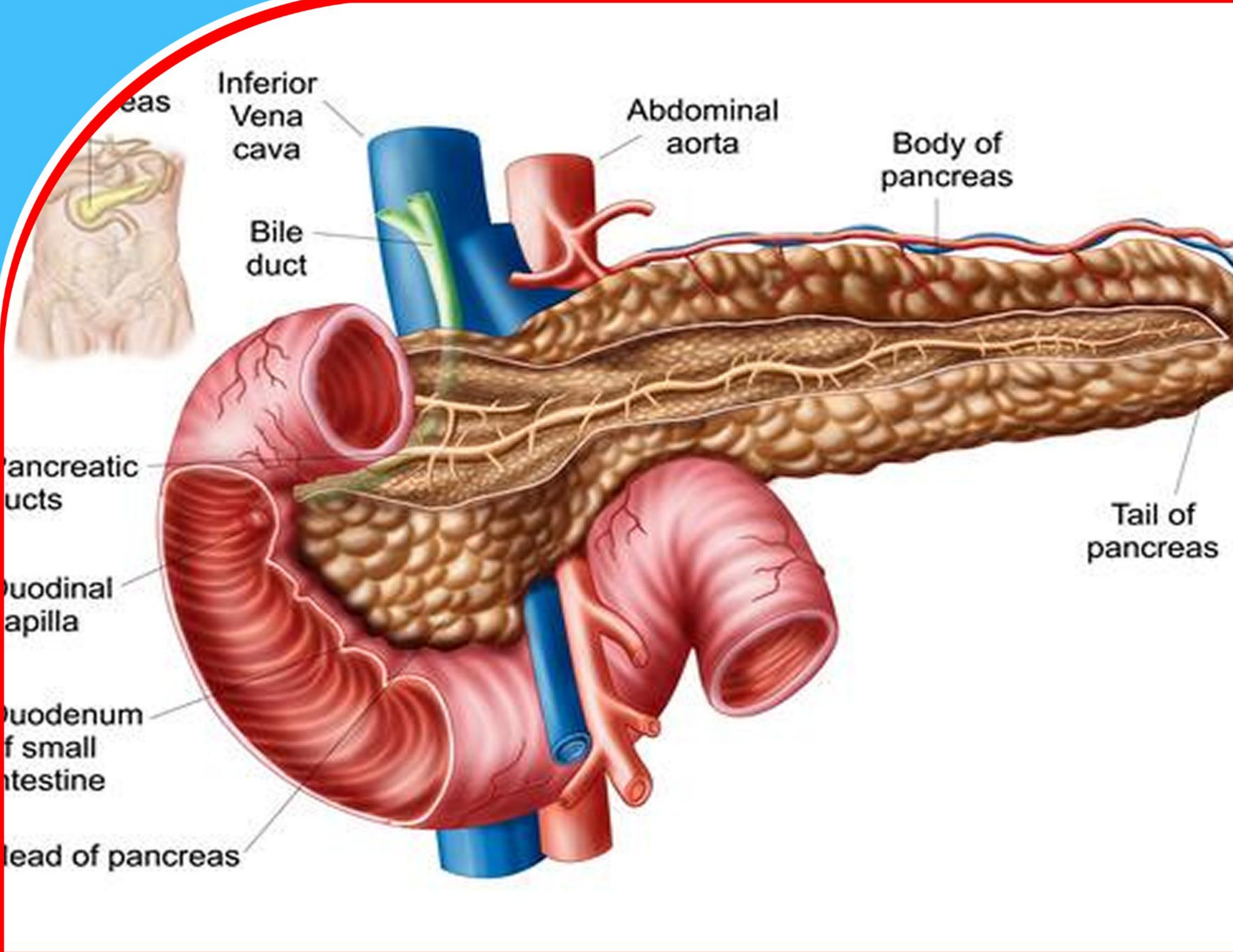
European Journal of
Biology
(EJB)



Decision Tree in Biology

Komal Shazadi

AJPO

European Journal of Biology
ISSN 2709-5886 (Online)
Vol.6, Issue 1, pp 1-15, 2021

www.ajpojournals.org

Decision Tree in Biology

Komal Shazadi

MS Biomedical Sciences, School of Mechanical and Manufacturing Engineering, National University of Sciences and Technology (NUST), Islamabad, Pakistan

Email: komalshazadi2020@gmail.com

ABSTRACT

Purpose: Human biology is an essential field in scientific research as it helps in understanding the human body for adequate care. Technology has improved the way scientists do their biological research. One of the critical technologies is artificial intelligence (AI), which is revolutionizing the world. Scientists have applied AI in biological studies, using several methods to gain different types of data. Machine learning is a branch of artificial intelligence that helps computers learn from data and create predictions without being explicitly programmed.

Methodology: One critical methodology in the machine is using the tree-based decision. It is extensively used in biological research, as it helps in classifying complex data into simple and easy to interpret graphs. This paper aims to give a beginner-friendly view of the tree-based model, analyzing its use and advantages over other methods.

Finding: Artificial intelligence has greatly improved the collection, analysis, and prediction of biological and medical information. Machine learning, a subgroup of artificial intelligence, is useful in creating prediction models, which help a wide range of fields, including computational and systems biology. Contribution and future recommendation also discussed in this study.

