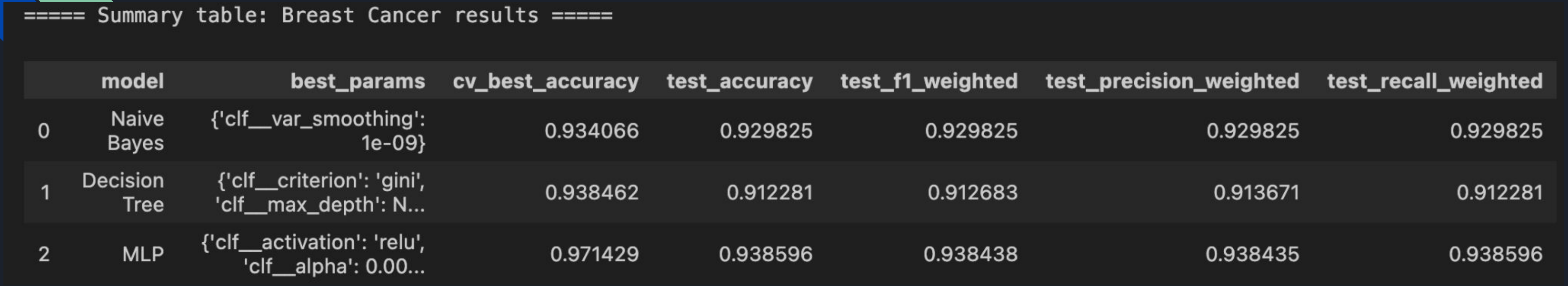
Comparison with trained models:

- Naive Bayes / Decision Tree: test accuracy ≈ **0.97**, F1 ≈ **0.97**
- MLP: test accuracy ≈ **0.80**, F1 ≈ **0.80**
- GPT (ICL): **outperforms all three** on this small, clean dataset.

For low-dimensional, well-separated data like Iris, a large LLM can **quickly infer the decision rule from a handful of examples**, matching or beating fully trained ML models on the test split.



Breast Cancer Results (Medium Complexity)



```
===== Summary table: Breast Cancer results =====
```

	model	best_params	cv_best_accuracy	test_accuracy	test_f1_weighted	test_precision_weighted	test_recall_weighted
0	Naive Bayes	{'clf__var_smoothing': 1e-09}	0.934066	0.929825	0.929825	0.929825	0.929825
1	Decision Tree	{'clf__criterion': 'gini', 'clf__max_depth': N...	0.938462	0.912281	0.912683	0.913671	0.912281
2	MLP	{'clf__activation': 'relu', 'clf__alpha': 0.00...	0.971429	0.938596	0.938438	0.938435	0.938596

CV best accuracies:

- Naive Bayes: **0.9341**
- Decision Tree: **0.9385**
- MLP: **0.9714**

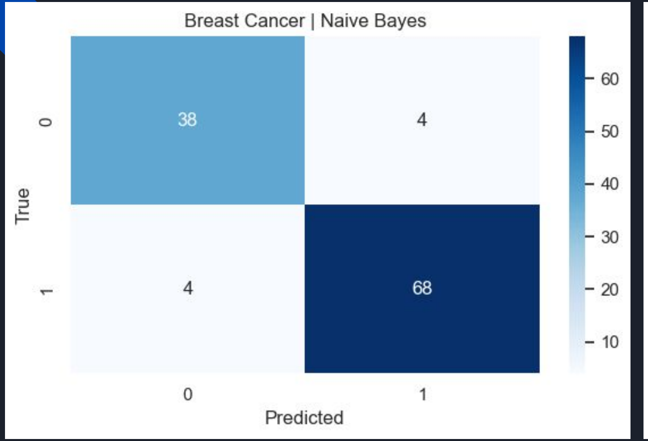
Test performance (114-sample test split):

