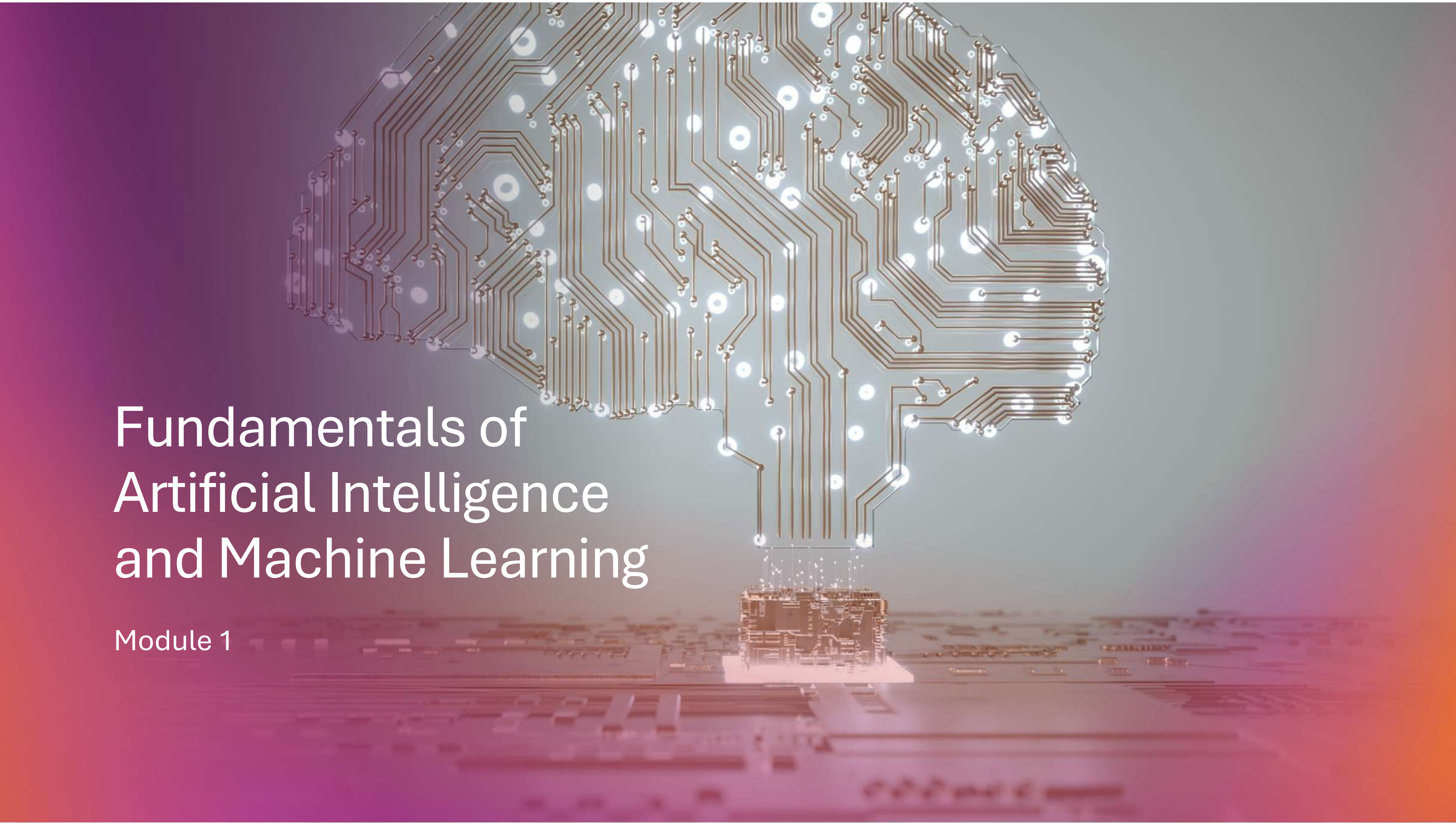


Fundamentals of Artificial Intelligence and Machine Learning

Module 1

Course Learning Objectives

- Understand the basics of Machine Learning and Data-Analytic thinking in a business context.
- Learn to build and evaluate simple models for decision-making.
- Solve business problems to classification tasks.
- Effectively communicate AI concepts and insights in the workplace.

Introductions

Victoria has a diverse background in software development and music education. She holds a Bachelor of Music degree from Temple University and a Master of Science in Business Analytics from Rutgers University, where she graduated Summa Cum Laude.

