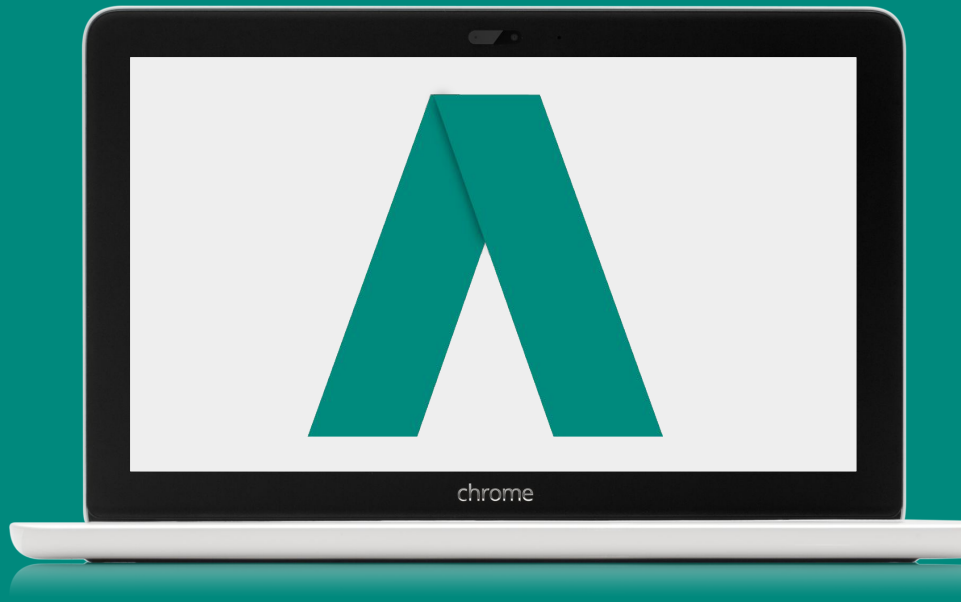
Machine Learning: Classification

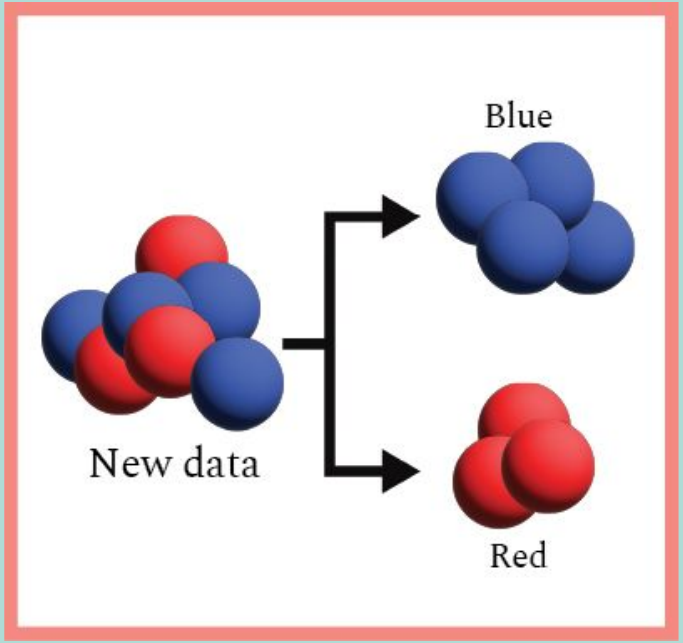


Classification

Classification involves categorizing or labeling data into predefined categories based on their features.

