




Welcome to the CoGrammar

Tutorial: Decision Trees and Random Trees

The session will start shortly...

Questions? Drop them in the chat. We'll have dedicated moderators answering questions.



Data Science Session Housekeeping

- The use of disrespectful language is prohibited in the questions, this is a supportive, learning environment for all - please engage accordingly.
(Fundamental British Values: Mutual Respect and Tolerance)
- No question is daft or silly - **ask them!**
- There are **Q&A sessions** midway and at the end of the session, should you wish to ask any follow-up questions. Moderators are going to be answering questions as the session progresses as well.
- If you have any questions outside of this lecture, or that are not answered during this lecture, please do submit these for upcoming Academic Sessions. You can submit these questions here: [Questions](#)

Data Science Session Housekeeping cont.

- For all **non-academic questions**, please submit a query: www.hyperiondev.com/support
- Report a **safeguarding** incident: www.hyperiondev.com/safeguardreporting
- We would love your **feedback** on lectures: [Feedback on Lectures](#)

Skills Bootcamp

8-Week Progression Overview

Fulfil 4 Criteria to Graduation

✓ Criterion 1: Initial Requirements

Timeframe: First 2 Weeks

Guided Learning Hours (GLH):

Minimum of 15 hours

Task Completion: First four tasks

Due Date: 24 March 2024

✓ Criterion 2: Mid-Course Progress

60 Guided Learning Hours

Data Science - **13 tasks**

Software Engineering - **13 tasks**

Web Development - **13 tasks**

Due Date: 28 April 2024

Skills Bootcamp Progression Overview

✓ Criterion 3: Course Progress

Completion: All mandatory tasks,
including Build Your Brand and
resubmissions by study period end
Interview Invitation: Within 4 weeks
post-course
Guided Learning Hours: Minimum of
112 hours by support end date
(10.5 hours average, each week)

✓ Criterion 4: Demonstrating Employability

Final Job or Apprenticeship
Outcome: Document within 12
weeks post-graduation
Relevance: Progression to
employment or related
opportunity

**SKILLS
FOR LIFE**

SKILLS BOOTCAMPS



Department
for Education

CoGrammar

Tutorial: Decision Trees and Random Trees

May 2024

Learning objectives

- ❖ Recap: Understand and implement **Decision Trees** and **Random Trees** models
- ❖ Using Python **scikit-learn** library for **regression** and **classification** tasks

