
Feasibility Analysis

Decision tree

Decision tree is a classification and regression algorithm based on tree structure for decision analysis(Loh, 2011). It gradually divides the data into smaller subsets according to the partitioning of features, and finally forms a tree that is used to predict the target variable.

Loh, W.-Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1), pp.14–23. doi:https://doi.org/10.1002/widm.8.

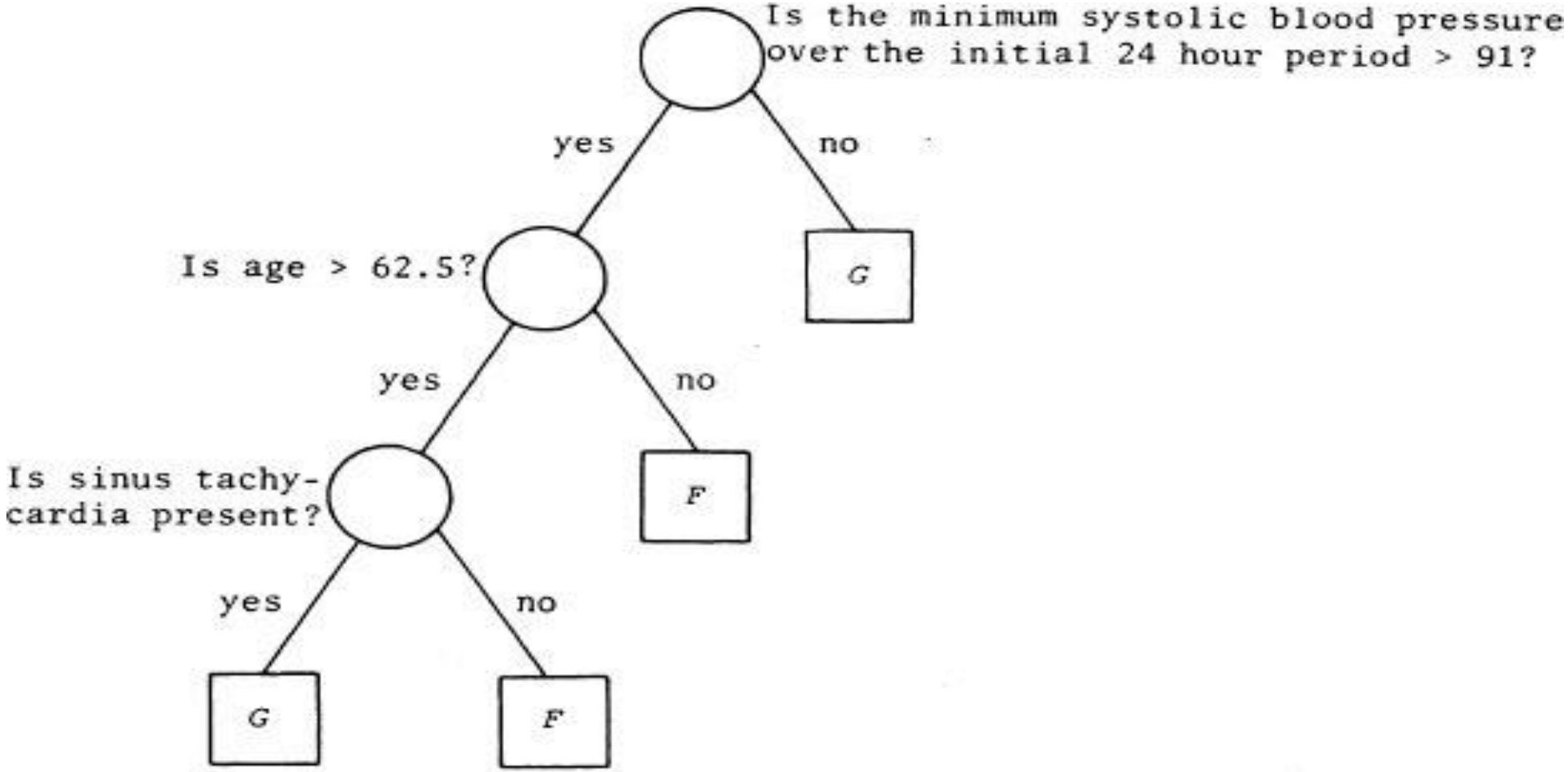
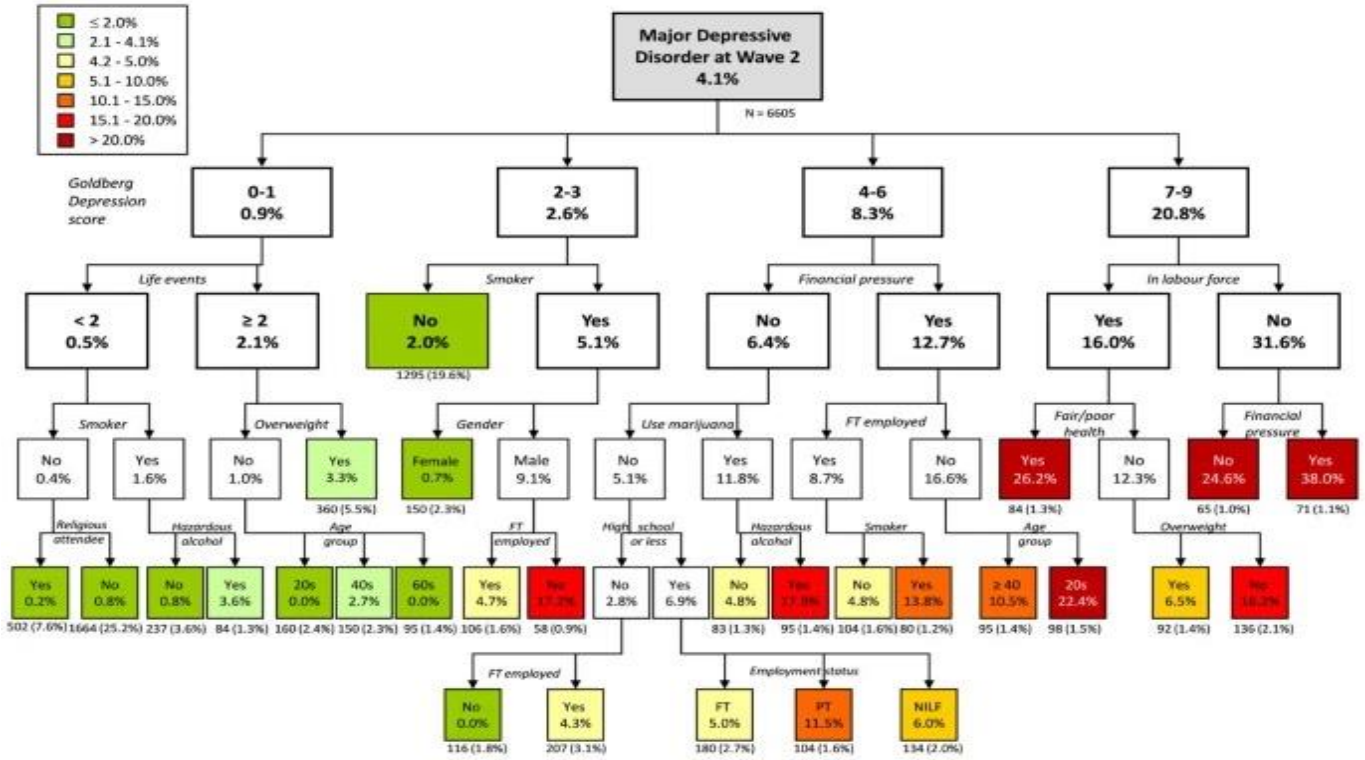


