



Welcome to this session: Task Walkthrough - Tasks 17 - 21

The session will start shortly...

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



Safeguarding & Welfare

We are committed to all our students and staff feeling safe and happy; we want to make sure there is always someone you can turn to if you are worried about anything.

If you are feeling upset or unsafe, are worried about a friend, student or family member, or you feel like something isn't right, speak to our safeguarding team:



Ian Wyles
Designated Safeguarding
Lead



Simone Botes



Nurhaan Snyman



Rafiq Manan



Ronald Munodawafa



Tevin Pitts

Scan to report a
safeguarding concern



or email the Designated
Safeguarding Lead:
Ian Wyles

safeguarding@hyperiondev.com

Skills Bootcamp Data Science

- 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](#)

Skills Bootcamp Data Science

- 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](#)**
- If you are hearing impaired, please kindly use your computer's function through Google chrome to enable captions.

Learning Outcomes

- ❖ **Explain the key principles** of supervised learning and machine learning models.
- ❖ **Apply Linear Regression** to predict continuous outcomes.
- ❖ **Use Logistic Regression** for binary classification tasks.
- ❖ **Construct Decision Trees** for decision-based learning.
- ❖ **Utilize Random Forest Classifiers** to improve classification accuracy.
- ❖ **Evaluate model performance** using appropriate metrics.

Lecture Overview

- Presentation of the Task
- Machine Learning
- Linear Regression
- Logistic Regression
- Decision Trees
- Task Walkthrough



Task Walkthrough

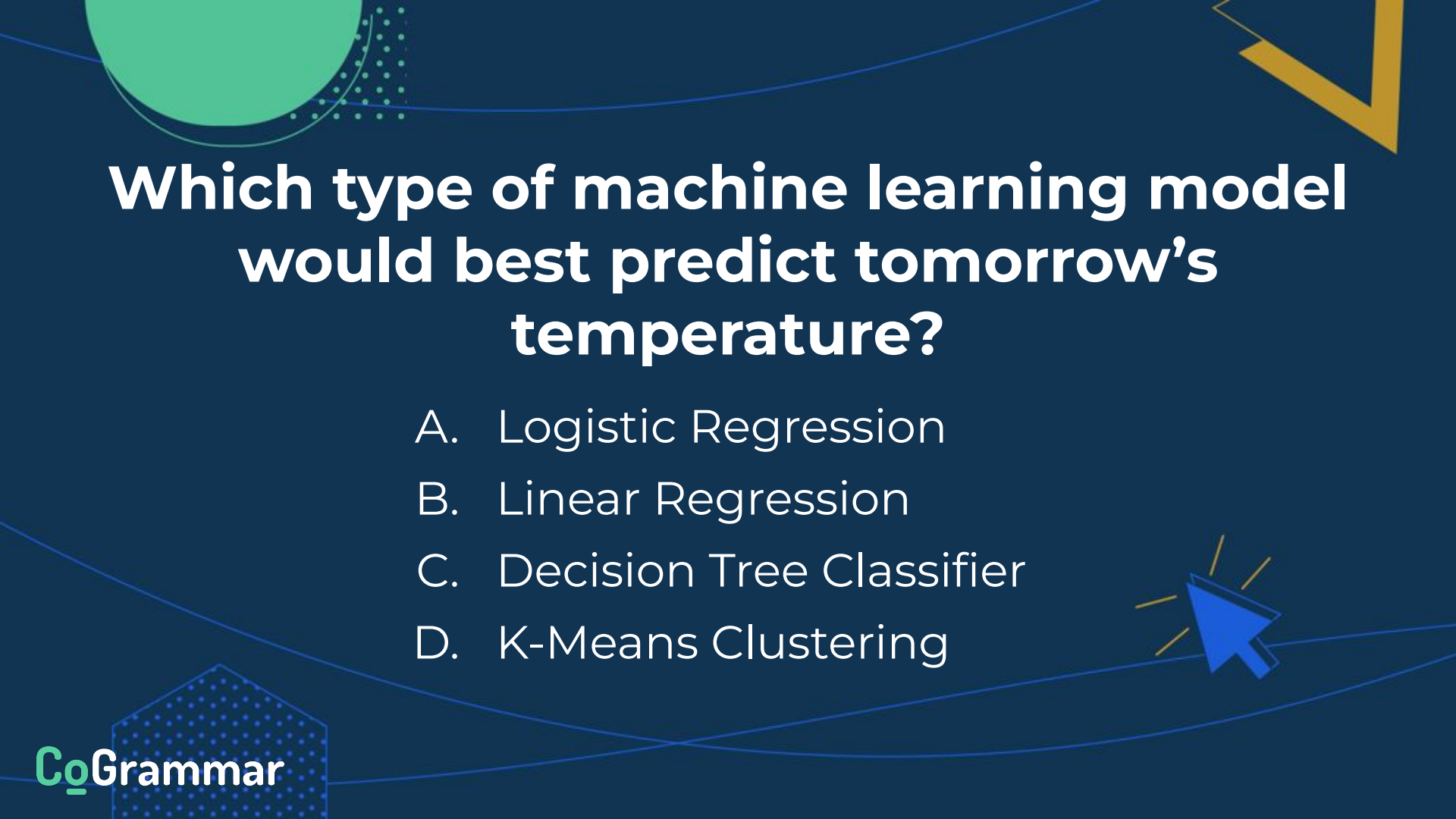
You work for a telecom company that wants to improve customer retention. Your team's goal is to predict whether a customer will churn (cancel their service) based on their usage and demographics. You will:

- ❖ Use **Linear Regression** to predict a **customer's monthly charges** based on their demographics and service plan.
- ❖ Use **Logistic Regression** to classify **whether a customer will churn or not (Yes/No)**.
- ❖ Use **Decision Trees and Random Forest Classifiers** to improve **classification accuracy**.



What is the key difference between supervised and unsupervised learning?

- A. Supervised learning requires labeled data, while unsupervised learning does not
- B. Supervised learning is faster
- C. Unsupervised learning uses decision trees
- D. Supervised learning can only classify text

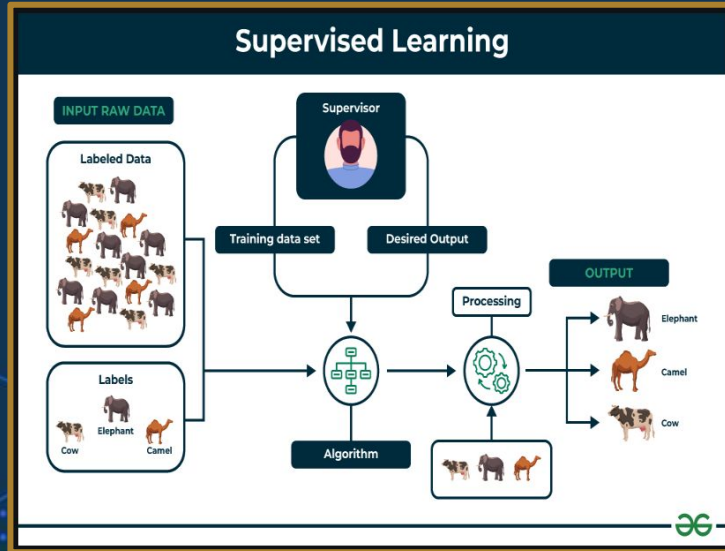


Which type of machine learning model would best predict tomorrow's temperature?

- A. Logistic Regression
- B. Linear Regression
- C. Decision Tree Classifier
- D. K-Means Clustering