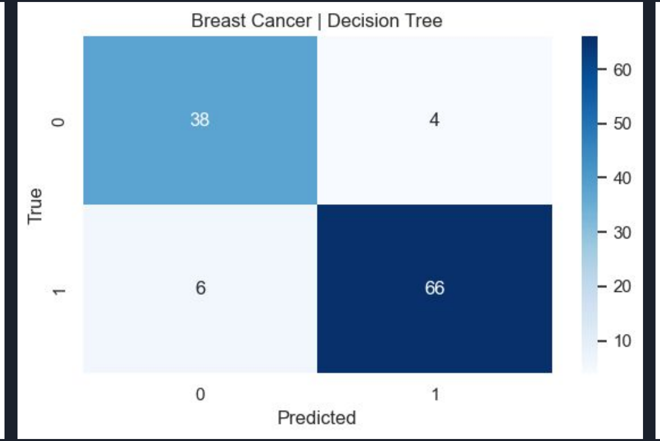
- Naive Bayes: accuracy **0.9298**, F1-weighted **0.9298**
- Decision Tree: accuracy **0.9123**, F1-weighted **0.9127**
- MLP: accuracy **0.9386**, F1-weighted **0.9384**

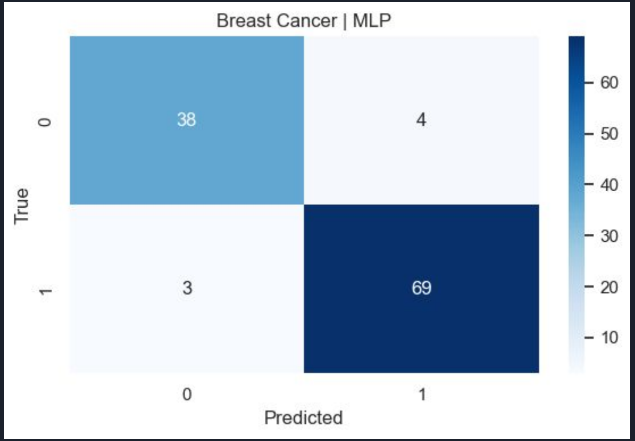
Error analysis:

- All three models make **few mistakes on malignant cases**, which is important clinically.
- MLP gives the **best overall F1**, at the cost of being less interpretable and more expensive to train.

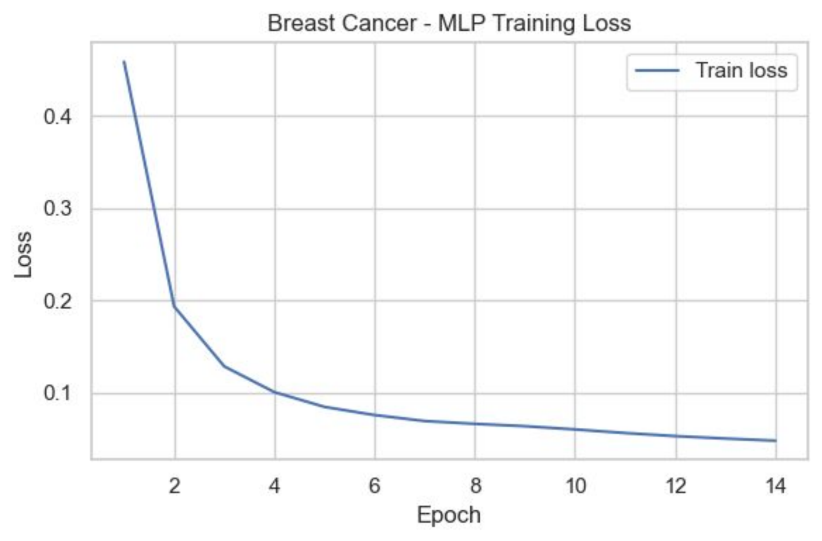
Breast Cancer Dataset – Confusion Matrices (Naive Bayes, Decision Tree, MLP)

