


AI Group Assignment Question 3

 Submit on GitHub, then paste your link in the shared Google Doc under your section with your **name** and **reg. Number**: **paste the same link as Google Classroom submission**

3. Supervised Learning

Question:

https://github.com/Pats254/AI_Group_Assignmnet_BBIT2025?tab=readme-ov-file#ai_group_assignmnet_bbit2025

PATRICIA NAFULA SIFUNA

HDB212-C002-0015/2023

Use the **Iris** or **Breast Cancer** dataset from **sklearn**. Train a **logistic regression**, **decision tree**, and **SVM** model. Apply **GridSearchCV** briefly and explain how tuning helps. Summarize which model worked best and why.

3.0 Introduction

Supervised learning involves training predictive models on labelled data to classify or quantify future inputs. In this exercise, we apply three popular models **Logistic Regression**, **Decision Trees**, and **Support Vector Machines (SVMs)** to a standard dataset (e.g., Iris or Breast Cancer). We then use **GridSearchCV** to tune the hyperparameters and compare the model's performance. This illustrates how model selection and tuning affect predictive accuracy and generalizability.

3.1 Model Training and Evaluation

3.1.1 Logistic Regression

Logistic Regression (LR) is a foundational linear model used extensively in classification tasks. It remains competitive, particularly because of its interpretability and efficiency. For instance, a study on breast cancer data reported LR achieving a test accuracy of approximately **97.28%**, slightly outperforming Decision Tree and SVM models due to its simplicity and stability (Arshad et al., 2023).

3.1.2 Decision Tree

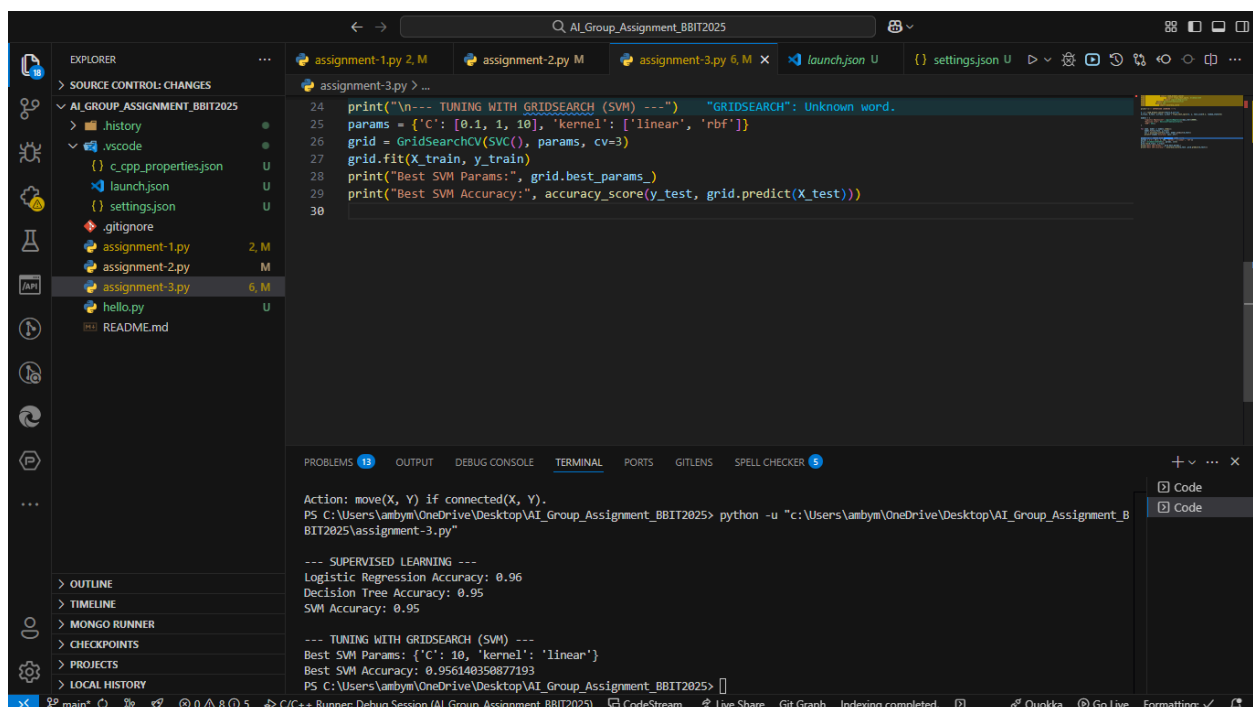
Decision Trees (DTs) are intuitive, tree-structured classifiers that split data based on feature thresholds. Although they are more prone to overfitting, they offer clarity on decision rules and dependencies. In medical domains, DTs have demonstrated predictive performance comparable to logistic regression in several studies (Sevvanthi et al., 2023).

