Decision Trees And Ensemble Methods

C5.0 and Random Forest Algorithms

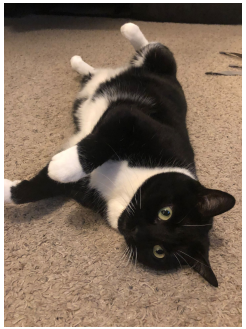
Team Matzinger

Selena Halabi, Kewei Ye, Jordan Smiley, Lexi Wittstadt, and our pets

Loki

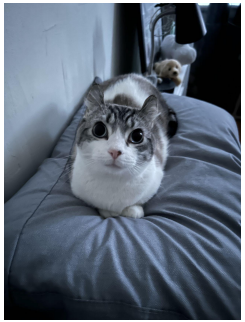
Rebel

Rogue



Ophelia

Anna



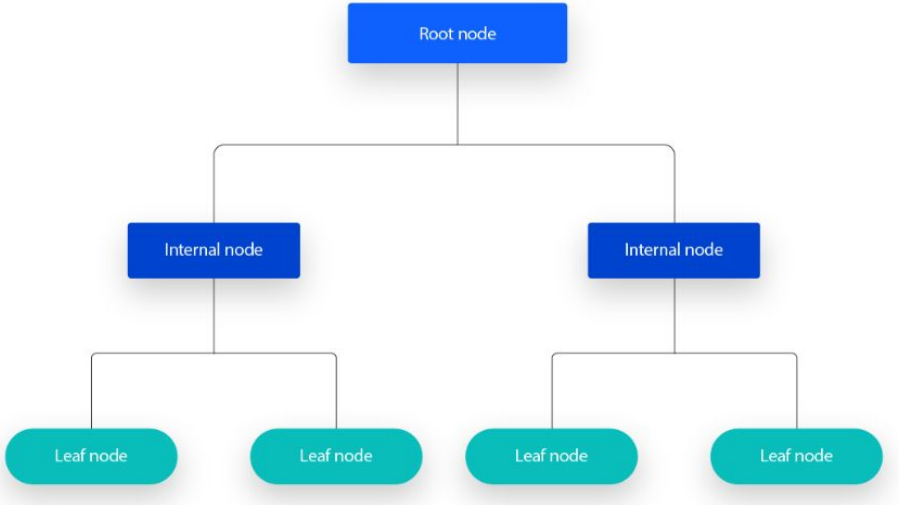
Bagel

Lola

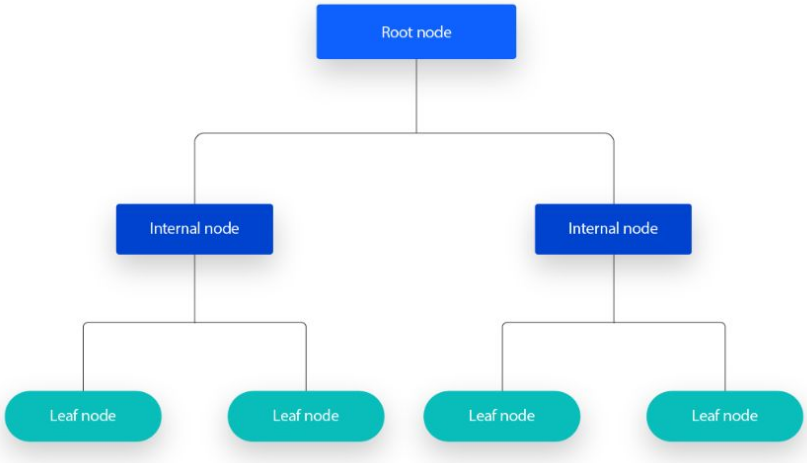
What is a decision tree?

A non-parametric supervised learning algorithm

- used for both classification and regression tasks
- has hierarchical, tree structure



https://www.ibm.com/topics/decision-trees



https://www.ibm.com/topics/decision-trees

Terminologies

Root Nodes: represents the original choice or feature from which the tree branches; the highest node

Internal (Decision) Nodes: node whose choices are determined by the values of particular attributes; branches on these nodes that go to other nodes

Leaf (Terminal) Nodes: choices or forecasts are decided upon; no more branches off these

Branch (Edge): Links between nodes that show how decisions are made in response to particular circumstances.

How is a decision tree formed?

A decision tree **recursively partitions** the data based on the values of different attributes.