Keywords: *Artificial intelligence, tree-based, machine learning, biology analysis*

1. INTRODUCTION

Understanding of the human body is one of the most vibrant fields in Science, in an attempt to understand how the body functions to give it maximum care. The human body's complexity has led biologists to try to find new methods through which they examine the human body. The research has been due to the increased amount of biological data that can be analyzed, giving new insights into human molecular biology and understanding human genomes (Bhardwaj et al., 2017). The advanced methods have led to the growth of imaging techniques, new mass spectrometry methods, high-output sequencing methods, etc., made possible by the advanced technology and marrying technology and biology.

Among the many methods used in biological analyses is the use of artificial intelligence (Mathur, 2018). Artificial intelligence has already proved its significance in various fields such as decision making in banking (Donepudi, 2017). Moreover, it has provided solutions to multiple problems encountered in retail pharmacy (Donepudi, 2018). The methodology offers great opportunities, especially to anyone who wants to learn more about human microbiology due to how images can be manipulated and image prediction. Machine Learning (ML) is among the best tools which are used in biological analysis. The high number of tools and methods that can be utilized to create models, train data and present the information make it an ideal tool.

Supervised learning is the most common method used to manipulate large sets of data, especially where the input and the output are known. Since the outcome of the information depends on the 'smartness' of the model, which is directly proportional to the amount of data used to train, supervised learning is well suited for learning medical information (Olson et al., 2016). The information generated from such models is used to handle personalized medical care, drug discovery, manufacturing, radiography and radiotherapy, epidemic outbreak prediction, clinical research, and disease identification and diagnosis, among others. This paper focuses on the use of decision tree-based methods family of algorithms ("Tree-based learners," 2020). The algorithms are suited for this scenario due to their popularity in medicine and extensive use in biological data computation. Their popularity among medics makes it among the most researched methods in biological studies. The research aimed at addressing a research gap in the use of the decision tree-based algorithms in medicine.

1.1 Research Gap

Extensive research has been done on the use of machine learning in medicine (Sidey-Gibbons, 2019; Darcy, Louie & Roberts, 2016). The field attracts professionals from the fields of medicine and computer science who often use their high expertise in giving their insights on how machine learning can be used in medicine. While such research has been instrumental in addressing medicine challenges that have taken decades to solve, it has created great handles for young medical professionals who want to conduct research using these powerful ML tools (Ahmad et al., 2018). Therefore, there is a huge gap in giving a beginner-friendly explanation of how machine learning works, especially decision tree-based technologies.

1.2 Objective

To provide a comprehensive beginner-friendly analysis on the application of decision tree-based machine learning models in medicine by providing a survey of their use in systems and computational biology.

2. LITERATURE REVIEW

To help beginners without computer science knowledge get started, this paper starts by introducing the importance of data, concepts of artificial intelligence, then goes deeper into machine learning, brief literature on supervised learning, and finally a deep dive into decision tree-based methods, which is the main agenda of the paper. Definition of artificial intelligence depends on the persons who asked the question. The layman's description of Artificial intelligence (AI) gives computers the capacity to think and make decisions without being explicitly programmed, such as robots that work without needing user interference. A more technical person would describe the under-the-hood description of a set of algorithms that can replicate themselves and produce results without explicit instruction. While those definitions are correct, the second definition will be the most important as understanding how the AI works will help in knowing how to apply it to solve the decision tree-based models. Technologies based on AI uses data to learn and predict information (Huskins et al., 2017). Data in this context is any collection of facts such as $20, true, 3.142. To solve a particular problem using any method in AI, specific data is collected; for instance, to solve a problem in computational and systems biology, data from that field is needed.

AI is one of the most important breakthroughs in human existence. From a technical viewpoint, AI has helped organizations make an informed decision, as they can access far-reaching data analytics. In medicine, which is the domain for this paper, AI has helped in serious breakthroughs in biological research and medical care administration. AI can be broken down into expert systems, machine learning, neural networks, robotics, natural language processing, and fuzzy logic.

Machine learning is the most common application of artificial intelligence, which helps systems automatically learn from past data and experience. The distinctive aspect of machine learning is that it allows computers to access data from various sources and create predictions from that data. Computers learn automatically from the data and make their predictions without human involvement (Khan & Tappen, 2013). However, the type of learning that the computers go through is dependent on the machine learning method used. There are three common methods used: supervised learning, analogous with a lecturer giving their students assignments and helping them in doing them, unsupervised learning where the teacher gives assignments to the student and leaves the student to determine patterns from the data while reinforced training is similar to when the teacher takes a child into a new environment and lets them learn by interacting with the environment, giving rewards and punishments until the student learns the correct thing.