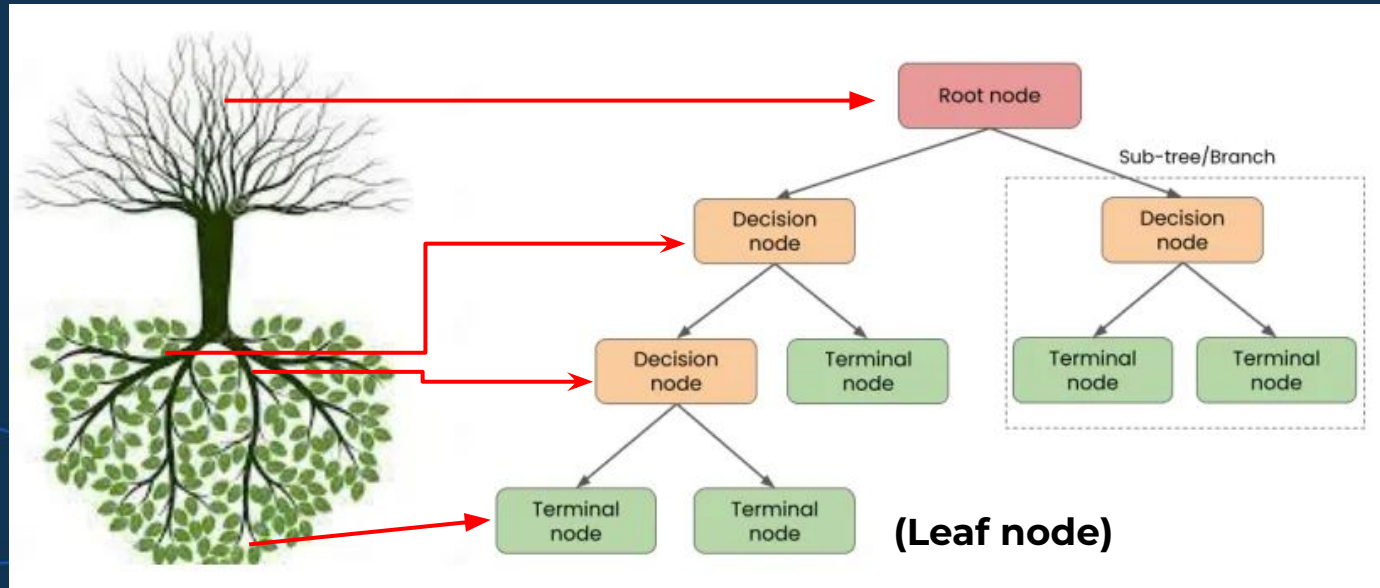
Decision Trees

Recap



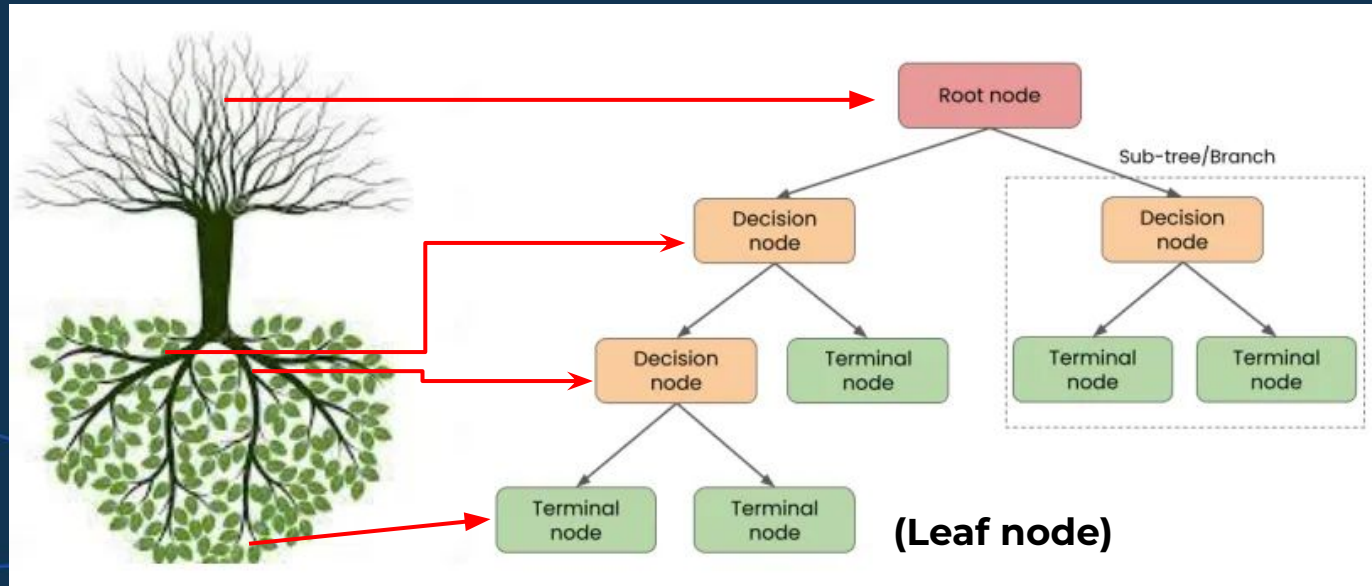
Decision Trees

- ❖ **Supervised** learning algorithm for **regression** and **classification**, **hierarchical**, **non-parametric**.
- ❖ Root node split into decision nodes (different decision points), branches (possible outcomes), leaf nodes (final class labels/predictions) cannot be further split.