- The algorithm selects the best attribute to split the data at each internal node, based on certain criteria
- Process continues until a **stopping criterion** is reached, such as reaching a maximum depth or having a minimum number of instances in a leaf node

Key Assumptions

1. Binary splits: each node divides the data into two subsets based on a single feature or condition
2. Recursive partitioning: each node can be divided into child nodes until stopping criterion met
3. Feature independence
4. Homogeneity: assumes homogeneous subgroups in each node
5. Top-down greedy approach: each split is chosen to maximize information gain or minimize impurity at the current node

Attribute Selection Measures

1. **Information Gain**
   - Entropy: measure of disorder/uncertainty in dataset

$$Information\ Gain = E(Y) - E(Y|X)$$

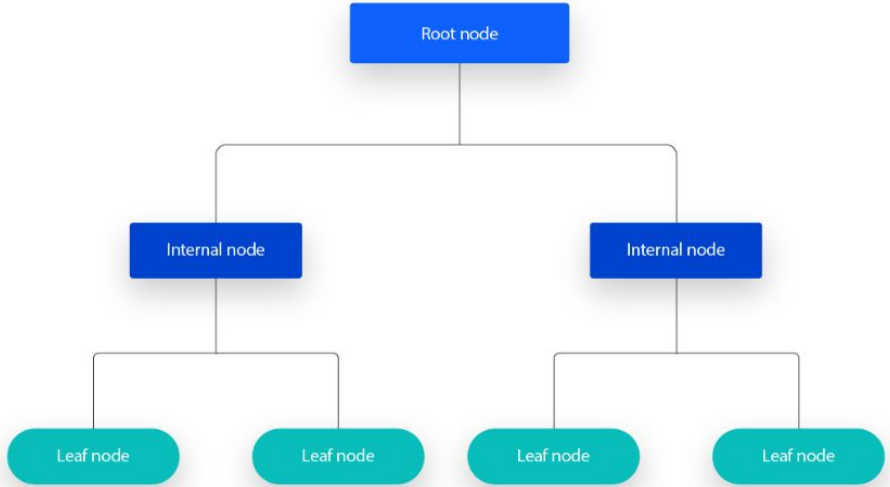
2. **Gini Index**: metric to measure how often a randomly chosen element would be incorrectly identified; measure of inequality/impurity of a distribution
   - Ranges from 0 to 0.5
     - 0 = pure set (all instances belong to the same class)
     - 0.5 = maximally impure set (instances are evenly distributed across classes)

$$Gini = 1 - \sum_{i=1}^{n} (p_i)^2$$

Why use a decision tree?

Decision trees are widely used in machine learning because of their

- High interpretability and versatility
- Proficiency with both numerical and categorical data
- Simple visualization

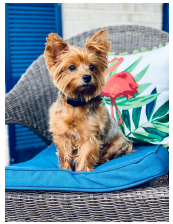


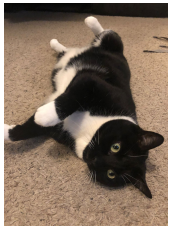
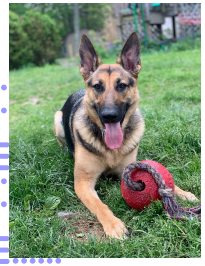
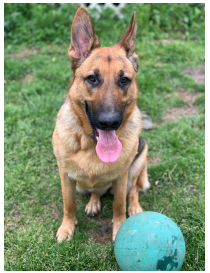
C5. 0 Algorithm

Entropy and Information Gain

- Splits a certain attribute to measure entropy reduction
    - Subtacts the entropy of the split from the original entropy
    - Makes it more organized bases on homogeneous grouping
- Builds a rule set (decision tree) to divide the subsamples determined by the initial split, then repeated
    - Optimizes information gain
- Predicts categorical outcomes based on the input features
- Chooses the best feature at each node
- Size and quality of the subgroups to determine the best splits
- Pruning mechanism to prevent overfit
    - Removes splits that are not useful

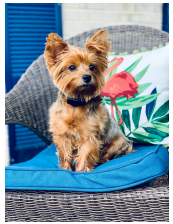
Example of C5.0 -Categorizing our pets!