Figure 1 : Categorize the risk level of medical patients

Breiman, L, Friedman, JH, Olshen, RA & Stone, CJ 1984, *Classification And Regression Trees*, 1st edn, CRC Press, Pacific Grove, Calif.



Example of Decision Tree Analysis

In a study analyzing risk factors for **major depressive disorder (MDD)**, the decision tree model was used to identify significant risk factors (Batterham, Christensen and Mackinnon, 2009)from 17 candidates such as gender, age, and smoking. The model segmented individuals into 28 subgroups, revealing varying risks of MDD.

Figure 2. Decision tree predicting the risk of major depressive disorder at the four year follow-up of the PATH cohort.

Batterham, P.J., Christensen, H. and Mackinnon, A.J. (2009). Modifiable risk factors predicting major depressive disorder at four year follow-up: a decision tree approach. *BMC Psychiatry*, 9(1). doi:https://doi.org/10.1186/1471-244x-9-75.

Algorithms for Building Decision Trees

Several algorithms exist for constructing decision trees, including:

Algorithm	Description
CART (Classification and Regression Trees)	Uses Gini index and Twoing criteria for variable selection.
C4.5	Utilizes entropy info-gain for selecting input variables.
CHAID (Chi-Squared Automatic Interaction Detection)	Employs Chi-square tests for categorical variables; J-way ANOVA for continuous/ordinal variables.
QUEST (Quick, Unbiased, Efficient, Statistical Tree)	Similar to CART, but optimized for speed and bias.