Types of machine learning

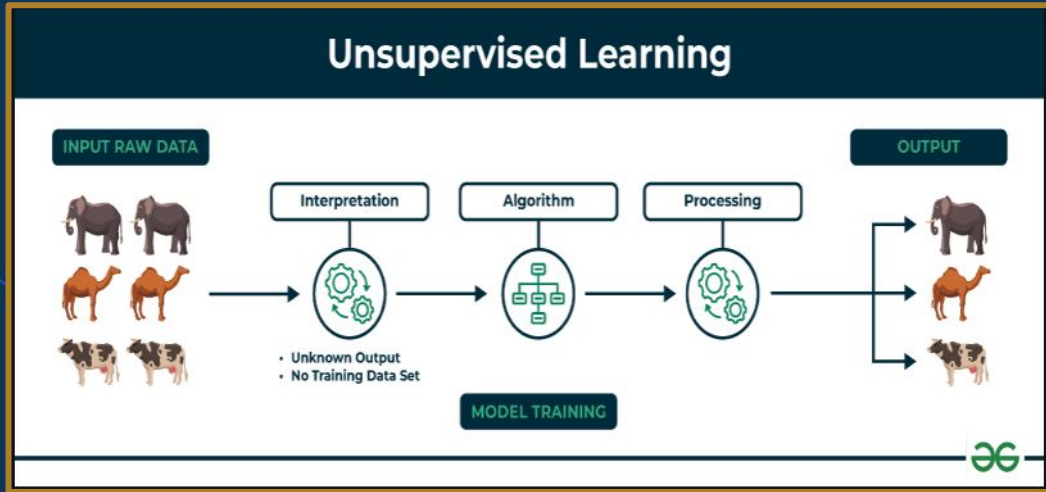
- ❖ **Supervised learning:** The computer learns from labelled data, where both input and output data are provided.



Source: [geeksforgeeks](https://www.geeksforgeeks.org/)

Types of machine learning

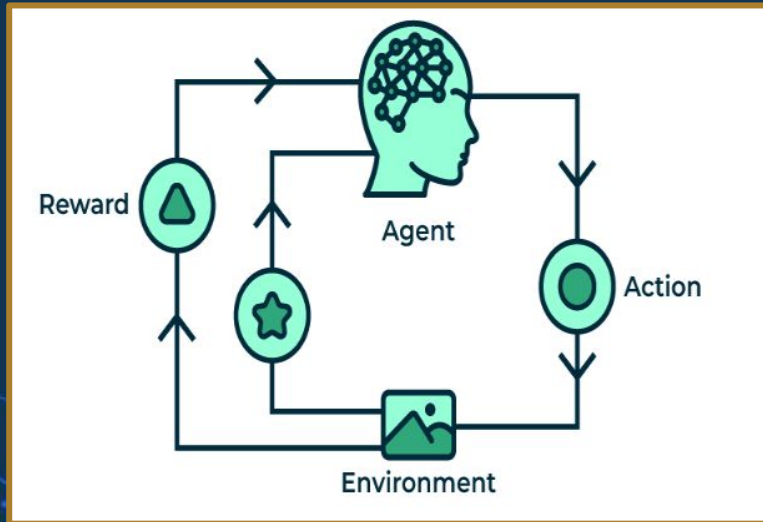
- ❖ **Unsupervised learning:** The computer learns from unlabeled data, discovering hidden patterns or structures on its own.



Source: [geeksforgeeks](https://www.geeksforgeeks.org/)

Types of machine learning

- ❖ **Reinforcement learning:** The computer learns through interaction with an environment, receiving rewards or penalties for its actions.



Source: [geeksforgeeks](https://www.geeksforgeeks.com/reinforcement-learning/)