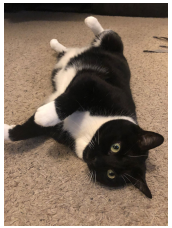
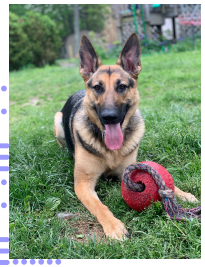
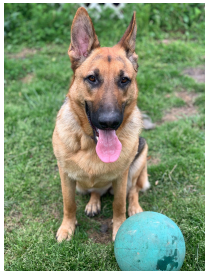


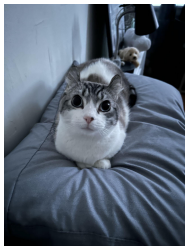




1. Overall Entropy of our pets - chaotic (there are a lot of them) let says we'll quantify by the number of pictures in each category
2. What method of dividing will lower the entropy the most? Color? Ear Shape? Size? Presence of whiskers? Presence of a toy?
3. Calculate the entropy of each option....
4. Choose the one that lowers the entropy

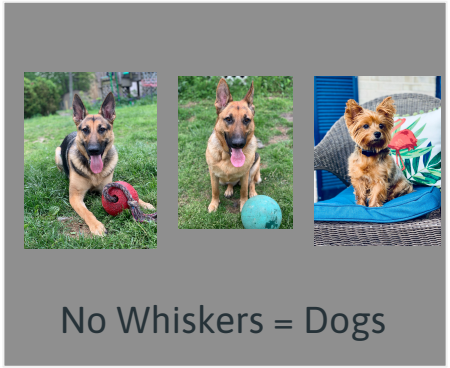
Example of C5.0 - Categorizing our pets !



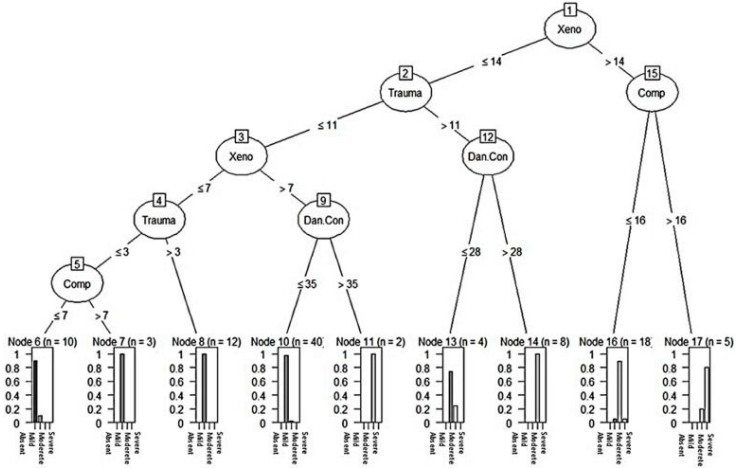






1. Overall Entropy of our pets - chaotic (there are a lot of them) let says we'll quantify by the number of pictures in each category
2. What method of dividing will lower the entropy the most? Color? Ear Shape? Size? Presence of whiskers? Presence of a toy?
3. Calculate the entropy of each option....
4. Choose the one that lowers the entropy



C5.0 For Real



Delgado-Gallegos, J. L., Avilés-Rodriguez, G., Padilla-Rivas, G. R., De Los Ángeles Cosío-León, M., Franco-Villareal, H., Nieto-Hipólito, J. I., de Dios Sánchez López, J., Zuñiga-Violante, E., Islas, J. F., & Romo-Cardenas, G. S. (2023). Application of C5.0 Algorithm for the Assessment of Perceived Stress in Healthcare Professionals Attending COVID-19. *Brain sciences*, *13*(3), 513. https://doi.org/10.3390/brainsci13030513

Factors used to classify healthcare professionals:
Fear of contamination, social economical, xenophobia, traumatic stress, compulsive stress

Application of C5.0 Algorithm for the Assessment of Perceived Stress in Healthcare Professionals Attending COVID-19

Juan Luis Delgado-Gallegos [1], Gener Avilés-Rodriguez [2], Gerardo R Padilla-Rivas [1], María De los Ángeles Cosío-León [3], Héctor Franco-Villareal [4], Juan Iván Nieto-Hipólito [5], Juan de Dios Sánchez López [5], Erika Zuñiga-Violante [5], Jose Francisco Islas [1], Gerardo Salvador Romo-Cardenas [5,*]

Editors: Giovanni Martinotti, Mauro Pettorruso, Domenico De Berardis, Drozdstoi Stoyanov