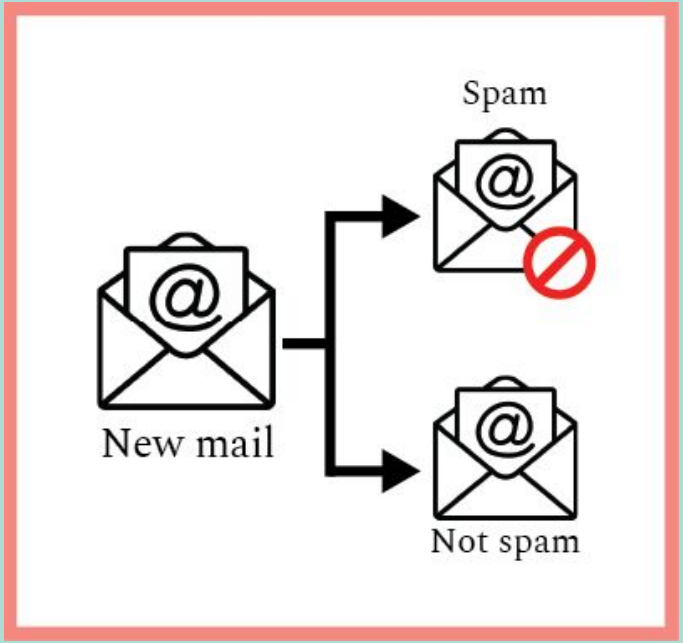
It is a supervised learning task that involves analyzing the connection between input features and predefined output labels.

In other words, the algorithm learns to classify new, unseen data points into the appropriate classes it has been trained on.

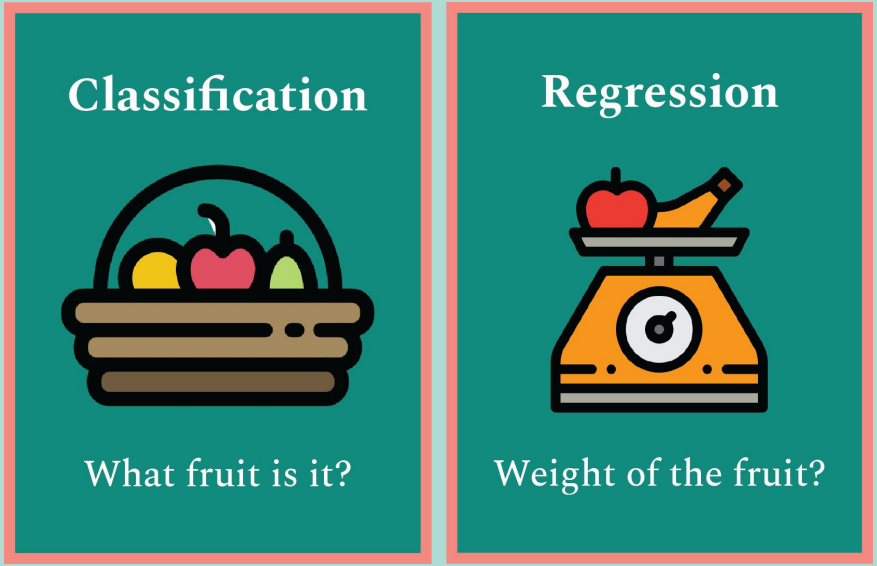


Simple Classification Problem.

Classifying mail as spam or not.



Difference between Classification and Regression



Imagine you have a basket of fruits, and you want to sort them.

Classification is like separating apples from oranges.

Regression is like guessing how much each fruit weighs.

In technical terms, classification deals with predicting a category (like apples or oranges), while regression deals with predicting a continuous value (like weight).

Use case in Finance

Fraud Detection: Determine if a credit card transaction is fraudulent or legitimate based on transaction details.
Classes: Fraudulent or Legitimate

Credit Scoring: Predicting the creditworthiness of loan applicants based on their financial history
Classes: Approved, Not Approved

Stock Price: Predicting stock price movements based on historical data and indicators
Classes: Increase, Decrease or No Change

Use case in Climate Science

Weather Pattern: Classifying different types of clouds or weather conditions from satellite images.
Classes: Clear Sky, Cloudy, Rainy, Snowy

Species Distribution: Predicting the distribution of plant and animal species based on environmental factors
Classes: Suitable Habitat, Unsuitable Habitat

Extreme Weather: Identifying the likelihood of extreme weather events like hurricanes or storms.
Classes: Storm, hurricanes, drought

Use case in Healthcare

Disease Diagnosis: Diagnosing diseases or medical conditions based on patient's medical test results
Classes: Healthy, Disease A, Disease B, etc

Medical Imaging: Classifying medical images (e.g., X-rays, MRI scans) to identify abnormalities, such as tumors.
Classes: Normal, Abnormal

Drug Response: Predicting how patients will respond to specific medications based on genetic information and medical history.
Classes: Positive, Negative or No response