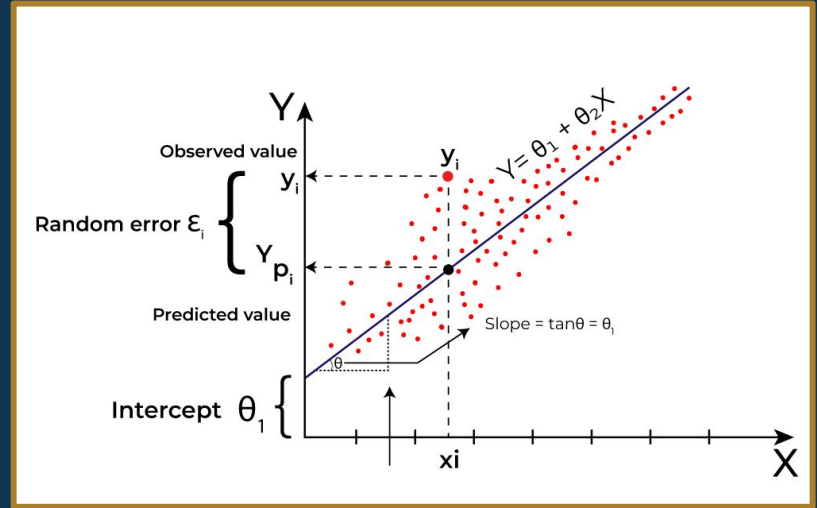
Types of Supervised Learning

- ❖ **Regression:** Predicting continuous numerical values, such as house prices or stock prices.
- ❖ **Classification:** Predicting discrete categories or classes, such as whether an email is spam or not.



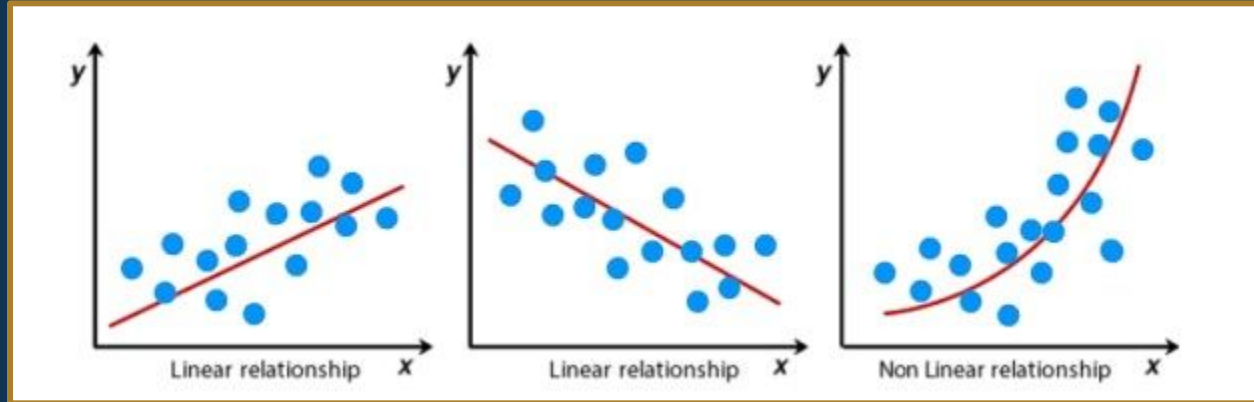
Simple Linear Regression

- ❖ Simple linear regression is a method to study the relationship between two variables: an independent variable (x) and a dependent variable (y).
- ❖ It helps us understand how changes in the independent variable affect the dependent variable.



Assumptions and Limitations of Simple Linear Regression

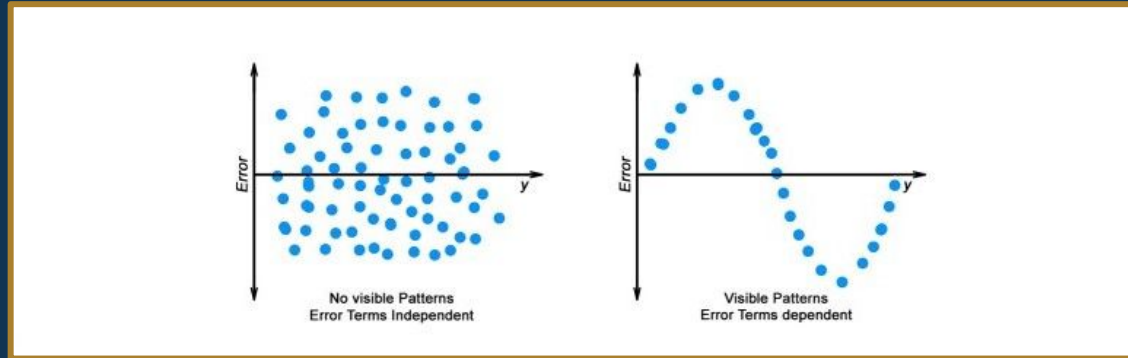
- ❖ **Linearity:** The relationship between x and y should be linear.