Decision Trees

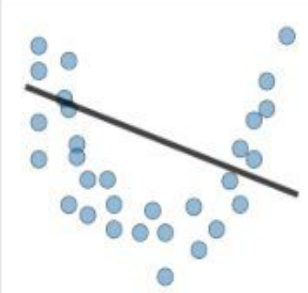

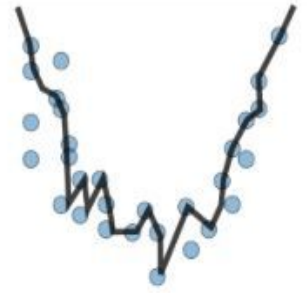
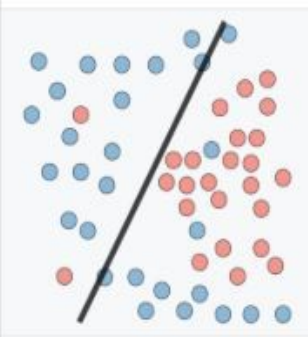
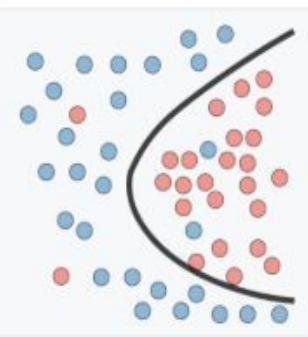
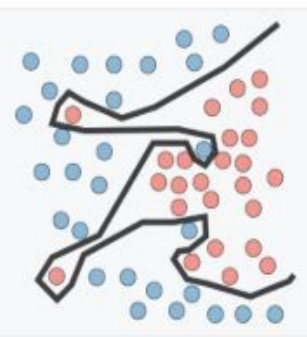
- ❖ **Splitting criteria:** Lower Gini impurity (classification) and residual reduction (regression)
- ❖ **Pruning:** done to prevent overfitting of the data, removes the nodes that contribute little to the model accuracy.



Overfitting and Underfitting

Bias-Variance Tradeoff

Finding the sweet-spot or the balance between underfitting and overfitting

	Underfitting	Just right	Overfitting
Symptoms	<ul style="list-style-type: none">• High training error• Training error close to test error• High bias	<ul style="list-style-type: none">• Training error slightly lower than test error	<ul style="list-style-type: none">• Very low training error• Training error much lower than test error• High variance
Regression illustration			
Classification illustration			
Possible remedies	<ul style="list-style-type: none">• Complexify model• Add more features• Train longer		<ul style="list-style-type: none">• Perform regularization• Get more data

Pre- and Post-Pruning

Pre-pruning

- ❖ Tunes hyperparameters (**max_depth**, **min_samples_leaf**, **min_samples_split**) prior to the training pipeline, get a robust model.
- ❖ **'Early stopping'** - **stops** the growth of the **decision tree** to reach its **full depth**, avoid producing leaves with small samples.
- ❖ Cross-validation error monitored at each step, if it is constant, stops growth.

Post-pruning

- ❖ **Decision Tree** model grows to its **full depth**, **tree branches** are **removed** to prevent the model from overfitting.
- ❖ Tune hyperparameter (**ccp_alpha** - **Cost Complexity Pruning**) to control the size of a tree, higher value leads to an increase in the number of nodes pruned.





Choose disadvantages of decision trees among the following.

1. Robust to outliers.
2. Factor analysis.
3. Prone to overfitting.
4. All of the above.




Choose disadvantages of decision trees among the following.

1. Robust to outliers.
2. Factor analysis.
- 3. Prone to overfitting.**
4. All of the above.



Which of the following statements is not true about the Decision tree?

1. Starts with a tree with a single leaf and assign this leaf a label according to a majority vote among all labels over the training set.
 2. Performs a series of iterations and on each iteration, it examine the effect of splitting a single leaf.
 3. Defines some gain measure that quantifies improvement due to the split.
 4. Among all possible splits, it either choose the one that maximises the gain and perform it, or choose not to split the leaf at all.
- 

Which of the following statements is not true about the Decision tree?

1. Starts with a tree with a single leaf and assign this leaf a label according to a majority vote among all labels over the training set.
2. Performs a series of iterations and on each iteration, it examine the effect of splitting a single leaf.
3. Defines some gain measure that quantifies improvement due to the split.
4. **Among all possible splits, it either choose the one that maximises the gain and perform it, or choose not to split the leaf at all.**

maximise the gain

What does the pruning method do?

1. Reduces the size of the tree.
2. Remove parts of the tree that do not provide power to classify instances.
3. Reduce the likelihood of overfitting.
4. All of the above.



What does the pruning method do?


1. Reduces the size of the tree.
2. Remove parts of the tree that do not provide power to classify instances.
3. Reduce the likelihood of overfitting.
4. **All of the above.**



Which parameters have a crucial effect on accuracy of Decision Tree Classifier? Select all that apply.