Author information · Article notes · Copyright and License information

PMCID: PMC10046351 PMID: 36979323

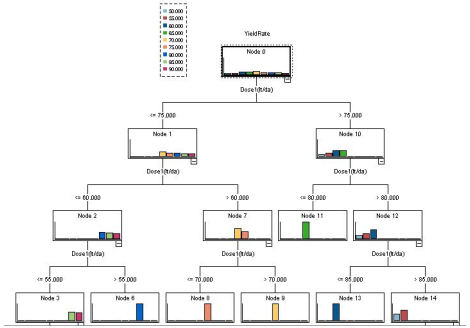
Abstract

Coronavirus disease (COVID-19) represents one of the greatest challenges to public health in modern history. As the disease continues to spread globally, medical and allied healthcare professionals have become one of the most affected sectors. Stress and anxiety are indirect effects of the COVID-19 pandemic. Therefore, it is paramount to understand and categorize their perceived levels of stress, as it can be a detonating factor leading to mental illness. Here, we propose a computer-based method to better understand stress in healthcare workers facing COVID-19 at the beginning of the pandemic. We based our study on a representative sample of healthcare professionals attending to COVID-19 patients in the northeast region of Mexico, at the beginning of the pandemic. We used a machine learning classification algorithm to obtain a visualization model to analyze perceived stress. The C5.0 decision tree algorithm was used to study datasets. We carried out an initial preprocessing statistical analysis for a group of 101 participants. We performed chi-square tests for all questions, individually, in order to validate stress level calculation ($p < 0.05$) and a calculated Cronbach's alpha of 0.94 and McDonald's omega of 0.95, demonstrating good

C5.0 Strengths & Weaknesses

Strengths:

- Scalable and efficient
- Handles both continuous and categorical characteristics
- Robust pruning mechanism
- Capable of handling noisy data
- Interpretable and intuitive



Weaknesses:

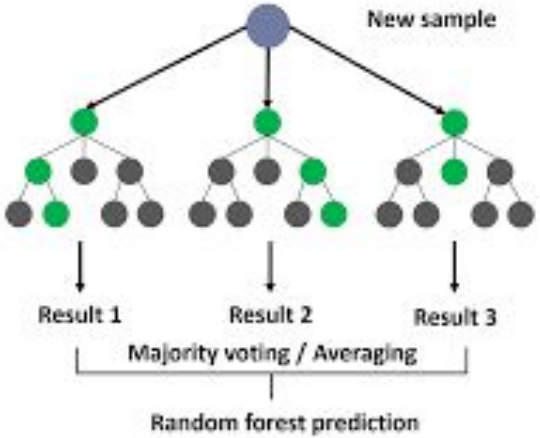
- Strongly connected qualities may cause the C5 algorithm to perform less effectively as it may place too much emphasis on one feature at the expense of others.
- Careful parameter selection is necessary
- Sensitive to missing values
- Complex nonlinear interactions may not be a good fit for decision trees.

Random Forest - Decision tree algorithm

Ensemble learning technique
Utilizes bagging/bootstrap sampling
Merges the results of multiple decision trees into one result
Can handle both continuous and categorical data

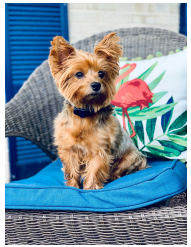
How it works:

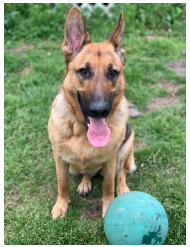
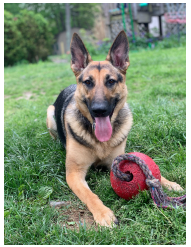
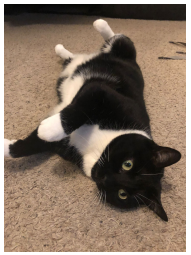
1. Select a random subset of the data set (size K)
2. Make a decision tree from this subset of data
3. Take a new subset of data (with replacement) and repeat, again and again!
4. For new data, compare the output of the decision trees/aggregate the data
5. The most common result wins!



Yehoshua, R. Random Forests. Medium.com.

Random Forest Example Using Our Pets!