Source: [Analytics Vidhya](#)

Assumptions and Limitations of Simple Linear Regression

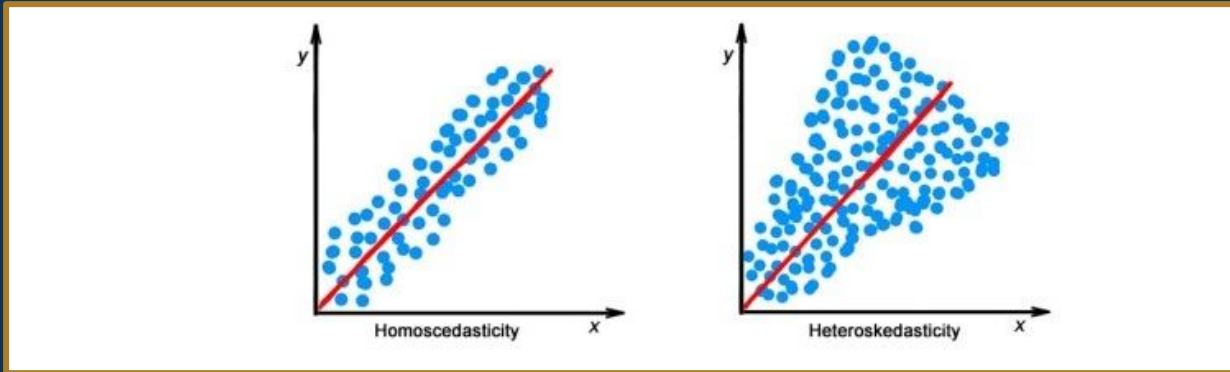
- ❖ **Independence:** The observations should be independent of each other.



Source: [Analytics Vidhya](#)

Assumptions and Limitations of Simple Linear Regression

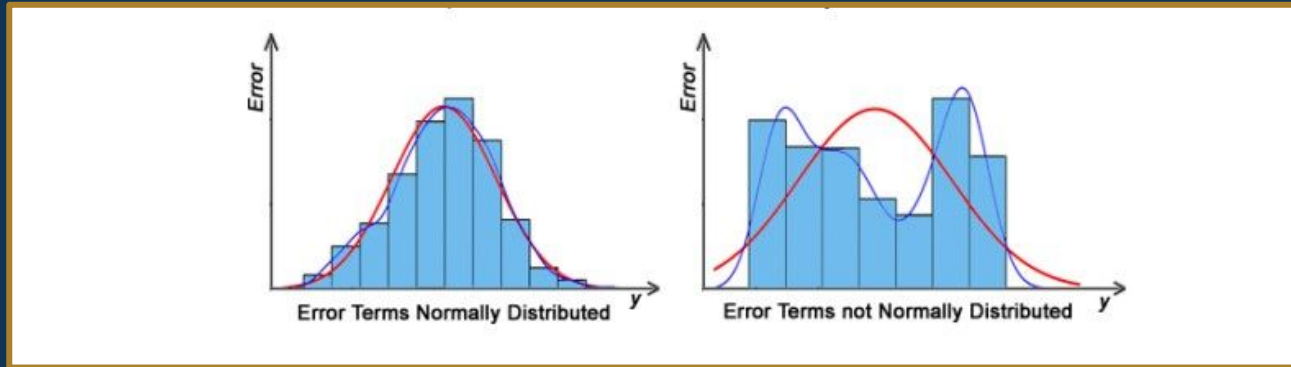
- ❖ **Homoscedasticity:** The variability of y should be constant across all values of x .



Source: [Analytics Vidhya](#)

Assumptions and Limitations of Simple Linear Regression

- ❖ **Normality:** The errors should be normally distributed.



Source: [Analytics Vidhya](#)

Evaluation Metrics

- ❖ Mean Squared Error (MSE):
 - MSE measures the average squared difference between the predicted and actual values.
 - A lower MSE indicates better model performance.
- ❖ R-squared (R^2) score:
 - R^2 represents the proportion of variance in the target variable that can be explained by the model.
 - An R^2 value closer to 1 indicates a better fit of the model to the data.