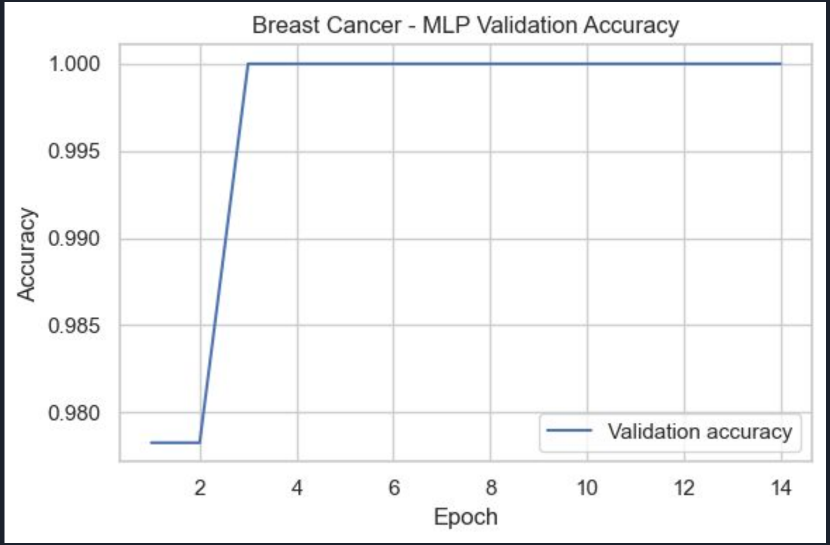






On the Breast Cancer Wisconsin dataset (42 benign, 72 malignant in the test set), all three models perform well, but the MLP is clearly best. Naive Bayes and the Decision Tree both correctly classify 38 benign cases, with overall accuracies of ~0.93 and ~0.91 respectively, but the tree has slightly more malignant cases misclassified as benign. The MLP achieves the highest test accuracy (~0.94) and the fewest false negatives (only 3 malignant predicted as benign), making it the most reliable and clinically safest model.





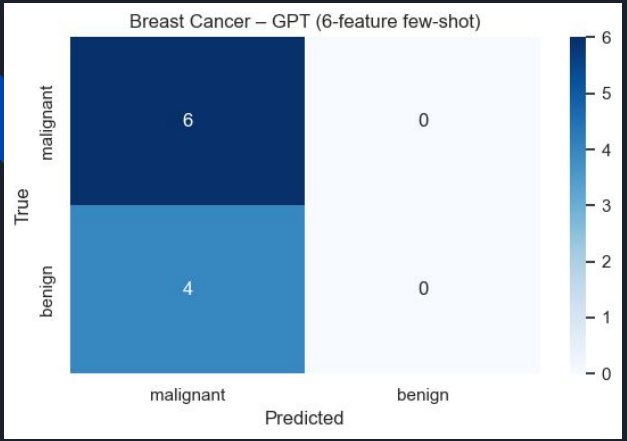
Model setup: MLP with one hidden layer of 50 neurons (ReLU) and learning rate 0.01, trained for 14 epochs.

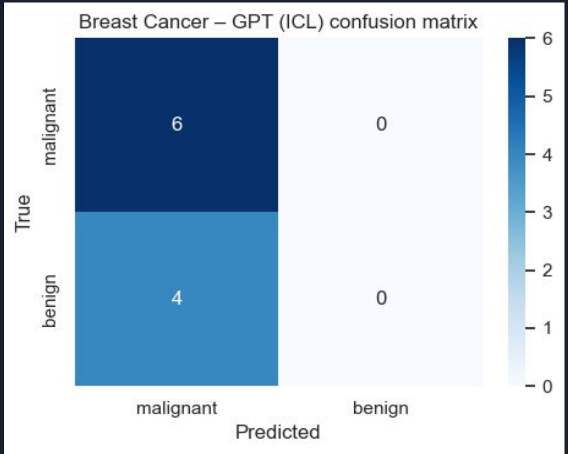
Training loss: Drops smoothly from **≈0.45 to <0.05**, showing stable optimisation without divergence.

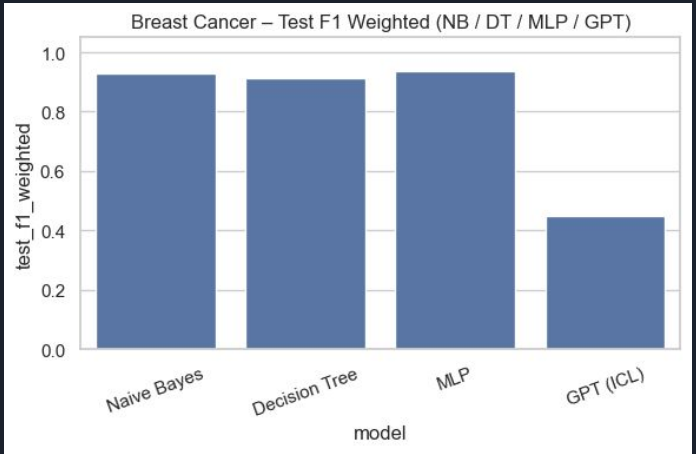
Validation accuracy: Reaches **≈0.98 by epoch 2** and becomes **~1.0 from epoch 3 onward**, staying flat afterwards.