2.1 Supervised Learning

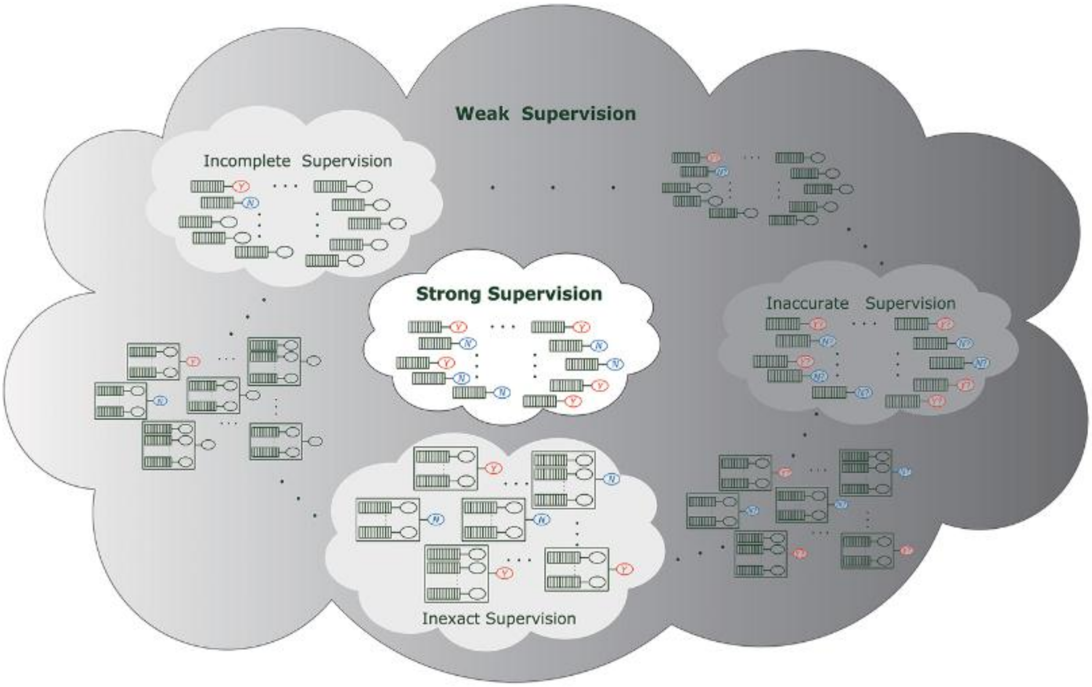


Figure 1. Learning process

The figure above shows the whole spectrum of supervised learning, the central umbrella under which all input-output pairs map new input to a new output. Supervision can be classified into weak supervision, intense supervision, and incomplete supervision, depending on the relationship between the input and the output (Mou et al., 2015). Incomplete supervision is when the model is given a small amount of data that cannot train a good learner. The model is then required to learn from that amount and classify the vast amount of unlabeled data.

Supervised learning can also take active learning mode, where the model assumes the user knows what the model wants. The model is designed to input from the user actively, called the teacher or the oracle of the information. Supervision can also be semi-supervised, which is among the most widely researched type of supervised machine learning (Nosratabadi et al., 2019).

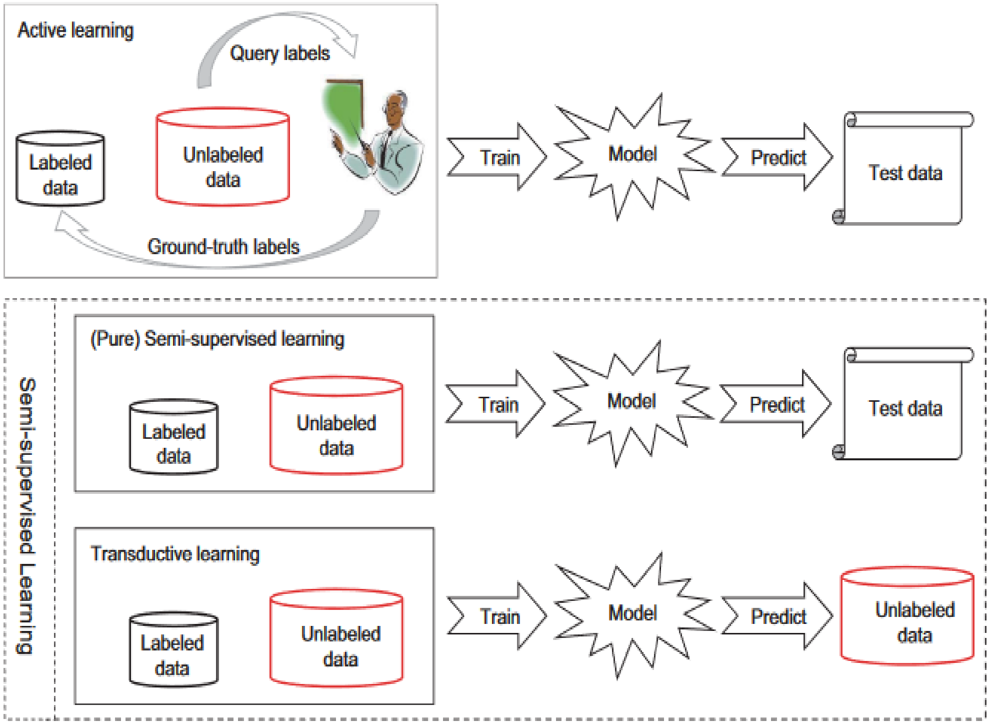


Figure 2. Supervised learning process

The main types of supervised learning are described in the above figure. As seen above, active learning and semi-supervised learning go through the same feeding data process into the model for training the predictions made (Olson & Moore, 2019). However, the active learning methods actively interacts with the user, while the semi-supervised learning does not need a cycle of data from the user. Moreover, as observed in the above figure, transductive training maps the predicted information as unlabeled data, while the other models map it into test data.

2.2 Decision Tree-Based Algorithms

Algorithms used in decision tree-based models are useful in the two steps in developing models of machine learning – training and prediction steps. Decision tree-based algorithms can be used for regression-based and classification-based algorithms, making them some of the most useful machine learning algorithms (Othman et al., 2018). The training model is determined in simple steps, which are predetermined during the training steps. The three central nodes or endpoints of decision tree-based algorithms are root node, decision node, and terminal node, as shown below.