Table 1. decision trees algorithms

Loh, W.-Y. (2014). Fifty Years of Classification and Regression Trees. *International Statistical Review*, 82(3), pp.329–348. doi:https://doi.org/10.1111/insr.12016.

Lewis, R. (2000). *An Introduction to Classification and Regression Tree (CART) Analysis Introduction to CART*. [online] Available at: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=6d4a347b99d056b7b1f28218728f1b73e64cbbac.

Song, Y.-Y. and Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry*, [online] 27(2). doi:https://doi.org/10.11919/j.issn.1002-0829.215044.

Advantage

Simplicity: They simplify complex relationships by dividing input variables into significant subgroups.

Interpretability: Easy to understand and interpret.

Non-parametric: No distributional assumptions.

Robustness: Effective with skewed data and resistant to outliers.

Limitations of Decision Trees

Overfitting: Particularly problematic with small datasets, affecting model generalizability.

Variable Correlation: Strong correlations among input variables may lead to misleading model statistics.

XGBoost

XGBoost (Extreme Gradient Boosting) is an efficient machine learning algorithm based on **Gradient Boosting Decision Tree** (GBDT), which is widely used in tasks such as classification, regression, sorting and anomaly detection. XGBoost outperforms traditional decision trees and random forests in computational efficiency(Nielsen,2016), prediction accuracy, and generalization, and is therefore widely used in Kaggle competitions and industrial applications

- https://github.com/dmlc/xgboost

Nielsen, D 2016, 'Tree Boosting With XGBoost - Why Does XGBoost Win "Every" Machine Learning Competition?', NTNU.

Chen, T & Guestrin, C 2016, 'XGBoost: A Scalable Tree Boosting System', in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 13-17-, ACM, New York, NY, USA, pp. 785–794.

XGBoost

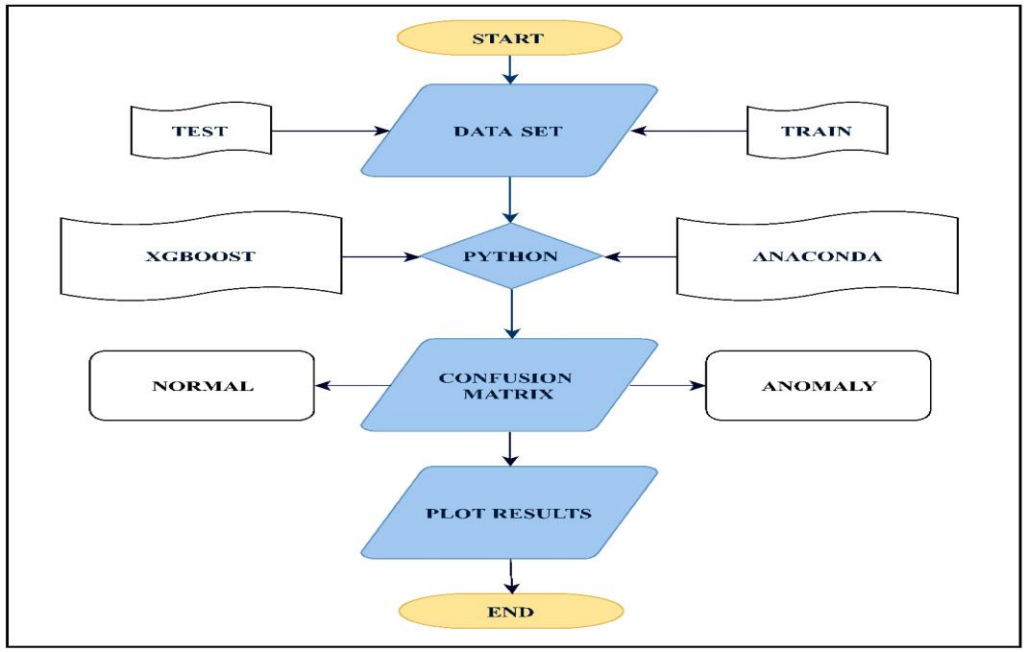


Figure 3. Flow of work chart.