Interpretation: The network learns a good decision boundary very quickly and does **not show obvious overfitting** within 14 epochs (no drop in validation accuracy while loss keeps decreasing).

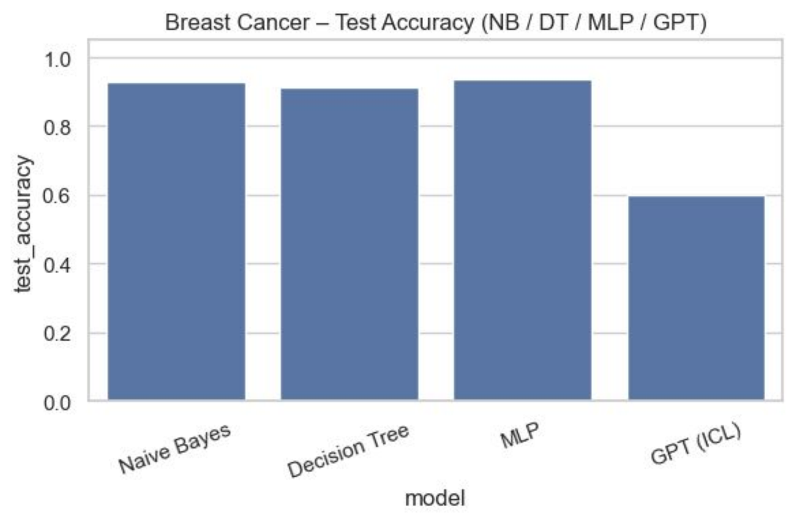
Link to metrics: This behaviour is consistent with the final test scores (accuracy ≈ **0.94**, weighted F1 ≈ **0.94**), indicating strong generalisation on this dataset.



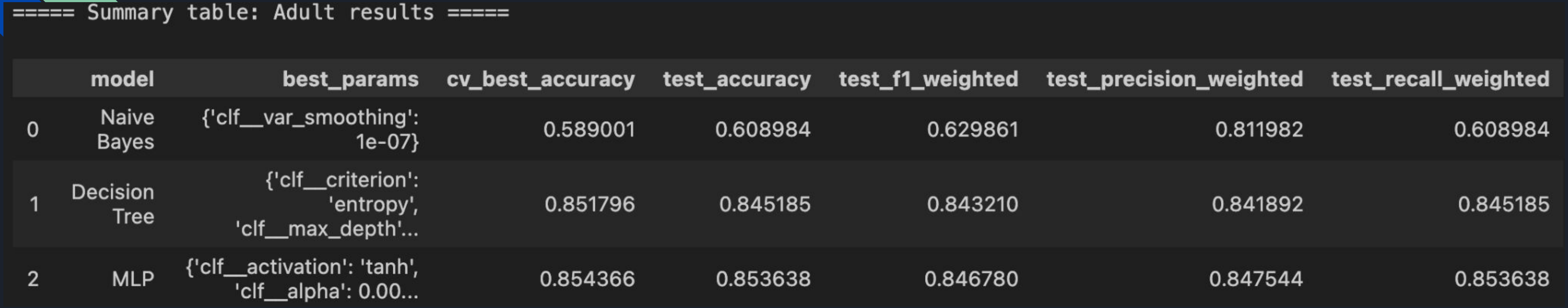




- **ICL setup:** GPT saw 6 few-shot examples with only 6 input features and had to predict labels for 30 Breast Cancer test cases.
- **Prediction pattern:** GPT classifies *every* case as malignant → catches all 6 malignant examples but mislabels 4 benign as malignant.
- **Performance vs ML models:** Accuracy ≈ **60%** and much lower F1 than Naive Bayes / Decision Tree / MLP (all ≈ **0.93–0.94**).
- **Cinclusion:** With limited features + small context, GPT behaves like a cautious but poorly calibrated classifier—good at avoiding missed cancers, but far worse overall than trained ML models.