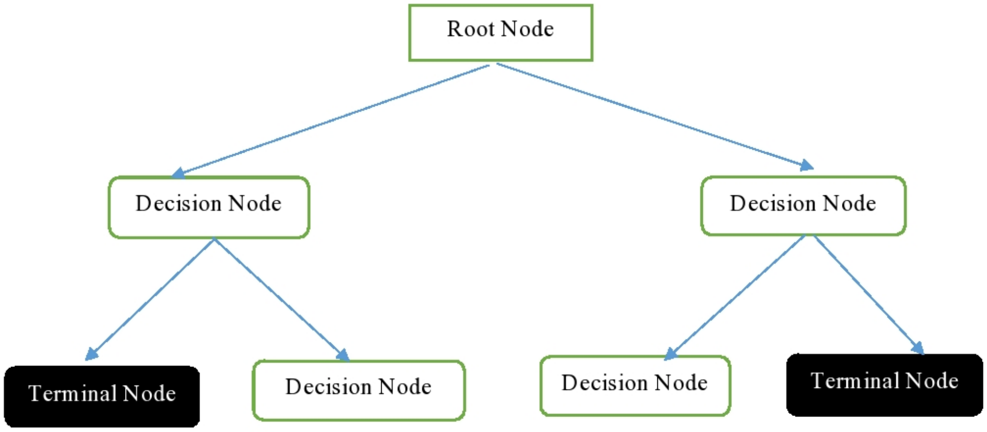


Figure 3. Decision Tree-Based Algorithms

The above diagram is a generic decision tree, starting from the root node and then branching off to the other nodes. The root node is the main entry point of the data, which gets classified into several terminals depending on the decision outcome. A decision node is a sub-node with different decisions, while a terminal/leaf node is the last decision to be made, where there are no other decisions (Huynh-Thu et al., 2010). Other terms used in the decision tree are pruning, which entails removing a decision node, splitting when a node is divided into two and, parent and child nodes representing the nodes that have been broken down.

2.3 Previous Work on Decision Tree-Based Algorithms

While this work focuses on creating a simplified version of Decision Tree-Based algorithms used in medicine, there has been extensive research on the subject (Darcy, 2016; Haider et al., 2008). However, much of the book targets individuals with extensive knowledge in both artificial intelligence and medicine. Historically, the method was invented as Concept Learning Systems by Hunt. However, the methods were expanded, popularized, and extended as Breiman, Friedman, Olsen through the book Classification and Regression of Trees (CART). The technique was also developed by Quinlan using the ID3 and C4.5 AI algorithms (Huynh-Thu et al., 2010). The algorithms were used primarily in medical research to analyze data from biological samples, although the methods have been extended to handle more complex problems besides classifying problems. The algorithm's initial set was used as shown by the tree below, mainly classifying a person as either sick or healthy. The decision tree usually had a corresponding decision boundary. In such a scenario, the decision tree and the corresponding decision boundary carry several decision points. The decision is made based on the threshold of the data provided, depending on the type of training that the algorithm has received.

There are many algorithms used in a decision tree, as demonstrated in the above diagram. The most critical approach in determining decision trees is finding the algorithm that makes it possible to give the highest prediction level based on the data provided (Pakdel & Herbert, 2017).

While it may seem evident that a more complex algorithm can offer a high accuracy prediction accuracy, that is usually not the case, although there are many exceptions. The scientist's goal in developing the model is to find the simplest model that gives the highest prediction level. To achieve that objective, all possible decision trees are grown on a heuristic basis. The greedy top-down recursive partitioning approach is the most common heuristic method that has been used in the past. The algorithm starts at the base node as the first layer where analysis is done to develop a decision, splitting into two parts. More splitting is done until the terminal node is reached. The method has been used extensively in the past in medical analysis.

3. METHODOLOGY

Several procedures are used in the generation of the learning algorithm. Those methods are done in increasing complexity, from the simplest one to the most challenging method. The methods are applied in the increasing order of need, where the most critical methods are done last since they significantly impact the algorithm.

3.1 Score Measure

Score measure is arguably the essential step in developing a decision tree. The score from each node determines the decision the algorithm will make and is used in making the whole tree.

	A	B	C	…	Z
1					
2					
…					
N					

The value obtained in each table is used in making the decision tree. The candidate with the best score is selected according to the preset criteria. Test discrimination and biases are added as the score measure is precisely designed to favor tests that create discrimination and biases on different candidates (Permanasari & Nurlayli, 2017, p. 76). A consistent formula is used to develop all the criteria for the decision table. The most common procedure is

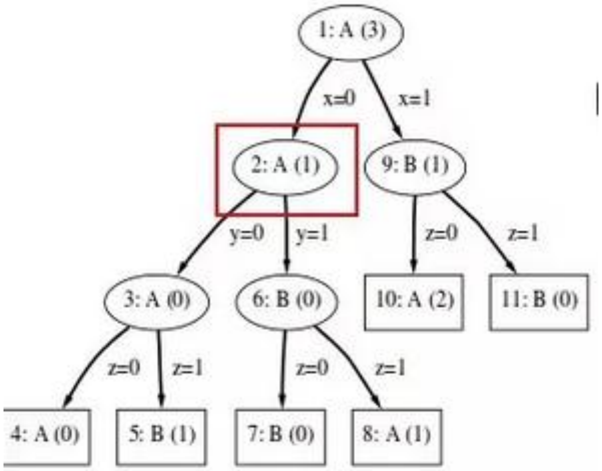
$$Score\ (S,T) = I(S) - \sum_{i=1}^{p} \frac{N_i}{N} I(S_i)$$

Where
T = Candidate Test
S = Sample Space for Local Learning of size N
I(S) = Measure of Impurity of Sample Space

The method can be manipulated using the Shannon's/logarithmic entropy formula or Gini's or quadratic entropy formula. For this paper's scope, how those formulas are applied is not essential, but a mention of what is mainly used to manipulate the tables.