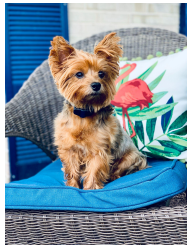




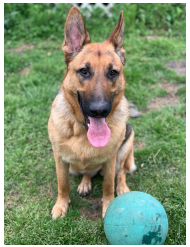
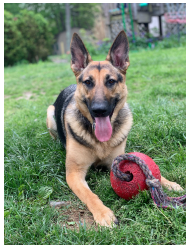
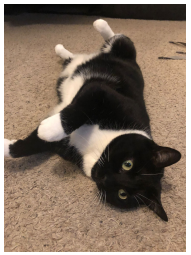
Let's say we are trying to determine if a pet will have an outgoing or shy personality.

Let's say the seven pets here are our "training set". We know their personalities already – it's a supervised training model

How can we do this with random forest?

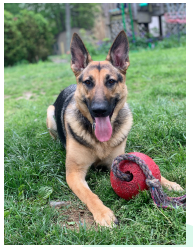
Random Forest Example Using Our Pets!



First, we pick a subset of the seven pets randomly. Let's say we pick 4, and these are the ones chosen:



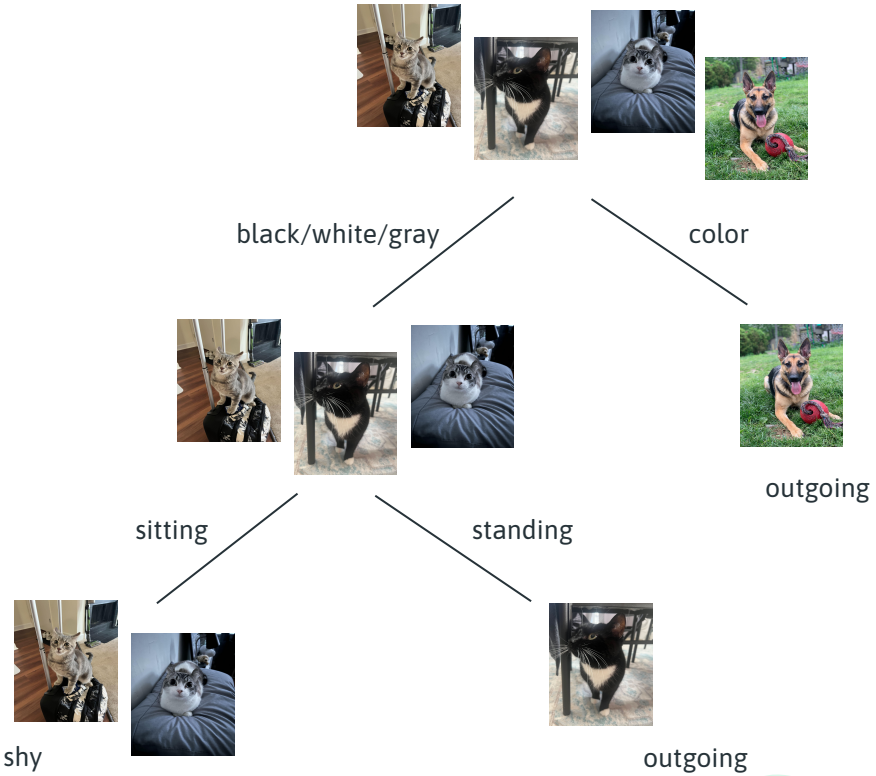



Random Forest Example Using Our Pets!

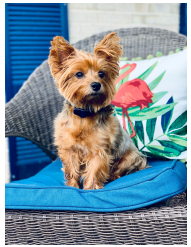
Now we create a decision tree!

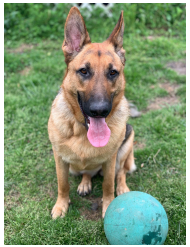
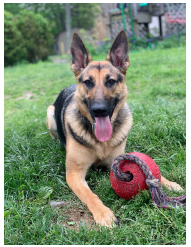
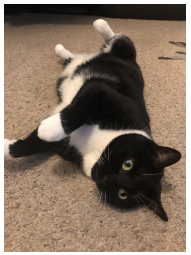
The model will create nodes using random characteristics.

The outcome is that two groups are classified as outgoing, and one group is shy.

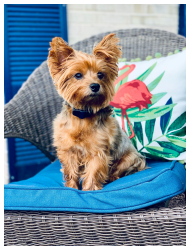


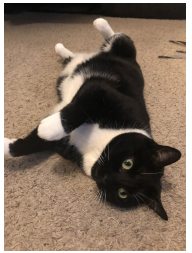
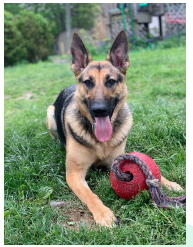
Random Forest Example Using Our Pets!