Adult Income Results (Large, Mixed-Type Dataset)



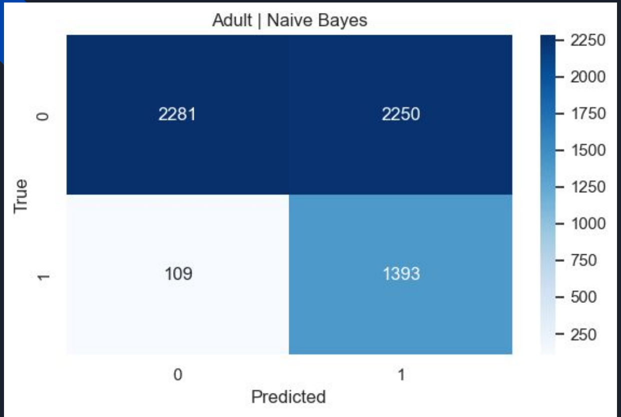
```
===== Summary table: Adult results =====
```

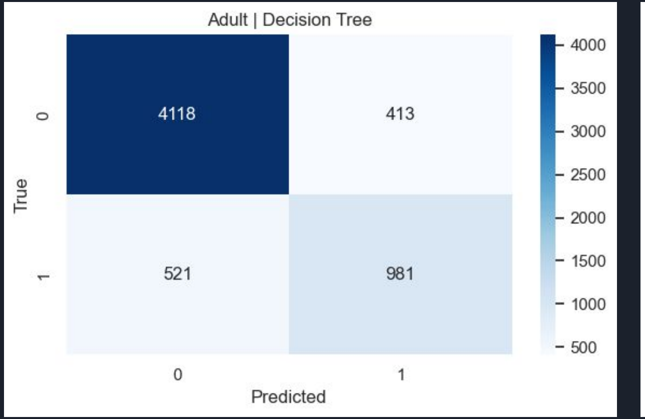
	model	best_params	cv_best_accuracy	test_accuracy	test_f1_weighted	test_precision_weighted	test_recall_weighted
0	Naive Bayes	{'clf__var_smoothing': 1e-07}	0.589001	0.608984	0.629861	0.811982	0.608984
1	Decision Tree	{'clf__criterion': 'entropy', 'clf__max_depth'...	0.851796	0.845185	0.843210	0.841892	0.845185
2	MLP	{'clf__activation': 'tanh', 'clf__alpha': 0.00...	0.854366	0.853638	0.846780	0.847544	0.853638

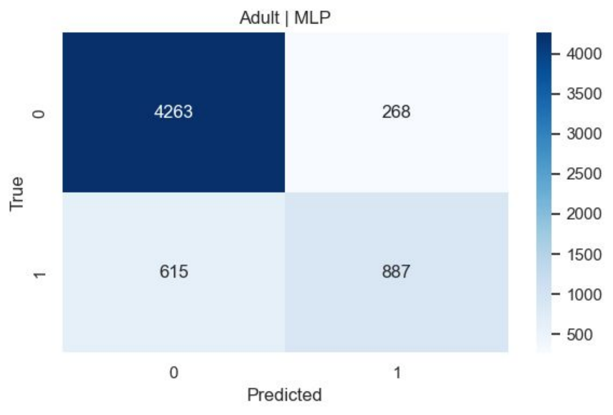
- **Best CV accuracies:**
  - Naive Bayes: **0.5890**
  - Decision Tree: **0.8518**
  - MLP: **0.8544**

- **Test performance (6033 test examples):**
  - Naive Bayes: acc **0.6090**, F1-weighted **0.6299**
  - Decision Tree: acc **0.8452**, F1-weighted **0.8432**
  - MLP: acc **0.8536**, F1-weighted **0.8468**

Adult– Confusion Matrices (Naive Bayes, Decision Tree, MLP)