1. max_features
2. criterion
3. min_samples_leaf
4. max_depth





Which parameters have a crucial effect on accuracy of Decision Tree Classifier? Select all that apply.

- 1. **max_features**
- 2. criterion
- 3. **min_samples_leaf**
- 4. **max_depth**



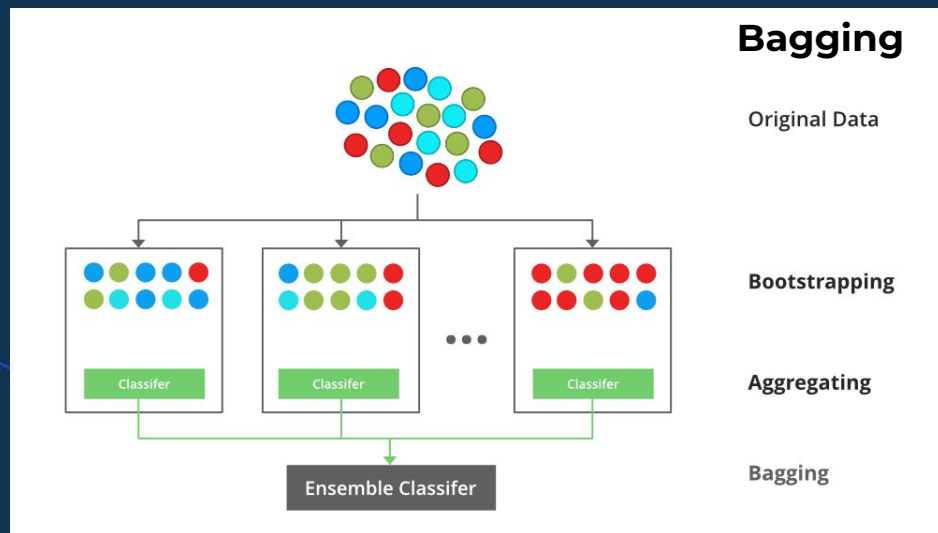
Random Trees

Recap



Random Trees

Ensemble methods aggregate the predictions of multiple classifiers / regressors into a single, improved prediction, mitigates overfitting.

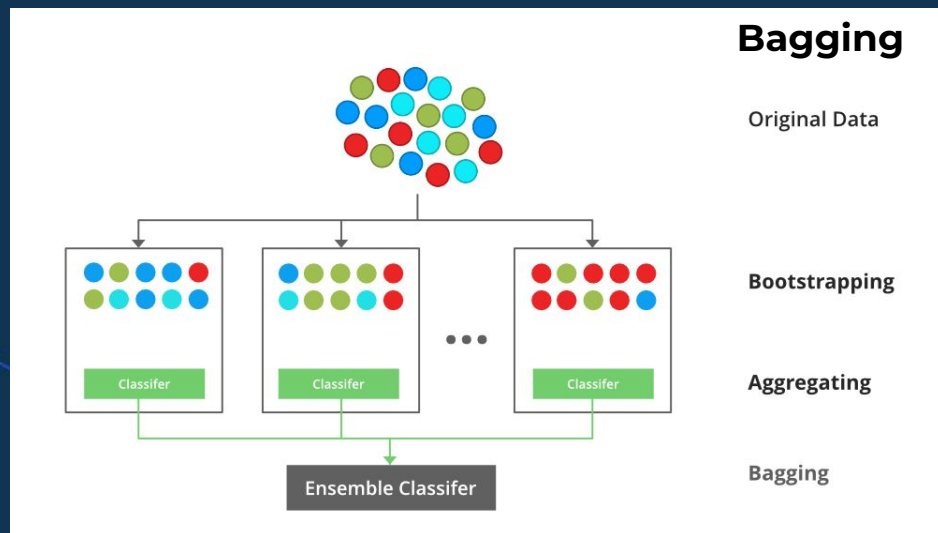


❖ **Bootstrapping:** Ensemble methods identify different patterns, reduces noise, and retain only the best patterns. Samples (size n) are drawn from a dataset (size N , $n < N$) repeatedly, with **replacement** (an instance can occur in more than one sample) -> **Bootstrap data set**.