In all machine learning models, the sample space has to satisfy a preset hypothesis that defines all the steps taken in the development of the algorithm, including selecting the candidate. In a standard decision tree algorithm, each candidate gets chosen only once per time (Permanasari & Nurlayli, 2017). The set provided is disseminated into all the possible values, which are treated as candidates. The candidates are now acted upon individually. The method of measuring scores is useful in the initial stages of working with models. Selecting each model individually is essential in ensuring successive methods can act upon it. Moreover, if the model achieves a threshold, it is optimized to provide even higher accuracy. Additional procedures are added to help improve the model's accuracy if it is insufficient at this stage.

3.2 Stop-splitting and Pruning



The decision tree can take any shape when subjected to the score measure formula. The scores may extend into unwanted sections, which usually affects the absolute accuracy of the model. The scenario is called a high variance state and exists due to the variance-bias tradeoffs. The state is undesirable due to its immense effects on the accuracy of the model. Two methods mainly used to remove the unwanted pieces are stop-splitting and pruning, whose duty is to prevent overfitting. The stop-splitting process is a simple algorithm applied in order to prevent the tree from splitting further after reaching a certain threshold. In contrast, pruning is used on an overgrown tree to reduce its complexity.

How, where, and when to apply stop-spitting and pruning methods are essential because their overall effect affects the model, making it more accurate, although they can make the models less accurate when misapplied. The standard criteria for stop-splitting are to preset a particular threshold value and applying the algorithm on any candidate that does not meet the threshold value. While the process helps increase the model's accuracy, it requires the user to be proactive in developing the model and have prefix values or meta-parameters that will need fixing. Pruning requires additional methods that will create a compromise for the nodes to determine which nodes should be thrown away. While pruning and stop-splitting may sound like simple techniques, there are several methods for each methodology for different scenarios.

Creating an accurate model, or a model with high accuracy may demand additional steps after pruning and stop-splitting. Pruning and stop-splitting methods can help increase the accuracy, although their effects are limited in most scenarios that require more methods. Additional methods used are randomization and boosting methods

3.3 Randomization Method

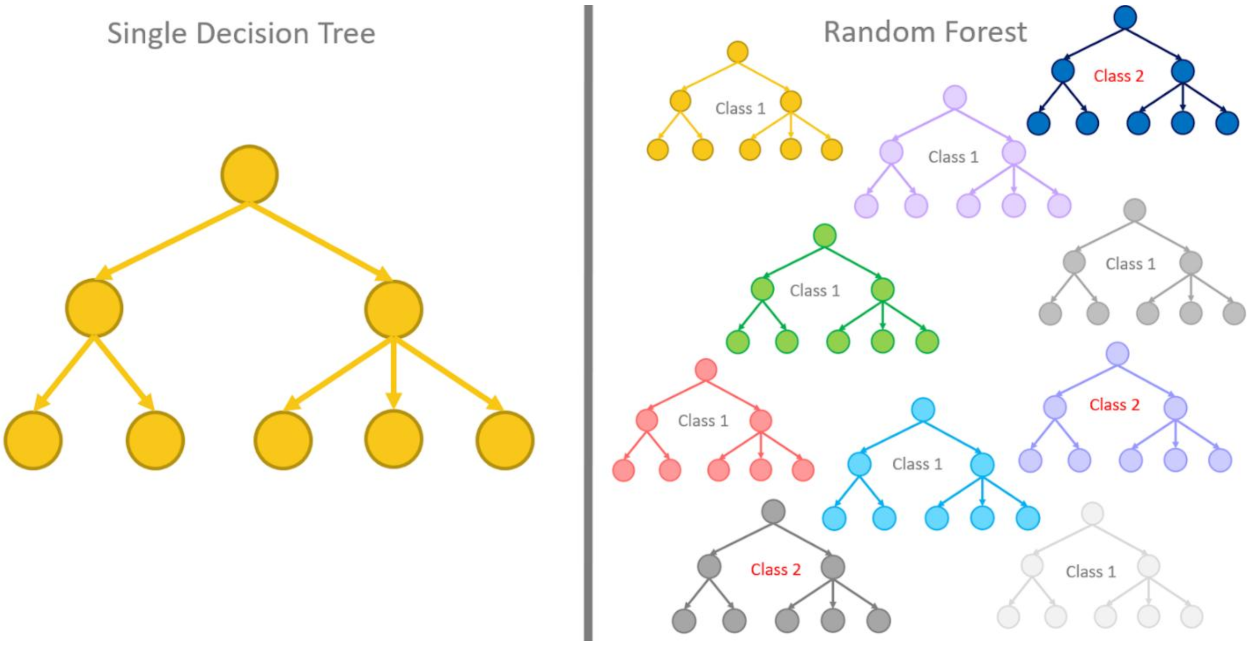


Figure 4. Tree-based Randomized method

The method is used to create perturbation to the tree-based process to generate several trees, as shown above. The tree on the left is broken down into small trees of different classes in a forest. The data is passed through all the trees to create an aggregation of predictions from each tree. The data generated is then used to create the final prediction. Generating the final prediction from the aggregated results can take different processes, depending on the data's variation and the accuracy of each aggregated result. When the variation in results is small, the common approach is to get the average prediction from individual trees. In contrast, most close predictions are taken when there is a huge variation in individual trees.