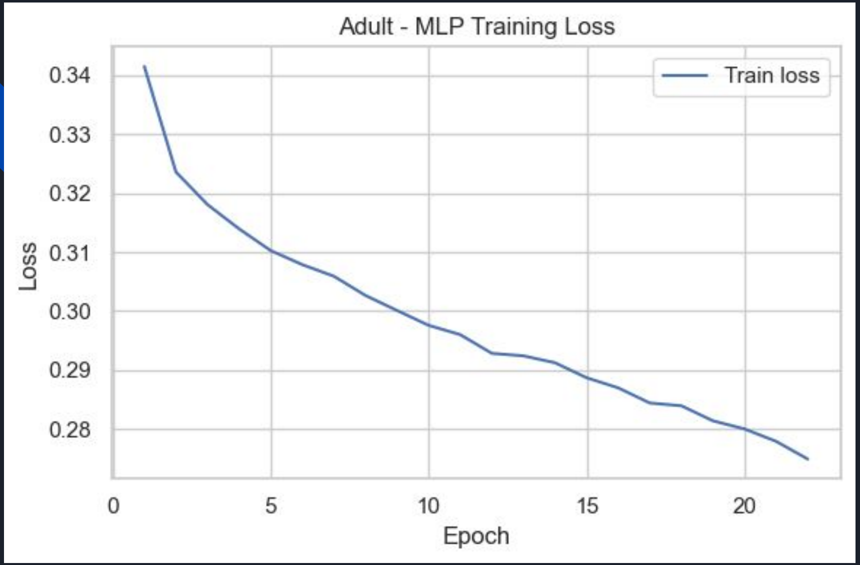


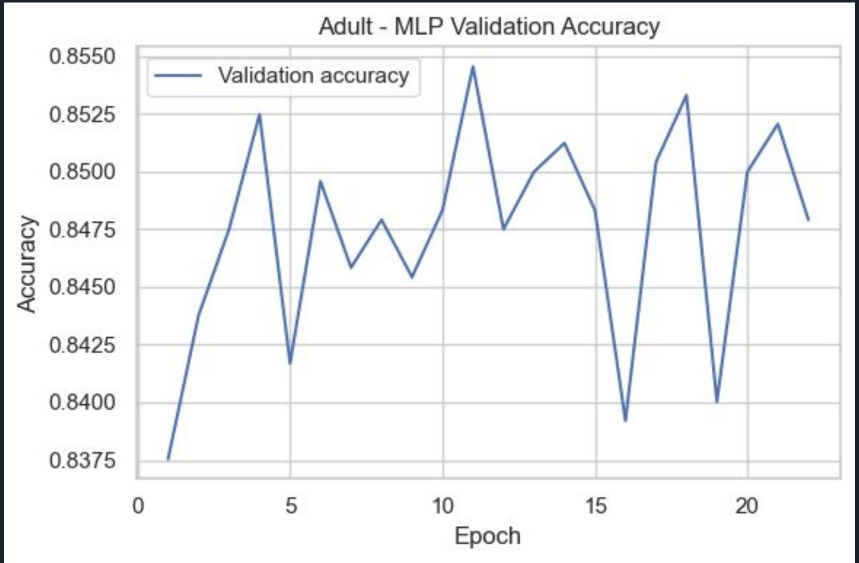
Setup: Binary income prediction on Adult dataset – class 0 = <=50K, class 1 = >50K.

Naive Bayes (Acc ≈ 0.61): Very unbalanced – many <=50K people (2250) are wrongly predicted as >50K, even though it catches most high-income cases.

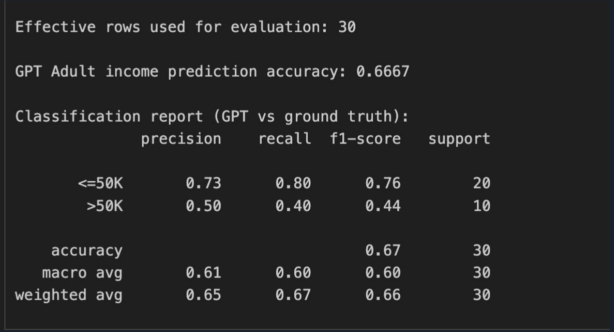
Decision Tree (Acc ≈ 0.85): Greatly reduces both false positives and false negatives (413 + 521), giving much more balanced behaviour across the two classes.

MLP (Acc ≈ 0.85, best F1): Fewest errors on the majority class and competitive recall on >50K; overall the strongest supervised model on Adult.