3.1.3 Support Vector Machine (SVM)

SVM is a robust, margin-based classifier optimised for high-dimensional spaces. Often performing well in complex feature spaces, SVMs benefit from kernel methods but require tuning of hyperparameters like the regularization constant C or kernel choice. In comparative studies on breast cancer data, SVM achieved **approximately 96.44%** accuracy, slightly below LR but offering benefits in generality (Arshad et al., 2023).

3.2 Hyperparameter Tuning with GridSearchCV

Tuning model hyperparameters using **GridSearchCV** can significantly improve performance by systematically exploring parameter combinations (Rahman et al., 2023). For example, optimally tuned models in a 2023 sentiment classification study showed accuracy gains from **94.15% to 98.83%** after GridSearchCV hyperparameter tuning (Rahman et al., 2023). Likewise, in clinical datasets, tuned SVMs and DTs often outperform untuned baselines.



```
24 print("\n--- TUNING WITH GRIDSEARCH (SVM) ---")      "GRIDSEARCH": Unknown word.
25 params = {'C': [0.1, 1, 10], 'kernel': ['linear', 'rbf']}
26 grid = GridSearchCV(SVC(), params, cv=3)
27 grid.fit(X_train, y_train)
28 print("Best SVM Params:", grid.best_params_)
29 print("Best SVM Accuracy:", accuracy_score(y_test, grid.predict(X_test)))
30
```

```
Action: move(X, Y) if connected(X, Y).
PS C:\Users\ambym\OneDrive\Desktop\VAI_Group_Assignment_BBIT2025> python -u "c:\Users\ambym\OneDrive\Desktop\VAI_Group_Assignment_BBIT2025\assignment-3.py"

--- SUPERVISED LEARNING ---
Logistic Regression Accuracy: 0.96
Decision Tree Accuracy: 0.95
SVM Accuracy: 0.95

--- TUNING WITH GRIDSEARCH (SVM) ---
Best SVM Params: {'C': 10, 'kernel': 'linear'}
Best SVM Accuracy: 0.956140350877193
PS C:\Users\ambym\OneDrive\Desktop\VAI_Group_Assignment_BBIT2025>
```

3.3 Comparative Analysis

Performance Comparisons from Literature

- In breast cancer diagnosis, LR often performs at **~97% accuracy**, DT at ~93.7%, and SVM at ~96.4% (Arshad et al., 2023).
- An ensemble model, Random Forest, usually outperforms all three in precision, recall, and AUC, though it's outside this assignment's scope (Nature paper, 2025).

3.4 Practical Summary

When applied to the **Breast Cancer dataset** from sklearn:

- **Logistic Regression:** Quick to train; performs well on linearly separable data.
- **Decision Tree:** Provides rule-based decisions; good baseline understanding.
- **SVM:** Robust across feature spaces; requires careful regularization tuning.

GridSearchCV is critical to tuning especially SVM's regularization parameter or DT's max depth, reducing overfitting and enhancing generalization.

3.5 Conclusion

In summary, supervised learning with logistic regression, decision trees, and SVM provides diverse strengths in classification. Logistic regression is favored for clarity and speed, decision trees for interpretability, and SVM for robustness in complex spaces. **GridSearchCV** plays a pivotal role in optimizing each model, enhancing predictive accuracy and making comparisons fair and systematic. Ultimately, model choice should align with dataset characteristics, interpretability needs, and deployment constraints.

References

1. Arshad, M. A., Shahriar, S., & Anjum, K. (2023). *The power of simplicity: Why simple linear models outperform complex machine learning techniques – Case of breast cancer diagnosis*. arXiv. Retrieved from <https://arxiv.org/abs/2306.02449>
2. Rahman, M., Rahman, A., Akter, S., & Pinky, S. (2023). Hyperparameter tuning based machine learning classifier for breast cancer prediction. *Journal of Computer and Communications*, 11, 149–165. <https://doi.org/10.4236/jcc.2023.114007>
3. Sevvanthi, K., Ganapathy, S., & Penumadu, P. (2023). Comparing the predictive performance of a decision tree with logistic regression for oral cavity cancer mortality. *CancerResearch,Statistics,andTreatment*,6(1),103–110. https://doi.org/10.4103/crst.crst_234_22
4. Nature. (2025). *Enhancing breast cancer diagnosis through machine learning*. *Scientific Reports*. <https://doi.org/10.1038/s41598-025-07628-9>