Evaluation Metrics

- ❖ **Accuracy** is another commonly used metric for evaluating the performance of a machine learning model, particularly in classification problems.
 - **Accuracy = (Number of correct predictions) / (Total number of predictions) * 100%**
- ❖ While accuracy is more suitable for classification tasks, metrics like Mean Squared Error (MSE) and R-squared (R^2) are used for regression problems.

Logistic Regression

- ❖ **Linear regression** models make **predictions** for the datasets for which dependent variables have **continuous numerical values**.
- ❖ **Logistic Regression**
 - **supervised learning** algorithm
 - **classification** algorithm
 - dependent variables are **distinct, non-continuous, categorical**
- ❖ **Classification** - predicting **probability** of **categorical variables** for a given observation and assigning the observation to the category with the highest probability.

Assumptions of Logistic Regression

- ❖ The **independent variables** should **not be correlated** with each other i.e. the model should have little or **no multicollinearity**.
- ❖ The **dependent variable** must be **categorical** in nature.
- ❖ The relationship between the **independent variables** and the **log odds** of the dependent variable should be **linear**.
- ❖ There should be **no outliers** in the dataset.
- ❖ The data sample size should be **sufficiently large**.

Accuracy

Accuracy of classifier: Total number of correct predictions by the classifier divided by the total number of predictions.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

For virus example, **Accuracy = 96%**

According to the **Accuracy** value, the model “can predict sick people 96% of the time”. However, it is **predicting the people who will not get sick with 96% accuracy while the sick are spreading the virus.**

Better to measure how many **positive cases we can predict correctly** to arrest spread of the contagious virus or **out of the correct predictions**, how many **are positive cases** to check the reliability of the model.

Precision and Recall

- ❖ **Precision:** tells us how many of the correctly predicted cases actually turned out to be positive, determine whether the model is reliable or not.
- ❖ **Recall:** how many of the actual positive cases we were able to predict correctly with our model.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

For virus example, Precision = 50%, Recall = 75%

For virus example, 50% percent of the correctly predicted cases turned out to be positive cases. Whereas 75% of the positives were successfully predicted by the model.

Precision and Recall

- ❖ **Precision**: useful in cases where **False Positive** is a greater concern.
- ❖ *Music or video recommendation systems, e-commerce websites.*
- ❖ *Wrong results could lead to customer churn and be harmful to the business.*

- ❖ **Recall**: useful in cases where **False Negative** trumps.
- ❖ *Medical cases where it does not matter whether a false alarm flag is raised, but the actual positive cases should not go undetected.*

For **contagious virus example**, the **Confusion Matrix** is more insightful measure in such critical scenarios.

Recall, assessing the ability to capture all actual positives, emerges as a **better metric**. **Accuracy** proves **inadequate** as a metric for the model's evaluation.

Avoid mistakenly releasing an infected person into the healthy population, potentially spreading the virus.

F1-score

- ❖ Cases where there is no clear distinction between whether Precision is more important or Recall.
- ❖ **F1-score**: harmonic mean of **Precision** and **Recall**, gives a combined idea about these two metrics, appropriate metric for imbalanced dataset.
- ❖ **Maximum** when **Precision** is **equal** to **Recall**.
- ❖ Use in combination with other evaluation metrics.

$$F_1 = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

Classification Trees

Decision tree models where the **target variable** uses a **discrete set of values**, **classification** problems, determine whether an event happened or didn't happen, involving a “yes” or “no” outcome. Each **node**, or **leaf**, represent **class labels** while **branches** represent conjunctions of **features** leading to class labels.

- ❖ The **root node (Outlook)** has two or more **decision nodes (Sunny, Overcast and Rainy)** with other **predictors (Windy, Humidity)**.
- ❖ The **leaf node (Play golf)** is the **target**, and represents a classification of decision.