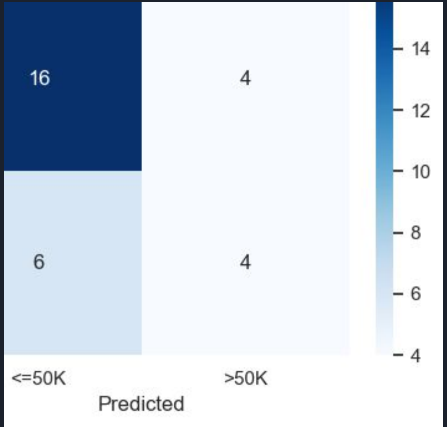


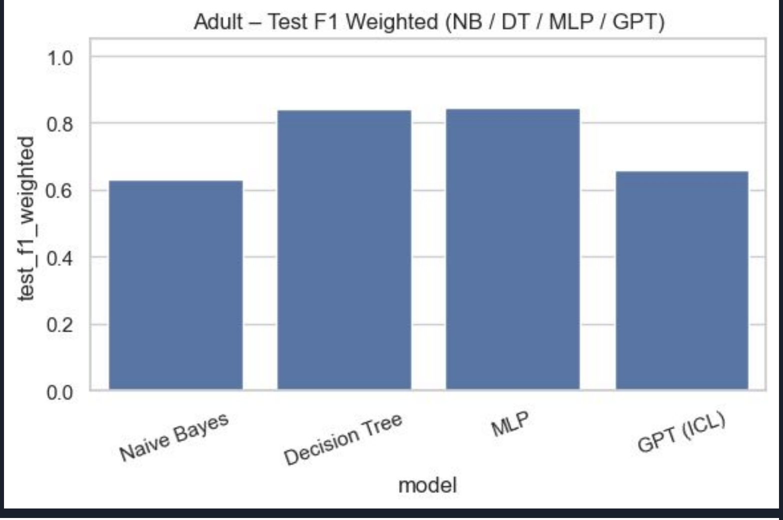


- Training loss decreases steadily from ~0.34 to ~0.28 over 22 epochs, showing the MLP is consistently learning a better decision boundary.
- Validation accuracy stays in a narrow band around **84–85.5%**, with small fluctuations due to mini-batch / stochastic optimisation.
- No strong overfitting trend: validation accuracy does not collapse even when training loss keeps going down.
- Overall, the chosen number of epochs gives a good trade-off between convergence and stability on the Adult dataset.

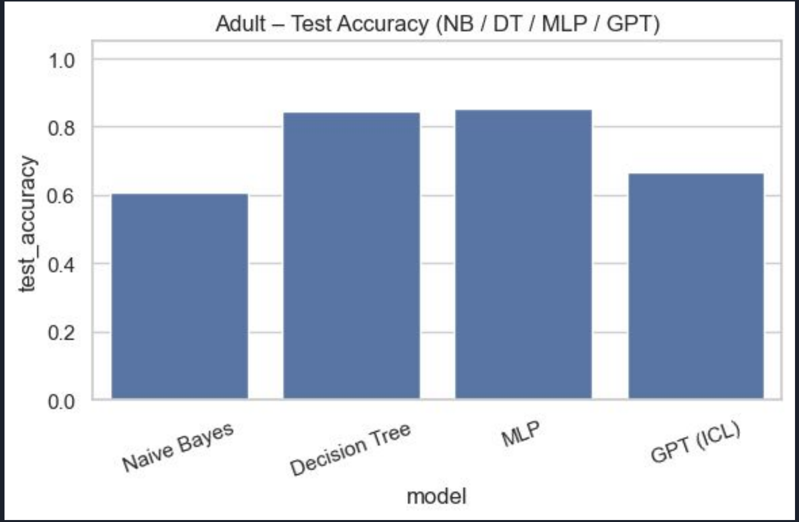


```
Effective rows used for evaluation: 30

GPT Adult income prediction accuracy: 0.6667

Classification report (GPT vs ground truth):
              precision    recall  f1-score   support

       <=50K       0.73      0.80      0.76        20
        >50K       0.50      0.40      0.44        10

    accuracy                           0.67        30
   macro avg       0.61      0.60      0.60        30
weighted avg       0.65      0.67      0.66        30
```