The most generic method in randomization is bagging since it can be used to randomize any learning algorithm. It is mainly used as it acts as a variance reducer without affecting the model's final accuracy. It is advantageous as it can be extended to perfectly predict the learning model, and they are less dependent on the learning sample, therefore creating a minimal bias in effect. Additionally, the method creates much smoother classification boundaries when compared to the original tree.

3.4 Boosting Methods

For a beginner, boosting methods can seem similar to the randomization methods, as they try to solve the problem using the same methodology of combining predictions of several models. However, the boosting method is very different under the hood because its objective is to 'boost the performance of weak performance model.' In contrast, randomization methods try to improve low-bias and high-variance data. The boosting is done sequentially, with the new tree trying to offset the previous model's challenges. Even the boosting method diagram is very different from the one from randomization, as shown below.

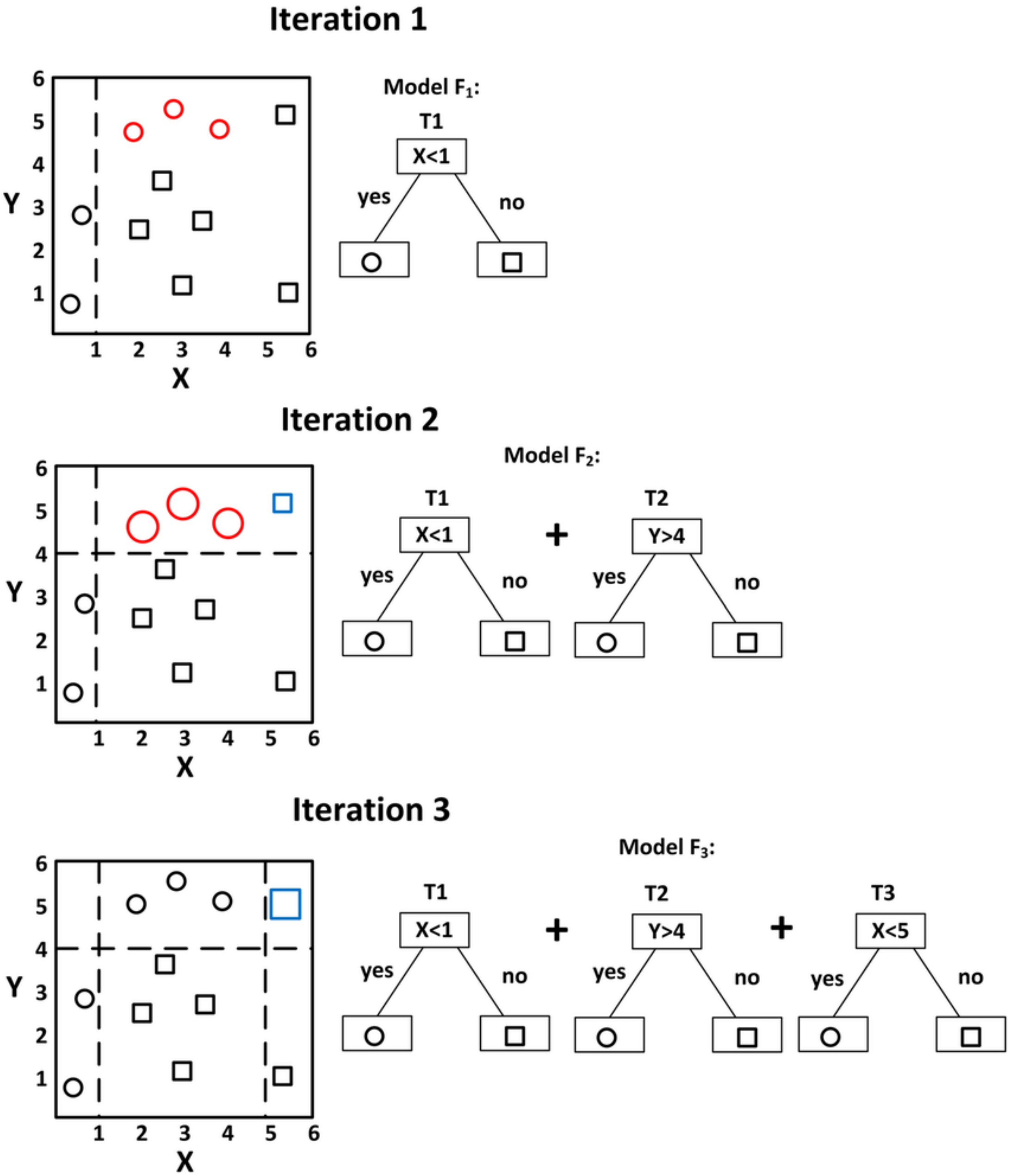


Figure 5. Boosting Method

In the first iteration, the decision tree is constructed normally, while the subsequent tree is constructed iteratively, considering the challenges of the previous model. The most common model is AdaBoost, designed by Freund and Schapire, which focuses on correcting the previous model's mistakes, assuming that the weighted values are cumulative. By taking into account the errors of the first model, the algorithm ensures that such errors will not affect the second model's accuracy, making it very accurate.

4. DISCUSSION

Decision tree methods are instrumental in the prediction of biological models. While a single tree is considered the truest method for prediction, its accuracy may be affected by several noises that need to be handled. Conventionally, more techniques are employed to increase the model's accuracy to help in prediction, as that is the whole essence of machine learning (Suresh et al., 2019). Pruning and stop-splitting methods are the first methods employed to try to increase the accuracy of the model. However, the methods are not as effective as needed because their effects, especially on the single tree, are not significant, requiring additional methods.