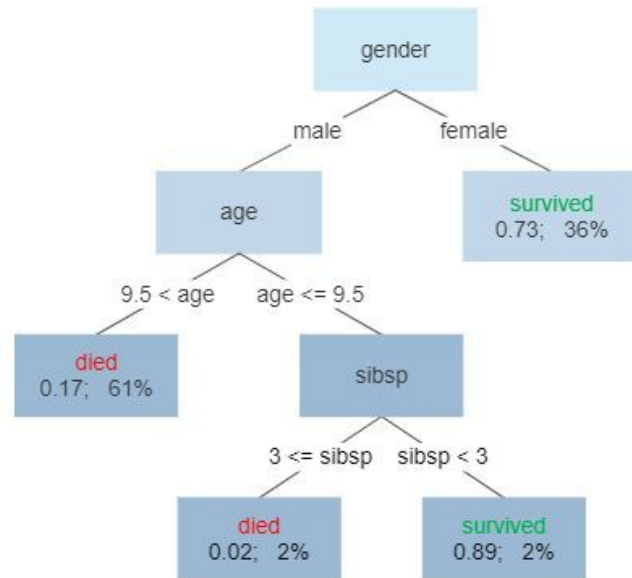


Regression Trees

Decision trees which **predict continuous values** as **targets** based on previous data or information sources. **Predicts** what is likely to happen, given previous behavior/trends.

- ❖ Survival of passengers on the Titanic. Figures under the leaves show the **probability of survival** and the **percentage of observations** in the leaf.
- ❖ "sibsp" is the number of spouses or siblings aboard.

Survival of passengers on the Titanic



CART algorithm

Classification and Regression Trees (CART) algorithm

- ❖ **Tree structure:** CART builds a tree-like structure with nodes and branches.
- ❖ **Nodes:** represent different decision points.
- ❖ **Branches:** represent possible outcomes.
- ❖ **Leaf nodes:** contain a predicted class label or value for the target variable.
- ❖ **Splitting criteria:** CART evaluates all possible splits and selects the one that best reduces the impurity of the resulting subsets.
 - **Gini impurity** (for **classification**, lower means purer subset) and **residual reduction** (for **regression**, lower means better model's fit to the data).
- ❖ **Pruning:** done to prevent overfitting of the data, removes the nodes that contribute little to the model accuracy.

Ensemble Methods

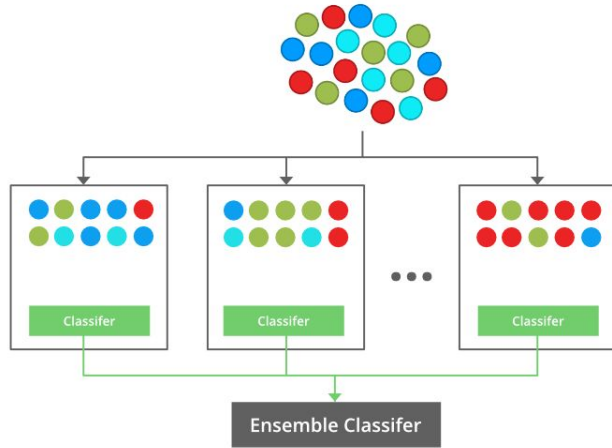
- ❖ **Decision Trees** are easy to understand, apply, interpret and visualise. However, they are **not very robust**, **small perturbations in the training data** could give rise to **substantially different predictions** at test time.
- ❖ Predictions of decision trees have very **high variance**. Ideally, we'd like our models to capture general patterns, not to be **so dependent on the data** they have trained on that a bit of noise or a different sample changes predictions entirely.

Ensemble Methods

- ❖ **Ensemble techniques** work like a group of diverse experts teaming up to make decisions, and create a more robust solution than any individual could achieve alone.
- ❖ Ensemble methods **aggregate the predictions of multiple classifiers/regressors** into a single, improved prediction.
- ❖ Aside from **Random Forests**, **ensemble methods** can and do get applied to methods **other than decision trees**, but trees can benefit in particular due to how flexible they are.