Now we pick a different subset of pets. We are allowed to repeat!



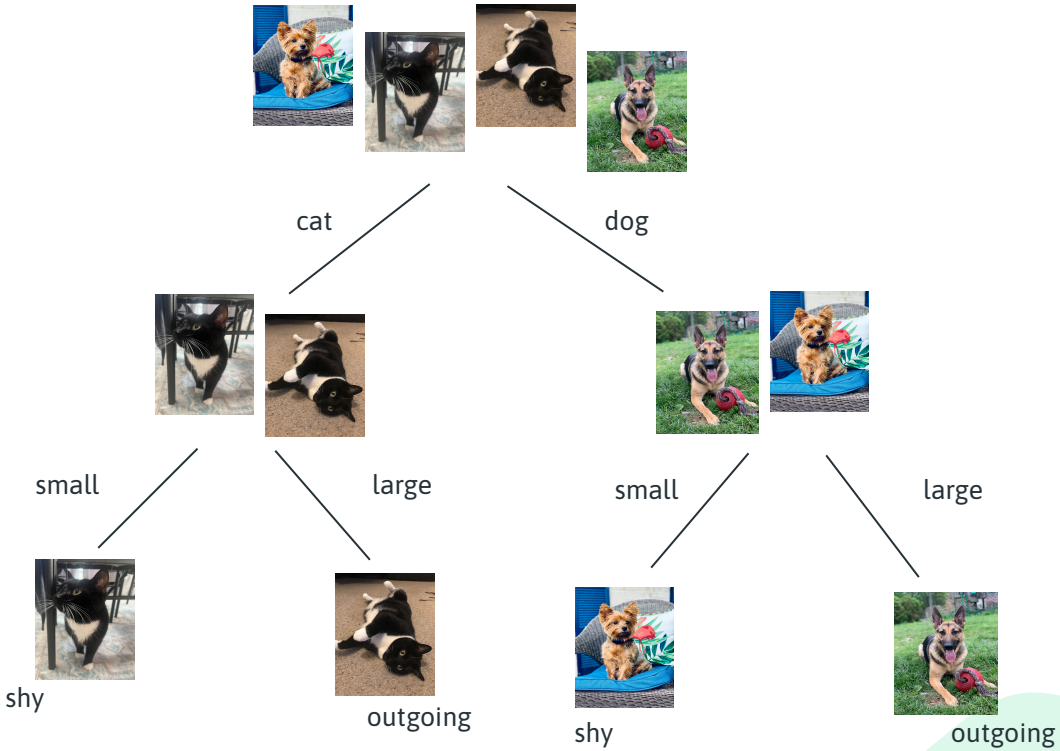



Random Forest Example Using Our Pets!

Now we create a decision tree (again)!

The model will create nodes using random characteristics.

The outcome is that two groups are classified as outgoing, and two groups are shy.



Random Forest Example Using Our Pets!

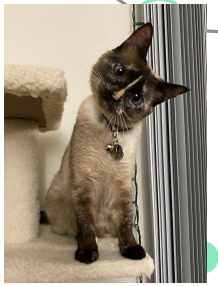
We do this a lot of times, to make a “forest” of trees.

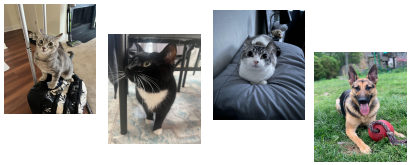
For our simple example, however, let’s say our whole model is these two trees.

Now we take new data points and run it through this collection.

Let’s use my sister’s cat Alasia!

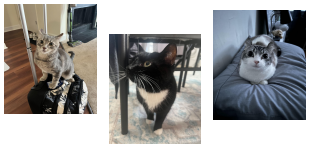
Random Forest Example Using Our Pets!