Boosting and randomization methods are applied to divide the single tree into many decision trees. The methods are instrumental in increasing the accuracy of the model to considerable numbers. However, the methods jeopardize the original data presented by the first tree model. The challenge of randomization methods affecting the overall accuracy of the models can be reduced by parallelizing the breakdown process, making the tree break breakdown as in the initial single tree case, while reducing the amount of time taken in the process. The boosting method creates a parallelizing challenge because the initial tree's outcome has to be used in the next tree. However, their overhead challenges are usually acceptable, as it increases the accuracy of the model.

4.1 Strengths and Limitations

Tree-based methods are potent in presenting any continuous functions with high precision, just like many supervised models. Their adaptation of single non-parametric and universal approximations helps provide relatively high accuracy prediction models when fed with sufficient complexity. Single trees can produce high consistent predictions without needing further modifications, although such happen under special conditions. Unlike many other models, tree models are useful in their predictability, making it easier to understand the prediction without external tools. They are highly predictable in construction and operation because they use the ranking mechanism to develop the best model (Rudin & Ustun, 2018). Moreover, their greedy approach makes the models very predictable in the way they process information. Improving the models' accuracy is possible by using forest models, which further break down a single tree into several trees like in randomization and boosting methods. It is no doubt that tree-based models are the go-to models for anyone trying to disseminate scientific data, such as biological experiment information, into useful predictions.

However, the method has several limitations, some of which may make the models hard to handle. The first challenge with tree-based models is high variance since the data making the single trees are selected from a mildly perturbed dataset, which can be very different. Compared with neural networks, tree-based methods become less competitive, as the high variance creates a high degree of inaccuracies (Rojas et al., 2018). When the models are tried using different samples, there is a high likelihood of the models to post different results. Such scenarios make the models less predictable and highly untrusted. The high variance affects the accuracy of the models for finely tuned models like vector machines. Consequently, tree-based models are not competitive when high accuracy is the main concern, as neural networks can post high accuracy.

Another possible challenge with tree-based is the storage requirements for the models. The number of nodes in an ensemble is higher than the original number of datasets in the model. Such high memory requirements become the limiting factor, especially on the amount of data that can be processed (Fishman, 2012). However, tree-based decision models' huge advantages and usability make them handy when dealing with simple, easy to use, and interpretable information is needed. Since most medical researchers are interested in using simpler to use models while giving a high degree of interpretation, the method is highly useful in the biological field.

4.2 Application in Computational and Systems Biology

4.2.1 Biomarker Discovery

Biomarkers such as mRNAs, peptides, and proteins are biomolecules that indicate biological conditions of interest like the stages of a disease or a differentiation. Biomarkers research and analysis require ranking and tissue classification, a scenario that can be effectively predicted using tree-based decisions. As established in genomics, it is critically important to distinguish between cancerous cells and healthy ones and distinguishing between different types of cancer (Shen et al., 2015). The decision tree boosting method can achieve those results within the shortest possible time. It has been established that tree-based decisions give the highest accuracy when dealing with data of this nature. The method promises research in identifying microbes, a problem challenging the medical professions for a long.

4.2.2 Sequence Annotation

Genomics is an interesting field of research because it helps provide effective medication, identifying traits hiding for a long and understanding of the human body. Identification and localization of transcriptional information inside human genomics have become possible through tree-based machine learning. The model helps determine any anomalies in the DNA and RNA sequence, coming in handy for genetics researchers.

4.2.3 Molecular Interaction

Interaction between different cell organelles is important in understanding human biology and identifying diseases and the treatment of those diseases. Supervised tree-based models help establish the human cells and identify tiny details in any form of an anomaly ("Proceedings of the 2018 ACM International Conference on bioinformatics, computational biology, and health informatics," 2018). Such methods are applied to identify specific types of proteins in human cells and classify them accordingly. Besides identifying the proteins and their classifications, the method has been in getting the precise location of any protein in question. There are other applications of tree-based machine learning models in biology, including their extensive use in genetics.

5. CONCLUSION

For a long time, biological research has faced several handles, especially in analyzing huge data. Such data obtained from human genomic include their genetic structure, protein structure, genomics sequence, and interaction of molecules. Additionally, medical information and care had posed several challenges, especially because analyzing human data was difficult. Artificial intelligence has greatly improved the collection, analysis, and prediction of biological and medical information. Machine learning, a subgroup of artificial intelligence, is useful in creating prediction models, which help a wide range of fields, including computational and systems biology.

Tree-based models are common in computational and systems biology. The models produce simple, easy to use and predictive models useful to young scientists who may not have expert knowledge in computer science and artificial intelligence while providing state of the art technology with high accuracy.

A single tree may provide high-quality forecast models, but it is not useful when it has a high disparity. Several methods are used to increase the accuracy of the models produced by the tree-based model. The first method applied is the pruning and stop-splitting methods, which reduce disparity in the data produced by a single tree. Additional methods are used to increase the accuracy of the models. Common methods are randomization and boosting methods, which reduce the single tree into a forest of trees. The methods produce high-quality results that can be used to predict with high accuracy.

Research implication

This study open new doors of research towards Machine learning as a branch of artificial intelligence that will be helpful to learn from data and create predictions without being explicitly programmed. Secondly, this paper will also help to improve the field of medicine by using artificial intelligence and specifically machine learning. This research will provide a comprehensive details and the application of decision tree-based machine learning models in medicine to the doctors, policy makers and for the medical group.