With 5 years of experience at Miles IT as Software Development Lead and Project Manager, Victoria's expertise lies in system architecture, machine learning, and data analytics. She excels in leading teams and is passionate about using technology to create impactful solutions while remaining committed to continuous learning.

About Yourself

What are you hoping to gain from this course?

What brought you to the WFD program? What is the next step of your career?

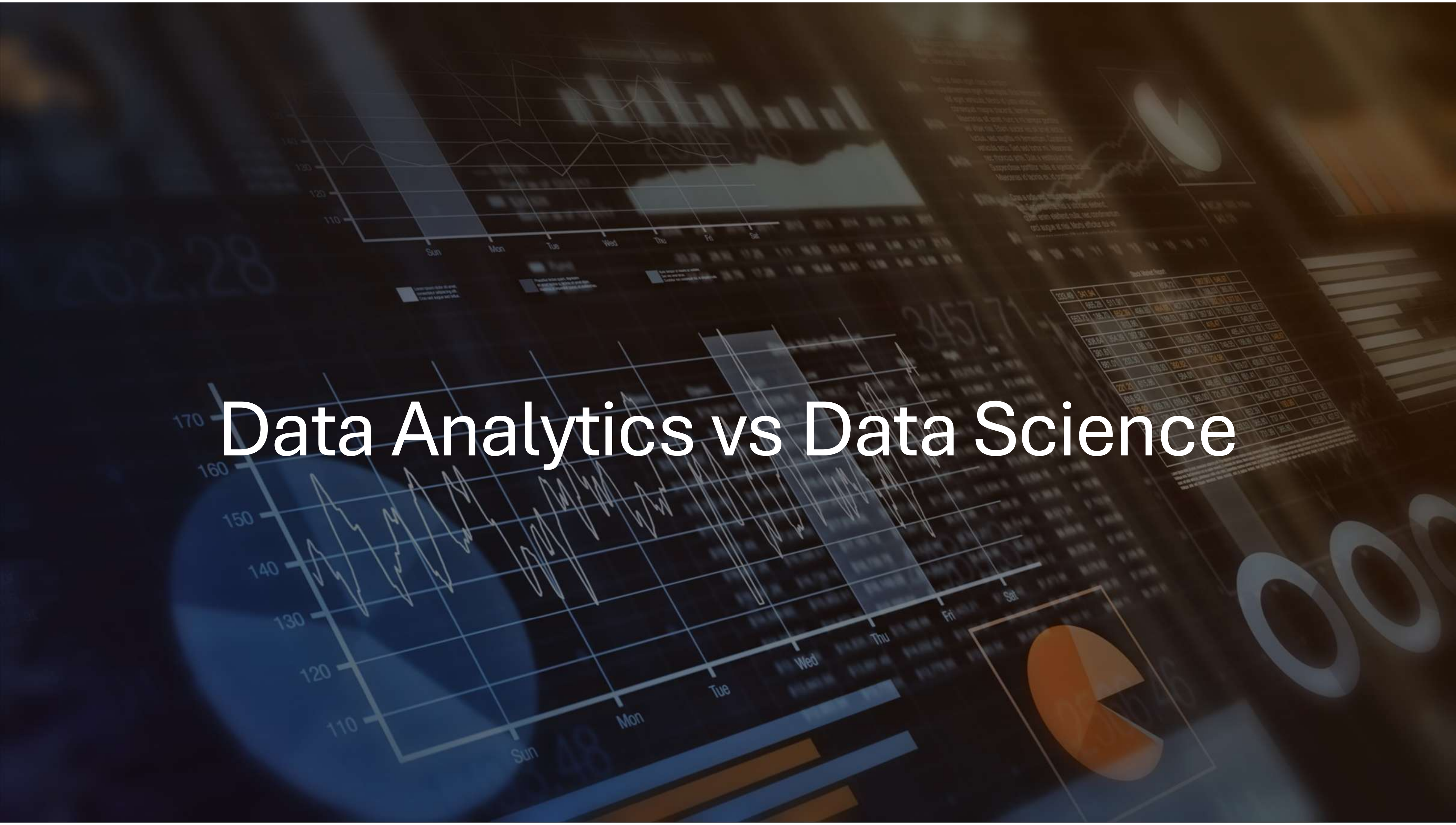
A hobby or fun fact!