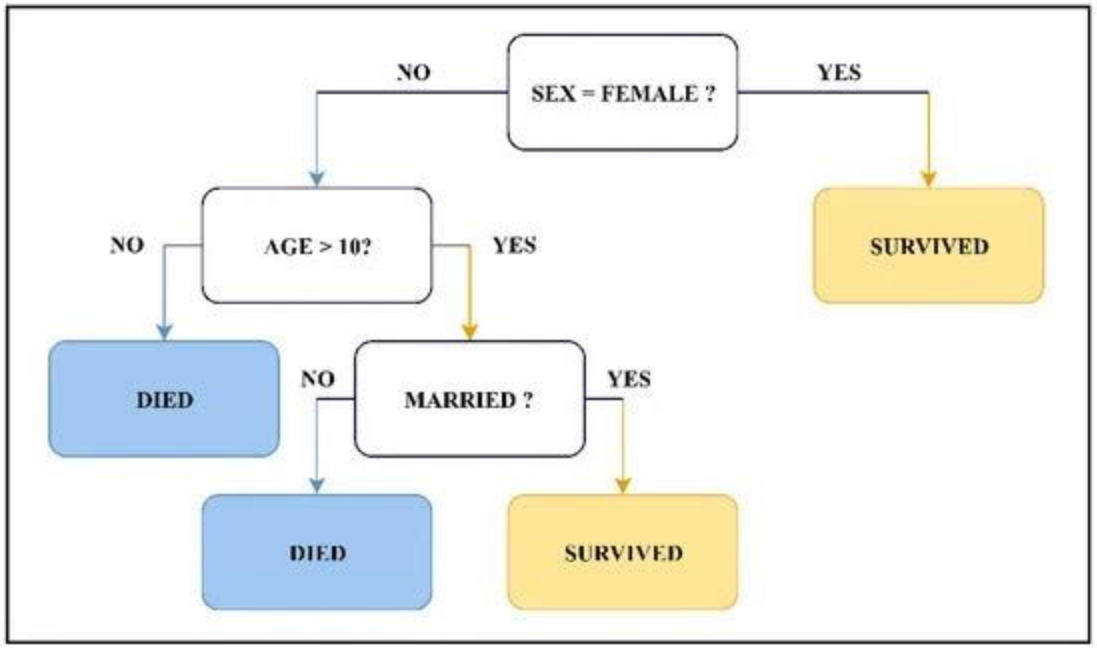


Figure 4. The working of Decision Trees.

Dhaliwal, S., Nahid, A.-A. and Abbas, R. (2018b). Effective Intrusion Detection System Using XGBoost. *Information*, 9(7), p.149. doi:https://doi.org/10.3390/info9070149.

XGBoost

1. First train the **first decision tree** to predict the outcome.
2. **Calculate the residual** (error) of the current tree and then train the next tree to fit the error.
3. **Iterate** until the error is minimized.

By iteratively training several weak decision trees, the new trees constantly correct the errors of the previous trees, and finally form a strong model.

The second derivative is used to optimize the loss function and improve the training speed and accuracy.

Advantage

Efficient computing (Dhaliwal, Nahid and Abbas, 2018) :Supports multi-threading and GPU acceleration, greatly reducing training time.

High prediction accuracy: Gradient Boosting with Second Order Approximation is used to improve the convergence speed and accuracy.

Prevent overfitting：Built-in regularization (L1/L2) is provided to reduce overfitting

Automatic processing of missing values(Sharma, 2018).

Flexible objective function: It is suitable for classification, regression, sorting (such as search engines), anomaly detection and other tasks.

Dhaliwal, S., Nahid, A.-A. and Abbas, R. (2018). Effective Intrusion Detection System Using XGBoost. *Information*, 9(7), p.149. doi:https://doi.org/10.3390/info9070149.

Sharma, N. (2018). *XGBoost. The Extreme Gradient Boosting for Mining Applications*. GRIN Verlag.

Limitations of XGBoost

- **Large consumption** of computing resources
- **Parameter tuning is complex** : Multiple parameters need to be adjusted (learning rate, depth of tree, subsampling rate, etc.)(Zhang, Jia and Shang, 2022).
- **Susceptible to noise**

Zhang, P., Jia, Y. and Shang, Y. (2022). Research and application of XGBoost in imbalanced data. *International Journal of Distributed Sensor Networks*, 18(6), p.155013292211069. doi:https://doi.org/10.1177/15501329221106935.

Conclusion --- Data size + Feature quantity

<10,000	10,000 – 100,000	100,000+
Decision tree	Random forest	XGBoost

<10	**10-100**	**100+**
Decision tree	Random forest	XGBoost

Breiman, L., Friedman, J., Stone, C.J. and Olshen, R.A. (2017). *Classification and Regression Trees*. Boca Raton Routledge Ann Arbor, Michigan Proquest.

Breiman, L. (2001). Random Forests. *Machine Learning*, [online] 45(1), pp.5–32. doi:https://doi.org/10.1023/a:1010933404324.

Chen, T & Guestrin, C 2016, 'XGBoost: A Scalable Tree Boosting System', in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, New York, NY, USA, pp. 785–794.