Future Recommendations

This study give direction to the researchers and practitioners for the future as machine learning is not limited to the specific field of life but it plays important role everywhere. This study was qualitative by nature, future study can be conducted by quantitative techniques. More recent examples can be added to enhance the findings. Using the modular code structure we offer, machine learning users who have focused on coding with the examples would be well-placed to further improve their skills working on more complicated datasets. Furthermore, an important concept of ML is also usefully illustrated by this data: more complicated algorithms do not inherently yield more useful predictions.

REFERENCES

Ahmad, M. A., Eckert, C., & Teredesai, A. (2018). Interpretable machine learning in healthcare. *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. https://doi.org/10.1145/3233547.3233667

Bhardwaj, R., Nambiar, A. R., & Dutta, D. (2017). A study of machine learning in healthcare. *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)*. https://doi.org/10.1109/compsac.2017.164

Darcy, A.M., Louie, A.K., & Roberts, L.W. (2016). Machine Learning and the Profession of Medicine. J Am Med Assoc, 315(6), 551. https://doi.org/10.1001/jama.2015.18421.

Donepudi, P. K. (2017). AI and Machine Learning in Banking: A Systematic Literature Review. *Asian Journal of Applied Science and Engineering* **6**(3): 157-162.

Donepudi, P. K. (2018). AI and Machine Learning in Retail Pharmacy: Systematic Review of Related Literature. *ABC Journal of Advanced Research* **7**(2): 109-112.

Fishman, N. (2012). Policy statement on antimicrobial stewardship by the society for healthcare epidemiology of America (SHEA), the Infectious Diseases Society of America (IDSA), and the pediatric infectious diseases society (PIDS). *Infection Control & Hospital Epidemiology*, *33*(4), 322-327. https://doi.org/10.1086/665010

Haider AH, Chang DC, Efron DT, Haut ER, Crandall M, Cornwell EE. Race and Insurance Status as Risk Factors for Trauma Mortality. Arch Surgery, 143(10), 945. https://doi.org/10.1001/archsurg.143.10.945.

Huskins, W. C., Fowler, V. G., & Evans, S. (2017). undefined. *Clinical Infectious Diseases*, *66*(7), 1140-1146. https://doi.org/10.1093/cid/cix907

Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., & Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE*, *5*(9), e12776. https://doi.org/10.1371/journal.pone.0012776

Khan, N., & Tappen, M. F. (2013). Discriminative dictionary learning with spatial priors. *2013 IEEE International Conference on Image Processing*. https://doi.org/10.1109/icip.2013.6738035

Mathur, P. (2018). Overview of machine learning in healthcare. *Machine Learning Applications Using Python*, 1-11. https://doi.org/10.1007/978-1-4842-3787-8_1

Mou, L., Peng, H., Li, G., Xu, Y., Zhang, L., & Jin, Z. (2015). Discriminative neural sentence modeling by tree-based convolution. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/d15-1279

Nosratabadi, S., Mosavi, A., Keivani, R., Faizollahzadeh ardabili, S., & Aram, F. (2019). State of the art survey of deep learning and machine learning models for smart cities and urban sustainability. https://doi.org/10.20944/preprints201908.0154.v1

Olson, R. S., & Moore, J. H. (2019). TPOT: A tree-based pipeline optimization tool for automating machine learning. *Automated Machine Learning*, 151-160. https://doi.org/10.1007/978-3-030-05318-5_8

Olson, R. S., Urbanowicz, R. J., Andrews, P. C., Lavender, N. A., Kidd, L. C., & Moore, J. H. (2016). Automating biomedical data science through tree-based pipeline optimization. *Applications of Evolutionary Computation*, 123-137. https://doi.org/10.1007/978-3-319-31204-0_9

Othman, M., Ratna, S., Tewari, A., Kang, A., & Vaisman, I. (2018). Machine learning classification of antimicrobial peptides using reduced alphabets. *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. https://doi.org/10.1145/3233547.3233657

Pakdel, R., & Herbert, J. (2017). Adaptive cost efficient framework for cloud-based machine learning. *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)*. https://doi.org/10.1109/compsac.2017.42

Permanasari, A. E., & Nurlayli, A. (2017). Decision tree to analyze the cardiotocogram data for fetal distress determination. *2017 International Conference on Sustainable Information Engineering and Technology (SIET)*. https://doi.org/10.1109/siet.2017.8304182

Proceedings of the 2018 ACM International Conference on bioinformatics, computational biology, and health informatics. (2018). https://doi.org/10.1145/3233547

Rojas, J. C., Carey, K. A., Edelson, D. P., Venable, L. R., Howell, M. D., & Churpek, M. M. (2018). Predicting intensive care unit readmission with machine learning using electronic health record data. *Annals of the American Thoracic Society*, *15*(7), 846-853. https://doi.org/10.1513/annalsats.201710-787oc

Rudin, C., & Ustun, B. (2018). Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice. *Interfaces*, *48*(5), 449-466. https://doi.org/10.1287/inte.2018.0957

Shen, D., Zhang, D., Young, A., & Parvin, B. (2015). Editorial: Machine learning and data mining in medical imaging. *IEEE Journal of Biomedical and Health Informatics*, *19*(5), 1587-1588. https://doi.org/10.1109/jbhi.2015.2444011

Suresh, A., Udendhran, R., & Balamurgan, M. (2019). Hybridized neural network and decision tree based classifier for prognostic decision making in breast cancers. *Soft Computing*, *24*(11), 7947-7953. https://doi.org/10.1007/s00500-019-04066-4

Tree-based learners. (2020). *Machine Learning Refined*, 443-470. https://doi.org/10.1017/9781108690935.019