❖ **Bagging (Bootstrap aggregation):** predictions of models fitted to sample are combined (majority voting/averaging) into a final prediction.

Random Trees

Ensemble methods aggregate the predictions of multiple classifiers / regressors into a single, improved prediction, mitigates overfitting.



- ❖ **Regular bagged trees:** many decision trees are built using bootstrapped training samples.
- ❖ **Random forests** provide an improvement over bagged trees by **decorrelating** the trees. Averaging many uncorrelated quantities leads to a **larger reduction in variance**. average of the resulting trees **less variable** and hence **more reliable**.

Decision Trees and Random Forests

Feature	Decision Tree	Random Forest
Randomness	No	Yes
Overfitting	More prone	Less prone
Bias-variance trade-off	Worse	Better
Interpretability	Easy	Challenging
Scalability	Less scalable	More scalable
Usage	Cases where transparency and interpretability are important	Cases with many features and where overfitting reduction are important

Decision Tree or Random Forest?

Problem: We want to predict whether a person will buy a new car based on their age and income. We want to train a machine-learning model with a dataset comprising 1000 entries.


- ❖ **Decision Tree:** model might split the data based on age first and then the income, resulting in a tree with several levels of splits. The resulting tree might be difficult to interpret, and there is a high chance of overfitting.
- ❖ **Random Forest:** model would create multiple decision trees, each trained on a randomly sampled subset of the data. Each tree splits the data differently, resulting in a collection of diverse trees. The model combines predictions of these trees to make final prediction. This results in better accuracy and less overfitting than using a single Decision Tree.



Decision Tree or Random Forest?

Problem: Predicting customer churn

A telecommunications company wants to predict which customers are likely to churn or cancel their service. The algorithm can take into account various features such as customer demographics, usage patterns, and customer service interactions.

- ❖ A Random Forest model would be better suited in this scenario due to its ability to handle a large number of features and reduce overfitting.
- 

Decision Tree or Random Forest?

Problem: Medical diagnosis

Model to diagnose medical conditions such as breast cancer, diabetes, or heart disease. The algorithm can take into account patient demographics, medical history, and laboratory test results. In this case,

- ❖ Decision Tree may be more suitable due to its ability to provide a transparent and interpretable model, which is important for medical practitioners.



Decision Tree or Random Forest?

Problem: Predicting stock prices

Random Forest or Decision Tree can be used to predict stock prices. The algorithm can take into account various financial indicators such as company earnings, dividends, and economic data.



- ❖ A Random Forest model would be better suited in this scenario due to its ability to handle noisy data and reduce overfitting.

Decision Tree or Random Forest?


Problem: Credit risk analysis



A bank can use a Random Forest or Decision Tree to assess the creditworthiness of loan applicants. The algorithm can take into account various factors such as credit history, income, and employment status.

- ❖ In this case, Random Forest may be more suitable due to its ability to handle a large number of features and reduce overfitting.