Ensemble Methods

Bagging



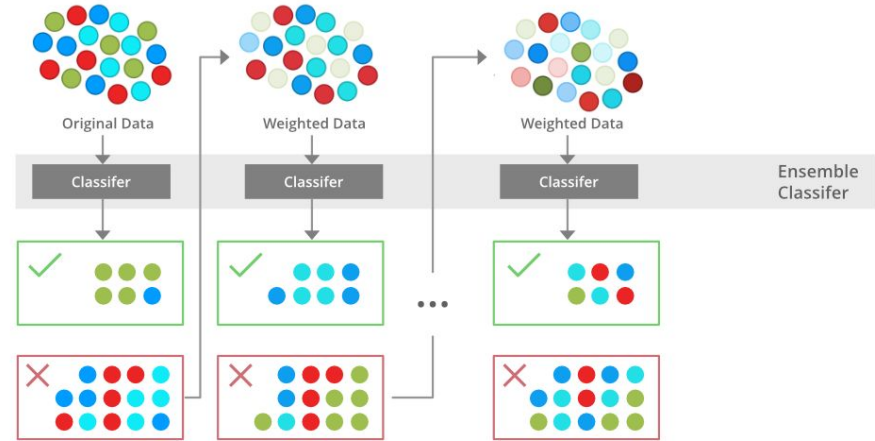
Original Data

Bootstrapping

Aggregating

Bagging

Boosting



Original Data

Weighted Data

Weighted Data

Classifier

Classifier

Classifier

Ensemble Classifier



Random Forests

Gradient Boosting, XGBoost, AdaBoost

Random Forests

- ❖ **Random Forests:** created from **many decision trees** during training phase using **random subset of the dataset** to measure a **random subset of features** in each partition.
- ❖ Randomness introduces **variability** among individual trees, **reducing** the risk of **overfitting** and **improving** overall **prediction performance**.
- ❖ Aggregates predictions of all trees, either by **voting** (**classification problems**) or by **averaging** (**regression problems**).
- ❖ **Collaborative decision-making process**, supported by **multiple trees** with their insights, gives **stable** and **precise results**.

Key Features of Random Forests

- ❖ **Diversity:** Not all attributes/variables/features are considered while making an individual tree; each tree is different.
- ❖ **Dimensionality reduction:** Feature space is reduced.
- ❖ **Parallelization:** Each tree is created independently out of different data and attributes, fully use the CPU to build random forests.
- ❖ **Train-Test split:** No train and test data splitting required as there is always 30% of data which is not seen by the decision tree.
- ❖ **Stability:** Stable as final results are based on majority voting/ averaging.

Differences with Decision Trees

Decision Trees	Random Forests
Can suffer from overfitting if allowed to grow without any control.	Created from subsets of data, and final output is based on average or majority ranking; overfitting is mitigated . Better bias-variance trade-off .
When a data set with features is taken as input by a decision tree, some rules formulated to make predictions .	Randomly selects observations, builds a decision tree, and takes the average result. Does not use any set of rules.
A single decision tree is faster in computation.	Comparatively slower .

Task Walkthrough

You work for a telecom company that wants to improve customer retention. Your team's goal is to predict whether a customer will churn (cancel their service) based on their usage and demographics. You will:

- ❖ Use **Linear Regression** to predict a **customer's monthly charges** based on their demographics and service plan.
- ❖ Use **Logistic Regression** to classify **whether a customer will churn or not (Yes/No)**.
- ❖ Use **Decision Trees and Random Forest Classifiers** to improve **classification accuracy**.



What does R^2 score measure in Linear Regression?

- A. The percentage of variance explained by the model
- B. The number of features in the dataset
- C. The classification accuracy
- D. The probability of an event occurring



What is the main advantage of a Random Forest Classifier?

- A. It is always more accurate than any other model
- B. It reduces overfitting by combining multiple decision trees
- C. It only works for numerical data
- D. It does not require training data

Summary

- ★ **Linear Regression** predicts continuous values.
- ★ **Logistic Regression** classifies binary outcomes.
- ★ **Decision Trees** handle both classification and regression with clear decision rules.
- ★ **Random Forest Classifiers** improve accuracy by reducing overfitting.

CoGrammar

Q & A SECTION

**Please use this time to ask
any questions relating to the
topic, should you have any.**

Thank you for attending



CoGrammar



Department
for Education