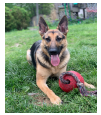




black/white/gray

color

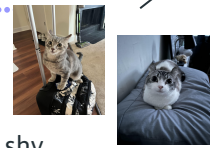




outgoing

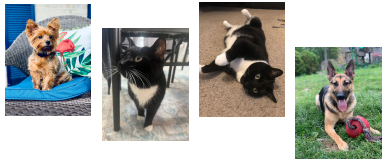
sitting

standing



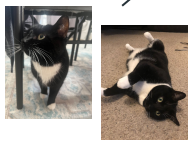
shy

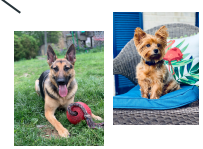
outgoing



cat

dog





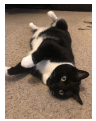
small

large

small

large

outgoing



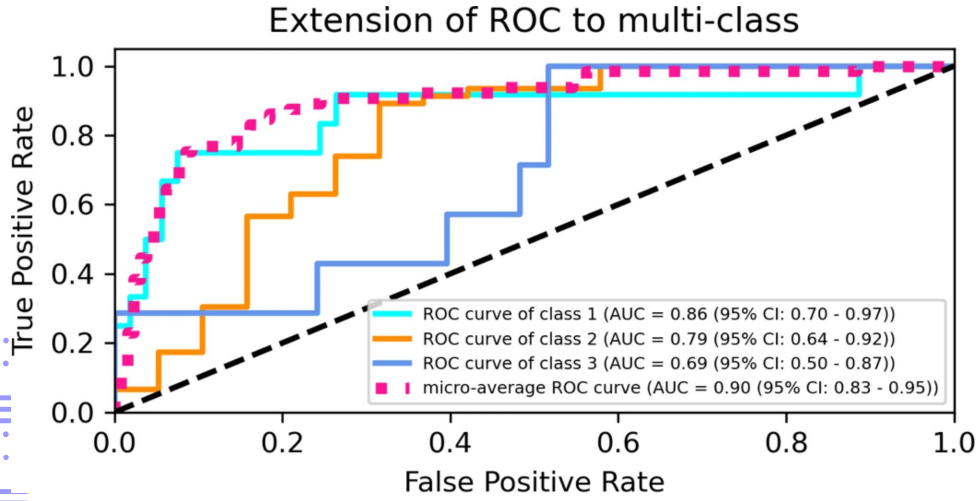
shy

outgoing

shy

Alasia is classified as outgoing! (This is definitely true, so our model is good!)

Random Forest - Real World Example



Class 1: COVID-19 positive
Class 2: Mycoplasma pneumoniae positive
Class 3: co-infected

nature › scientific reports › articles › article

Article | Open access | Published: 30 September 2024

Using random forest and biomarkers for differentiating COVID-19 and *Mycoplasma pneumoniae* infections

Xun Zhou, Jie Zhang, Xiu-Mei Deng, Fang-Mei Fu, Juan-Min Wang, Zhong-Yuan Zhang, Xian-Qiang Zhang, Yue-Xing Luo & Shi-Yan Zhang

Scientific Reports **14**, Article number: 22673 (2024) | Cite this article

422 Accesses | Metrics

Abstract

The COVID-19 pandemic has underscored the critical need for precise diagnostic methods to distinguish between similar respiratory infections, such as COVID-19 and *Mycoplasma pneumoniae* (MP). Identifying key biomarkers and utilizing machine learning techniques, such as random forest analysis, can significantly improve diagnostic accuracy. We conducted a retrospective analysis of clinical and laboratory data from 214 patients with acute respiratory infections, collected between October 2022 and October 2023 at the Second Hospital of Nanping. The study population was categorized into three groups: COVID-19 positive (n = 52), MP positive (n = 140), and co-infected (n = 22). Key biomarkers, including C-reactive protein (CRP), procalcitonin (PCT), interleukin- 6 (IL-6), and white blood cell (WBC) counts, were evaluated. Correlation analyses were conducted to assess relationships between biomarkers within each group. The random forest analysis was applied to evaluate the discriminative power of these biomarkers. The random forest model demonstrated high

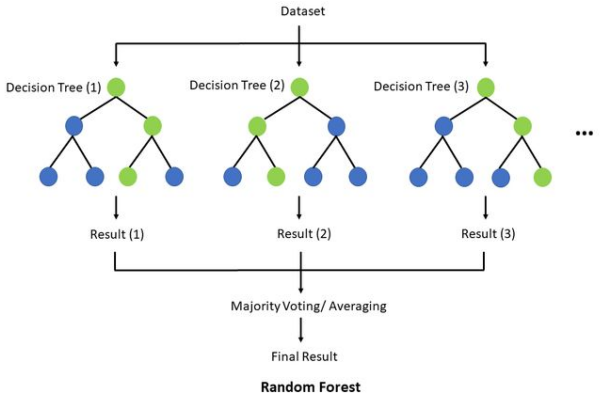
Random Forest Strengths & Weaknesses

Strengths:

- High accuracy
- Robustness to noise
- Non-parametric nature
- Estimating feature importance
- Handles missing data and outliers
- Handles both numerical and categorical data
- Generally avoids overfitting (but it can happen)

Weaknesses:

- Computational Complexity
- Memory usage
- Prediction time
- Lack of interpretability

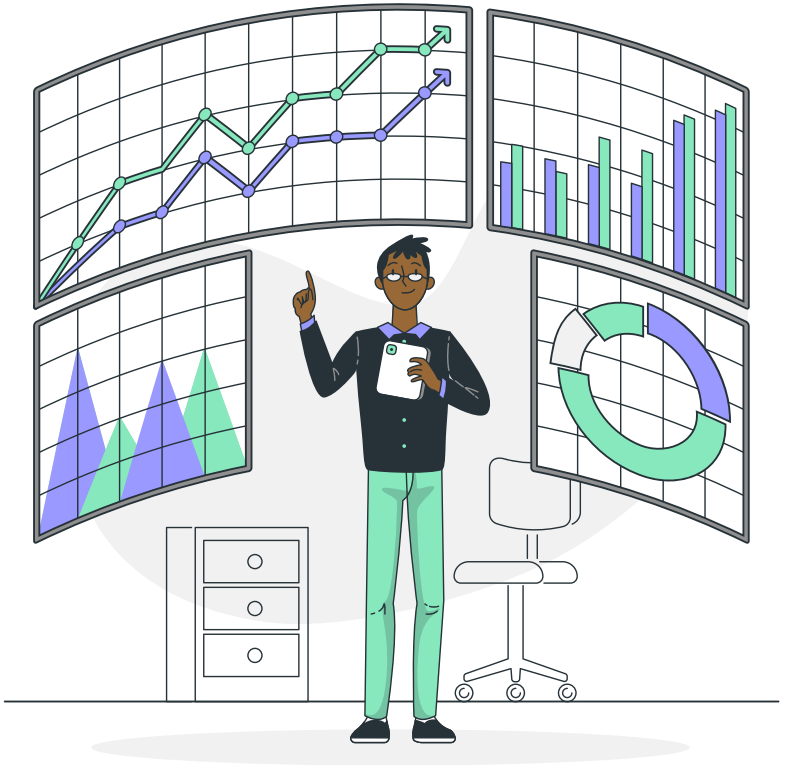


Key differences

Feature	C5.0	Random Forest
Ensemble Approach	Usually single decision tree	Ensemble of decision trees
Handling of Missing Data	Sensitive to missing values	Can handle
Overfitting Resistance	Using pruning mechanisms	Aggregation of diverse decision trees
Feature Selection	The most information gain	Random
Interpretability	Interpretable and intuitive	Lack of interpretability
Prediction Time	Relatively shorter	Relatively longer

Thanks!

Do you have any questions?



Resources

GeeksforGeeks, (2023) C5.0 Algorithm of Decision Tree,, https://www.geeksforgeeks.org/c5-0-algorithm-of-decision-tree/

IBM Corporation, (2024), C5.0 node https://www.ibm.com/docs/en/cloud-paks/cp-data/5.0.x?topic=modeling-c50-node

GeeksforGeeks. (2024). Random Forest Algorithm in Machine Learning. https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/

GeeksforGeeks. (2024). GeeksforGeeks. (2024). What are the Advantages and Disadvantages of Random Forest? https://www.geeksforgeeks.org/what-are-the-advantages-and-disadvantages-of-random-forest/

Delgado-Gallegos, J. L., Avilés-Rodriguez, G., Padilla-Rivas, G. R., De Los Ángeles Cosío-León, M., Franco-Villareal, H., Nieto-Hipólito, J. I., de Dios Sánchez López, J., Zuñiga-Violante, E., Islas, J. F., & Romo-Cardenas, G. S. (2023). Application of C5.0 Algorithm for the Assessment of Perceived Stress in Healthcare Professionals Attending COVID-19. *Brain sciences*, *13*(3), 513. https://doi.org/10.3390/brainsci13030513

IBM, (2024). What is a decision tree? https://www.ibm.com/topics/decision-trees

What is Decision Tree? [A Step-by-Step Guide]

Analytics Vidha, (2024) https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/#h-decision-tree-assumptions.