First Session

- Data Analytics vs Data Science
- Machine Learning and AI
- AI in Everyday Technology
- AI in Business
- Ethical Considerations – Discussion
- Setting up Machine Learning Environment
- Theory and Practice
  - *Classification*
  - *Decision Trees*
  - *Nearest Neighbors Algorithm*
  - *Famous Iris Data Set*

Types of Analytics

- Descriptive Analytics
  - Understanding past events by summarizing and interpreting historical data.
- Diagnostic Analytics
  - Attempting to understand why something happens. Identifies the root causes of specific events or patterns contributing to certain outcomes.
- Predictive Analytics
  - Uses historical data and statistical models to forecast future outcomes and trends.
- Prescriptive Analytics
  - Focuses on finding the best course of action for achieving desired objectives and outcomes.



Data Analytics vs Data Science

Data Analytics

- Interpreting data to identify trends, patterns, and insights.
- Helps in business and decision making by answering specific business questions.
- Focus is on descriptive and diagnostic data, frequently dealing with historical data, and performing statistical analysis.
- Tools include Tableau, and Power BI for visualizing and reporting data.

Data Analytics - Roles and Contributions to Business

- Identifying Insights and Trends
  - Example: Analyzing sales data to identify top-performing products.
- Supporting Decision-Making with Data
  - Example: Analyze campaign performance and determine which channels or strategies were most effective.
- Tracking and Measuring KPIs
  - Example: Analyzing customer conversion rates, average order value, customer satisfaction.
- Reporting and Business Intelligence
  - Example: Building a monthly performance report to show revenue trends, customer demographics, and monitoring of business health.

Data Science

- Data science is broader and more advanced than data analytics.
- Involves more advanced mathematical models, algorithms, and computational techniques.
- Not only analyses data but predicts future outcomes.
- Focus is on predictive and prescriptive analytics.
- Tools include machine learning algorithms like regression, clustering and deep learning, as well as programming languages like Python and R, and data manipulation frameworks.

Data Science - Roles and Contributions to Business

- Predicting Future Trends
  - Example: Using a predictive model to forecast demand for products, optimizing inventory management and avoid stockouts or overstocking.
- Creating Automated Solutions
  - Example: Creating recommendation engines that suggest products to customers based on their browsing history and preferences.
- Personalizing Customer Experience
  - Example: Streaming services might personalize recommendations based on a user’s history and preferences
- Building Predictive Models and Decision Systems
  - Example: Financial institutions might use data science to predict loan default rates, improving risk management strategies

Quick Quiz!

Which type of analytics is primarily focused on understanding past events?
a) Predictive Analytics
b) Descriptive Analytics
c) Prescriptive Analytics
d) Diagnostic Analytics

What does diagnostic analytics aim to do?
a) Summarize historical data
b) Identify the root causes of events or patterns
c) Forecast future outcomes
d) Find the best course of action