Use case in Marketing

Customer Segmentation: Segmenting customers into groups based on purchasing behavior to tailor marketing campaigns.
Classes: High, Medium or Low spending

Churn Prediction: Predicting which customers are likely to churn (stop using a service) based on usage patterns.
Classes: Churn, No Churn

Sentiment Analysis: Analyzing customer reviews to understand sentiment and feedback about products or services.
Classes: Positive, negative, or neutral sentiment

Use case in Education

Student Performance: Predicting student performance based on factors such as attendance, or previous grades
Classes: High, Average, or Low Performer

Learning Style Identification: Identifying the preferred learning styles of students to personalize educational content.
Classes: Visual, Auditory, Kinesthetic Learner

Automated Essay Grading: Classifying and grading student essays based on content, structure, and language.
Classes: Excellent, Good, Fair, or Poor

Use case in Sales

Lead Scoring: Predicting the likelihood of a lead converting into a customer based on their interactions with a company's website.
Classes: High or Low Likelihood

Customer Lifetime Value: Estimating the potential value of a customer over their entire relationship with a business.
Classes: High, Medium, or Low value

Product Recommendation: Recommending products to customers based on their purchase history and browsing behavior.
Classes: Recommended, Not Recommended

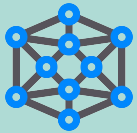
Steps to tackle a classification problem.

1. Collect data

Gather information about the problem you want to solve (e.g., wine type or wine quality).

2. Prepare data

Clean and organize the data, removing any errors or missing values.



3. Choose a model

Pick a regression model that suits your problem (e.g., Linear Regression, Logistic Regression).

Steps to tackle a classification problem.

4. Train the model

Teach your model using the data you've collected. It's like giving the model a set of practice questions.

5. Evaluate the model

Test your model on new data to see how well it predicts ice cream sales.

6. Fine-tune the model

If needed, adjust the model to make better predictions.

Steps to tackle a classification problem.

7. Make predictions.

Use your trained model to predict whether the weather is going to be clear, sunny or rainy.

Models (Algorithms) for classification.

There are numerous regression algorithms available for machine learning, each with its strengths and weaknesses.

Here is a list of some popular regression algorithms. We have highlighted the most relevant for this project:

- Logistic Regression
- Decision Trees
- Random Forest
- Support Vector Machines (SVM)
- K-Nearest Neighbors (KNN)
- Naive Bayes
- Gradient Boosting
- Neural Networks

[Top Machine Algorithms by DataCamp]

Strengths and Weaknesses

Logistic Regression

Strengths:

- Easy to understand and explain.
- Works well when you want to know the chances of something happening.
- Can handle big datasets without using too much computer power.

Weaknesses:

- Can only use straight lines, not good for complex patterns.
- Assumes that the relationship between features is simple.
- Gets easily affected by weird data points that don't fit the pattern.

Strengths and Weaknesses

Decision Trees

Strengths:

- Can learn tricky patterns in data, even if they're not straight lines.
- Can tell you which features are the most important for making decisions.
- Can handle big datasets without using too much computer power.

Weaknesses:

- Sometimes makes decisions that don't fit the bigger picture, leading to overfitting.
- Can change a lot if you just change a tiny bit of data, which makes them unstable.
- Can't capture all kinds of relationships, especially if they're very complicated.

Strengths and Weaknesses

Random Forest

Strengths:

- Combines many decision trees to give better answers and avoid overfitting.
- Gives more reliable results by averaging out the predictions from different trees.
- Can help you know which features are important in making decisions.

Weaknesses:

- Harder to understand than a single decision tree because it's a group of trees.
- Takes longer to train and make predictions compared to just one decision tree.
- Needs a lot of computer memory, especially with many trees or features.

A few evaluation metrics.

Evaluation metrics for classification in machine learning are used to measure the accuracy of a model's predictions compared to the actual values. The list below is not an exhaustive list.

Accuracy

Measures the proportion of correctly classified instances among all instances. It's a basic metric but can be misleading when classes are imbalanced.

Precision

Focuses on how many of the instances predicted as positive are actually positive. Helps avoid false positives.

Recall

Measures how many of the actual positive instances were correctly predicted as positive. Helps avoid false negatives.

F1-Score

Combines precision and recall into a single metric that balances their trade-off.

Metrics to Evaluate your Classification Model to take the right decisions