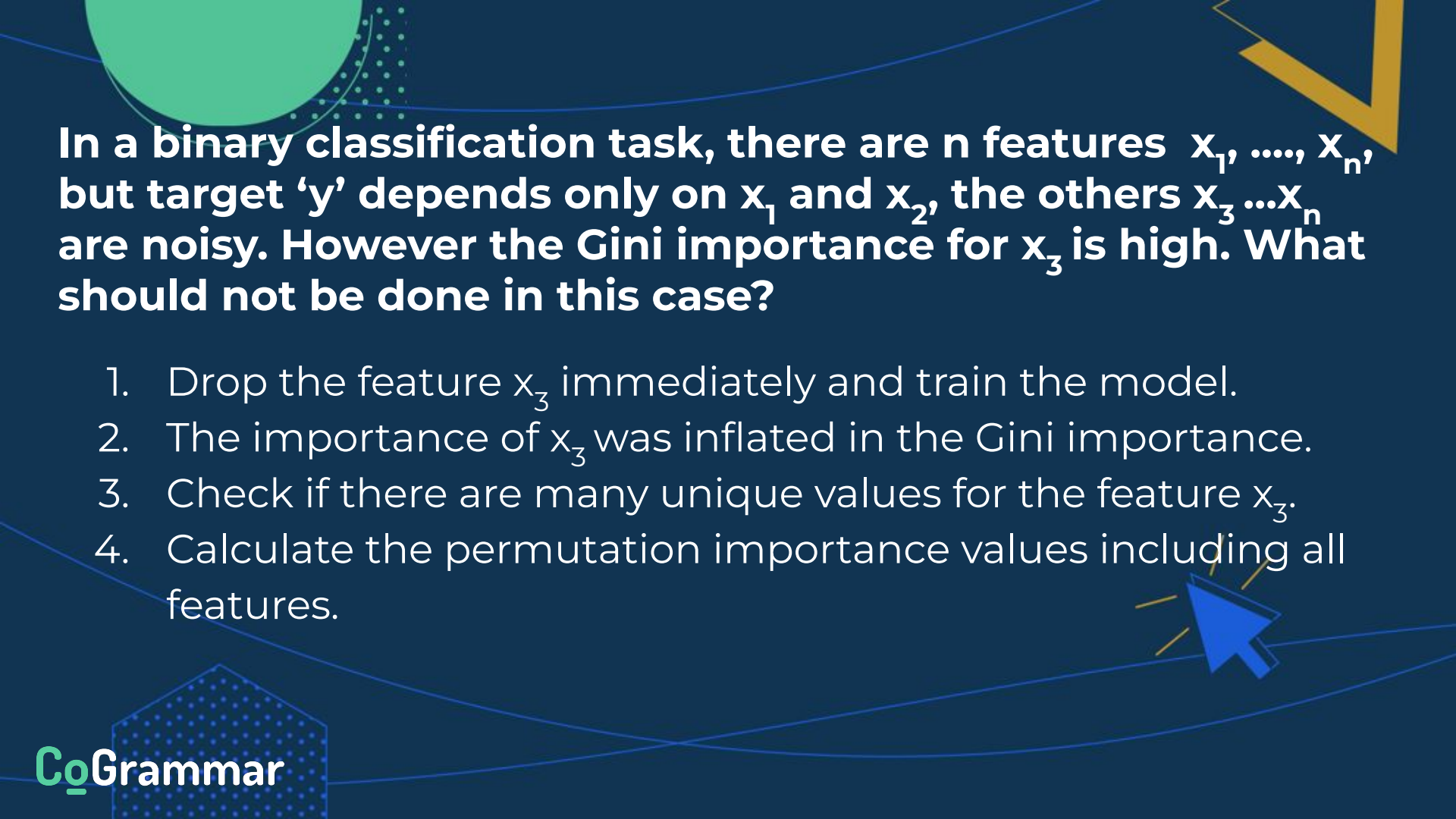
Which of the following statements is not true about Random forests?

1. It is a classifier consisting of a collection of decision trees
 2. Each tree is constructed by applying an algorithm on the training set and an additional random vector
 3. Each individual tree in the random forest does not have a class prediction
 4. The prediction of the random forest is obtained by a majority vote over the predictions of the individual trees
- 





Which of the following statements is not true about Random forests?

1. It is a classifier consisting of a collection of decision trees
 2. Each tree is constructed by applying an algorithm on the training set and an additional random vector
 - 3. Each individual tree in the random forest does not have a class prediction**
 4. The prediction of the random forest is obtained by a majority vote over the predictions of the individual trees
- 




In a binary classification task, there are n features x_1, \dots, x_n , but target 'y' depends only on x_1 and x_2 , the others $x_3 \dots x_n$ are noisy. However the Gini importance for x_3 is high. What should not be done in this case?

1. Drop the feature x_3 immediately and train the model.
 2. The importance of x_3 was inflated in the Gini importance.
 3. Check if there are many unique values for the feature x_3 .
 4. Calculate the permutation importance values including all features.
- 



In a binary classification task, there are n features x_1, \dots, x_n , but target 'y' depends only on x_1 and x_2 , the others $x_3 \dots x_n$ are noisy. However the Gini importance for x_3 is high. What should not be done in this case?

1. **Drop the feature x_3 immediately and train the model.**
 2. Check if importance of x_3 was inflated in the Gini importance.
 3. Check if there are many unique values for the feature x_3 .
 4. Calculate the permutation importance values including all features.
- 

Questions and Answers



Thank you for attending



Department
for Education

CoGrammar