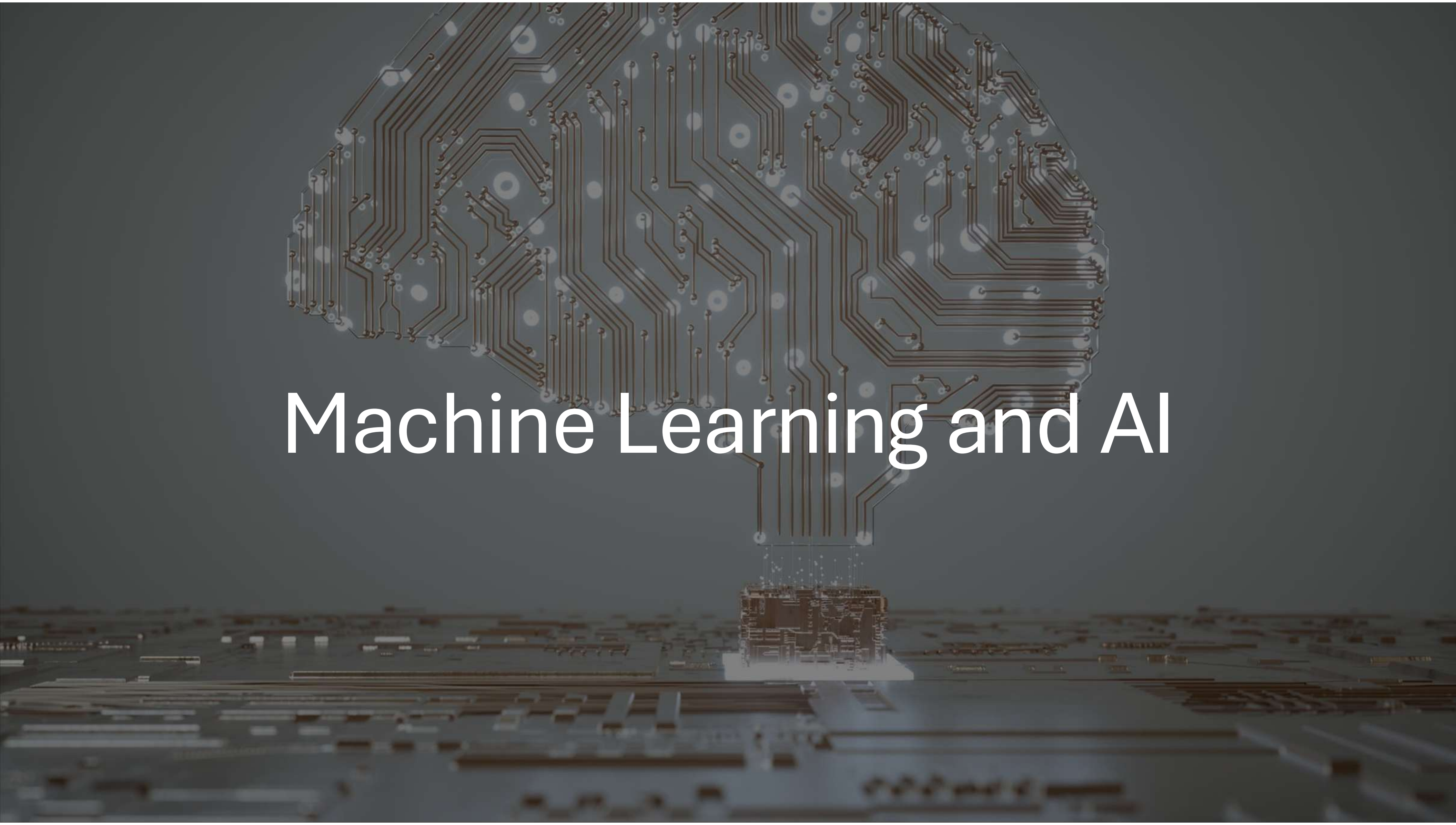
Which analytics method uses historical data to forecast future trends?
a) Descriptive Analytics
b) Diagnostic Analytics
c) Predictive Analytics
d) Prescriptive Analytics

Quick Quiz … continued

True or False?

1. Data science focuses on predictive and prescriptive analytics.
2. Machine learning algorithms are a key part of data analytics.
3. Tracking KPIs is an example of how data analytics contributes to business.
4. Data science often uses programming languages like Python and R.