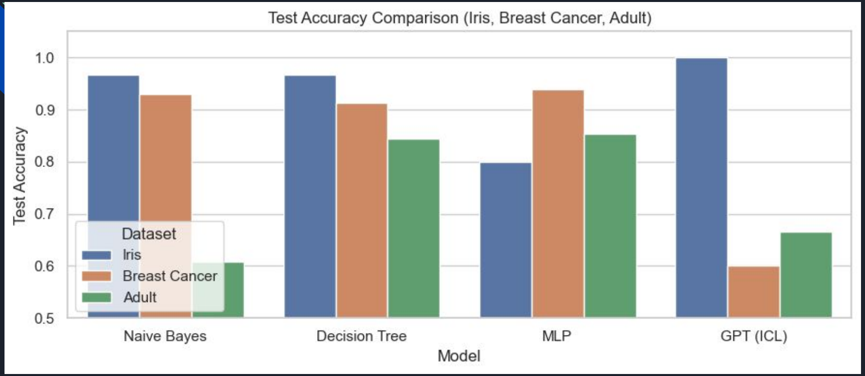


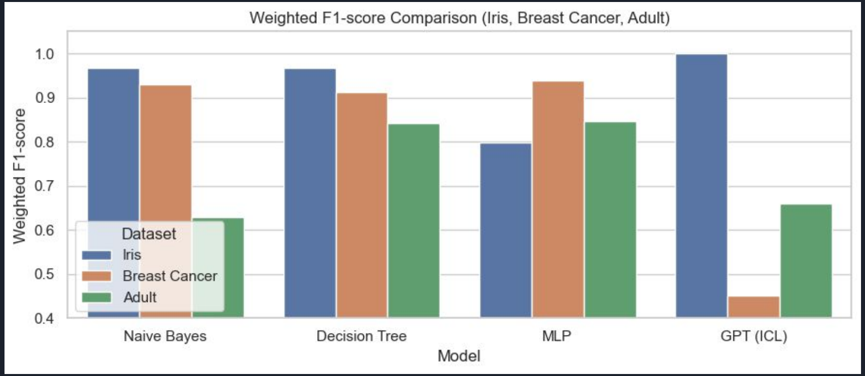


- Evaluated GPT with in-context learning on **30 Adult test examples** using only 6 key features and few-shot prompts.
- GPT reaches **~66.7% accuracy** and **F1 ≈ 0.66** – it correctly classifies 16/20 <=50K and 4/10 >50K cases.
- Confusion matrix shows a **bias toward the majority class (<=50K)**: 6 out of 10 high-income (>50K) samples are misclassified.
- Tuned Decision Tree and MLP (~0.85 accuracy, F1 ≈ 0.84–0.85) clearly outperform GPT, highlighting that **on large, imbalanced tabular data, supervised ML still beats ICL with an LLM**.



Cross-Dataset Comparison & Discussion





- **Effect of dataset size & complexity**
  - On **small, clean Iris**, even very simple generative/discriminative models are almost perfect; GPT also does very well with few-shot ICL.
  - On **Breast Cancer**, the tuned MLP gives the best F1; GPT struggles because it only sees 6 features and because of exact-match issues.
  - On **large, heterogeneous Adult**, tree and MLP clearly dominate; GPT's text-only reasoning cannot fully capture the tabular structure.
- **Interpretability:**
  - Naive Bayes & Decision Tree are **interpretable** (feature importance, decision paths).
  - MLP and GPT behave like **black boxes**.
- **Training and compute cost:**
  - Classical models are **cheap to train** locally.
  - GPT requires **no training**, but **inference is expensive** and limited by context length.
- **Robustness & generalisation:**
  - MLP + Decision Tree generalise well on large data when tuned with cross-validation.
  - GPT is sensitive to **prompt wording, feature selection, and formatting**.

Conclusions and Takeaways

Classical ML vs GPT (ICL):

- On structured tabular datasets, **properly tuned classical models** (NB/DT/MLP) generally **outperform GPT ICL**, especially on medium/large datasets.
- GPT can match or exceed performance on **very small, low-dimensional tasks** like Iris when the mapping is simple.

Role of In-Context Learning:

- ICL is powerful for **rapid prototyping** and tasks where labels can be described semantically, but it is not a drop-in replacement for supervised training on tabular data.
- Prompt design, feature selection, and consistent formatting are critical.

Model choice guidelines (from this study):

- For **small, clean** datasets → NB / DT already very strong.
- For **medium-sized medical data** → MLP gives best F1 but DT is a strong, interpretable baseline.
- For **large, mixed-type** datasets → tree-based or neural models tuned with CV are preferred.