Machine Learning and AI

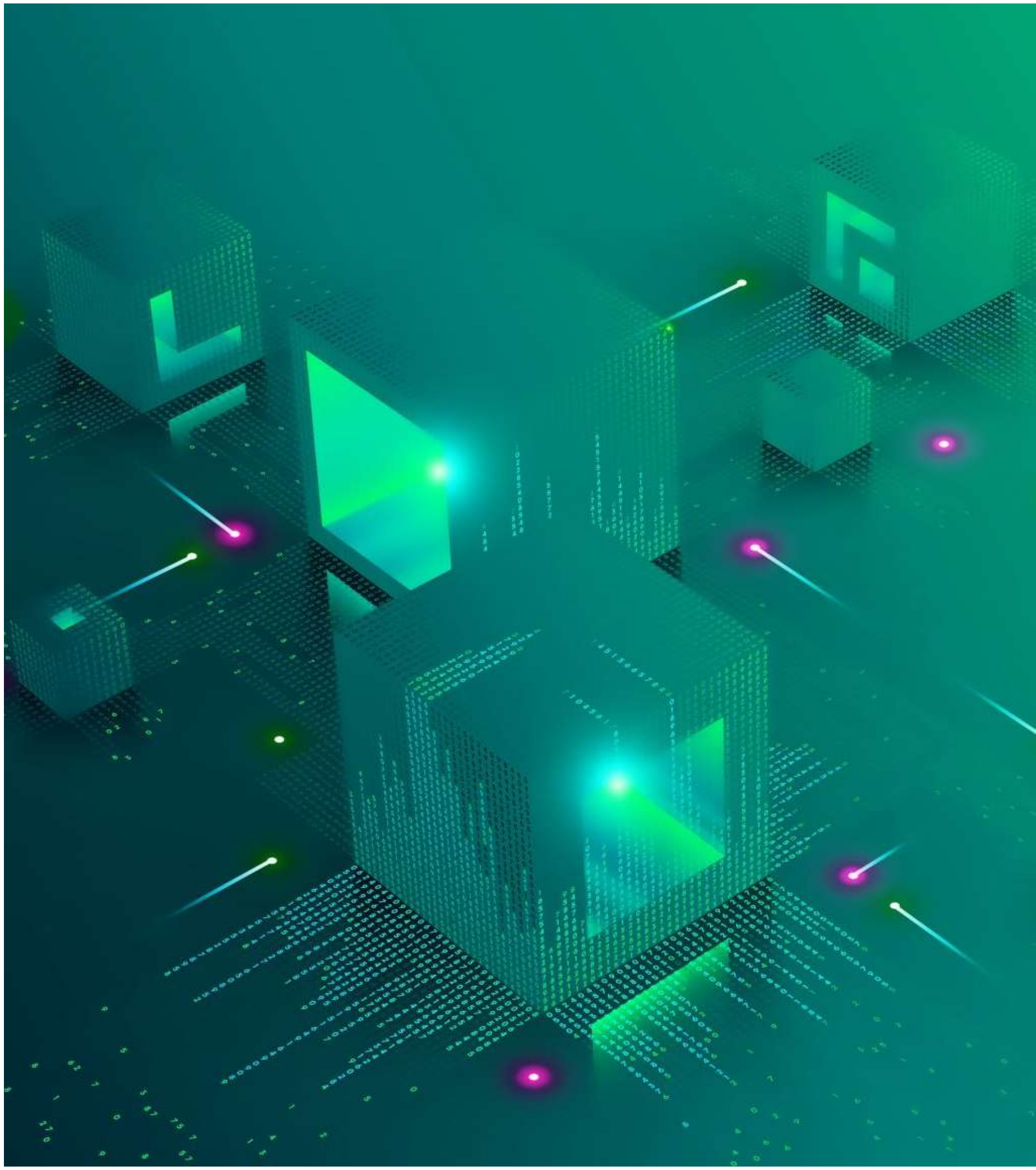


Artificial Intelligence

- Simulation of human intelligence by machines.
- Wide range of tasks like learning, reasoning, problem-solving, perception, and language understanding.
- AI systems are designed to make decisions and automate processes.
- Some AI systems can even adapt their models based on new data they process.

Machine Learning

- Subset of AI that focuses on developing algorithms that enable machines to learn from data, and improve their performance over time without being explicitly programmed.
- Instead of following a fixed set of rules, ML systems identify patterns and make decisions based on data.



Types of Machine Learning

- Supervised Learning
  - Learning from labeled data (input-output pairs)
  - Examples:
    - Spam detection
    - Image classification
- Unsupervised Learning
  - Learning from unlabeled data, identifying hidden patterns or structures withing the data.
  - Examples:
    - Customer segmentation in marketing
    - Fraud detection

Other Types of AI Subfields

- Natural Language Processing NLP
- Computer Vision
- Deep Learning (subset of machine learning based on neural networks)

Quick Quiz!

1. What is the main difference between supervised and unsupervised learning?
2. Give two examples of AI in everyday technology.
3. Name one application of supervised learning and one of unsupervised learning.

Customer Service

Chatbots and virtual assistants for 24/7 support

Marketing and Sales

Personalized recommendations

Predictive analytics for customer behavior

Finance

Fraud detection and credit scoring

Algorithmic Trading

Operations & Supply Chain Management

Inventory Optimization

Predictive Maintenance for Equipment

AI in Everyday Technology

- Virtual Assistants
- Predictive Text and Autocorrect
- Facial Recognition
- Content Recommendations
- Content Curation
- Chatbots
- Photo and Video Filters
- Shopping Recommendations
- Chatbots for Customer Service
- Fraud Detection
- Navigation Apps
- Ride Sharing Services
- Autonomous Vehicles
- Fraud Detections
- Robo Advisors
- Robo Traders
- Credit Scoring
- Healthcare Diagnostic Tools
- Health Monitoring Apps
- Virtual Health Assistants
- Spam Filters

Changing Role of Business Professionals

Decision-making	Using AI insights for better business strategies
Automation	Understanding how AI streamlines workflows
Collaboration	Working with data scientists and IT teams to implement AI
Ethical Considerations	Recognizing biases and ensuring ethical AI use
Competitiveness	Staying ahead by leveraging AI-driven tools

Predictive Models

Forecasting sales trends and demand

Sentiment Analysis

Gauging customer feedback from reviews or social media

Optimization

Pricing strategies based on competitor analysis

Market Segmentation

Identifying niche customer groups for tailored campaigns

Ethical Implications - Bias

- AI can inherit biases based on data they are trained on leading to discriminatory outcomes.
- Example: hiring which will favor certain demographics, facial recognition systems that perform poorly for certain racial or ethnic groups.
- Discussion topic: How do we ensure fairness and inclusivity in AI decision-making?

Ethical Implications – Privacy Concerns

- AI relies on vast amounts of personal data.
- Collecting this data may raise concerns about surveillance and misuse.
- Examples:
    - AI systems tracking online behavior for targeted ads
    - Governments using AI for mass surveillence
- Discussion topic: How do we balance innovation with protecting individual privacy?

Accountability and Transparency

- AI systems function as “black boxes” making their decisions difficult to understand.
- Examples:
    - Autonomous vehicles making life-or-death decisions in accidents.
    - Credit scoring algorithms denying loans without clear explanations
- Discussion topic: Who is responsible when AI systems make mistakes or cause harm?

Job Displacement

- Automation and AI-driven systems can replace human workers, leading to unemployment and economic inequality.
- Examples:
  - AI automating roles in manufacturing, customer service, and logistics.
  - Reduced demand for certain skill sets due to AI advancements.
- Discussion Topic: How do we reskill workers and ensure equitable job opportunities?

Manipulation and Misinformation

- AI-generate content such as deepfakes, can spread misinformation or manipulate public opinion.
- Examples:
  - Fake videos or audio clips used to harm reputations or influence elections
  - Social media algorithms amplifying false or biased information.
- Discussion topic: How do we detect and counteract AI-generated misinformation?

Ethical AI

- **Transparency:** Clear understanding of how systems work
- **Regulations:** Policies to govern AI development and deployment
- **Collaboration:** Multidisciplinary efforts involving technologists, ethicists, and policymakers
- **Accountability:** Assigning responsibility for AI outcomes

How Can You Get Started?

- Learn the basics: Understand key AI and ML concepts
- Leverage Tools: User-friendly AI platforms (Tableau, Power BI)
- Collaborate with Experts: Partner with a Data Science Team, Participate in an Open Source project
- Stay Informed: Keep up with industry Trends

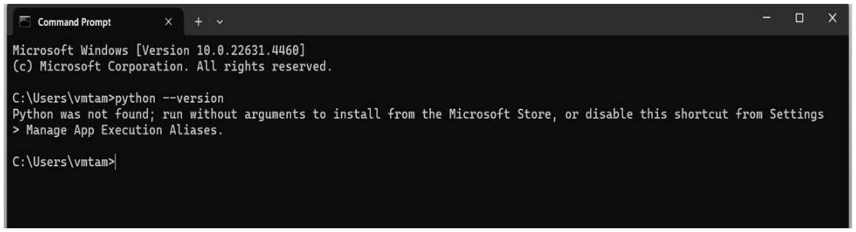
Setting up Scikit-learn

If you have trouble in the first class…

- It’s ok! We’re just going to start with some python basics.
- Email me after class with any issues that you might be having.
- https://www.online-python.com/
    - Try using this in the meantime!

Step 1: Install Python (if not already installed)

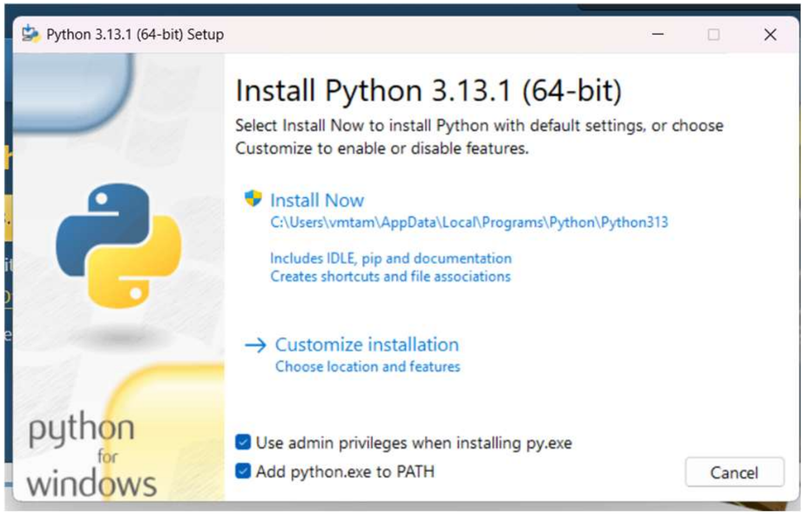
- Before proceeding with installation of any packages, ensure that Python is installed on your system. You can check if it’s already installed by running
  - python –version
- If python is not installed, you can download it from python.org



```
Microsoft Windows [Version 10.0.22631.4460]
(c) Microsoft Corporation. All rights reserved.

C:\Users\vmtam>python --version
Python was not found; run without arguments to install from the Microsoft Store, or disable this shortcut from Settings
> Manage App Execution Aliases.

C:\Users\vmtam>
```



Install Python 3.13.1 (64-bit)

Select Install Now to install Python with default settings, or choose Customize to enable or disable features.

C:\Users\vmtam\AppData\Local\Programs\Python\Python313

Includes IDLE, pip and documentation
Creates shortcuts and file associations

☑ Use admin privileges when installing py.exe
☑ Add python.exe to PATH

Step 2: Install Jupyter Notebook

- If you don’t already have Jupyter Notebook installed, you can install it using pip (Python’s package installer)
    - pip install notebook

Step 3: Install Scikit-learn

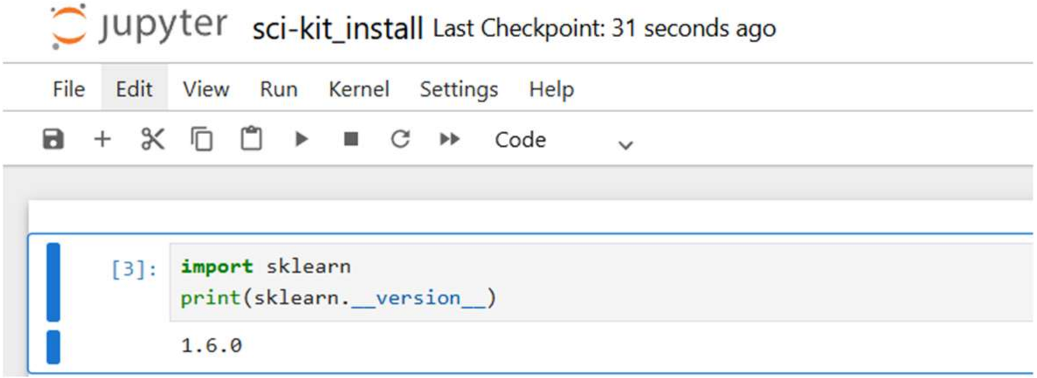
- To install scikit-learn run the following command in your terinal
  - pip install scikit-learn

Step 4: Launch Jupyter Notebook

- Once Jupyter Notebook and scikit-learn are installed, you can launch Juptyer by running
  - jupyter notebook
- This will open Jupyter in your web browser.
- Create a folder directory on your PC for this course.

Step 5: Import Scikit-learn in a Notebook

- After opening Jupyter Notebook, create a new Python notebook and try importing scikit-learn to make sure it's installed correctly

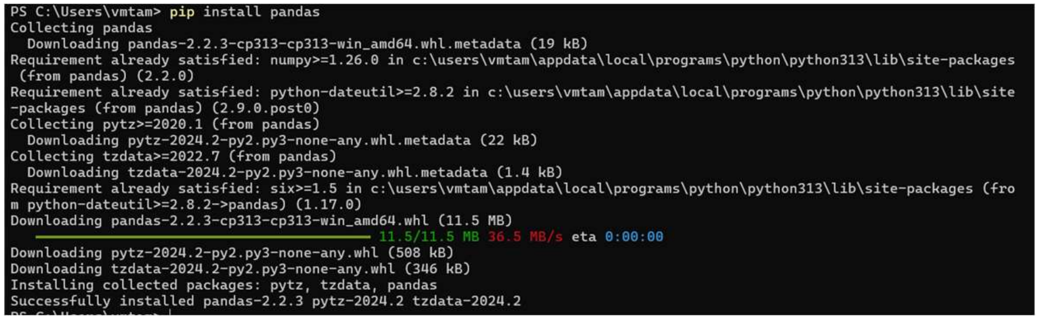


```
import sklearn
print(sklearn.__version__)
```

```
1.6.0
```

Step 6: Install common packages

- *You will need to be comfortable installing certain packages. Attempt to install the following common packages using the below command*
  - pip install pandas
- Repeat for numpy, matplotlib, imageio, mglearn



```
PS C:\Users\vmtam> pip install pandas
Collecting pandas
  Downloading pandas-2.2.3-cp313-cp313-win_amd64.whl.metadata (19 kB)
Requirement already satisfied: numpy>=1.26.0 in c:\users\vmtam\appdata\local\programs\python\python313\lib\site-packages
 (from pandas) (2.2.0)
Requirement already satisfied: python-dateutil>=2.8.2 in c:\users\vmtam\appdata\local\programs\python\python313\lib\site
-packages (from pandas) (2.9.0.post0)
Collecting pytz>=2020.1 (from pandas)
  Downloading pytz-2024.2-py2.py3-none-any.whl.metadata (22 kB)
Collecting tzdata>=2022.7 (from pandas)
  Downloading tzdata-2024.2-py2.py3-none-any.whl.metadata (1.4 kB)
Requirement already satisfied: six>=1.5 in c:\users\vmtam\appdata\local\programs\python\python313\lib\site-packages (fro
m python-dateutil>=2.8.2->pandas) (1.17.0)
Downloading pandas-2.2.3-cp313-cp313-win_amd64.whl (11.5 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 11.5/11.5 MB 36.5 MB/s eta 0:00:00
Downloading pytz-2024.2-py2.py3-none-any.whl (508 kB)
Downloading tzdata-2024.2-py2.py3-none-any.whl (346 kB)
Installing collected packages: pytz, tzdata, pandas
Successfully installed pandas-2.2.3 pytz-2024.2 tzdata-2024.2
PS C:\Users\vmtam>
```

Tools Overview

Python Intro

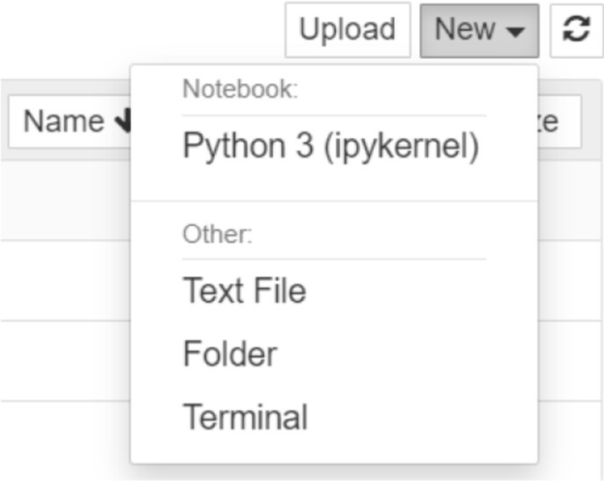
- Python Programming Language
  - w3schools Intro
- Used for
  - Server Side Web Development
  - Mathematics
  - System Scripting and automation
- Capabilities
  - Handling big data and complex analytics
  - Connection to databases and file modification
- Syntax Notes
  - Indentation matters!
  - Designed for readability
  - No semicolons – new lines complete a command

Jupyter Notebook

- How to Use Jupyter Notebook in 2024: A Beginner's Tutorial
- Web-based IDE that blends the code and the output
- Code is placed in a window, and the results are printed immediately, which is essential for data exploration, testing, and sharing of insights

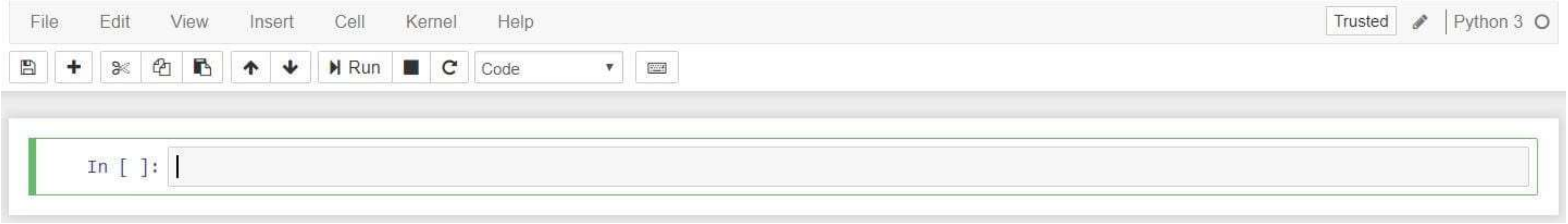
Create a new Jupyter Notebook

1. Select the directory you want to save the file in
2. Click New > Python 3 (ipbykernel)
3. Click 'Untitled' at the top, give the file a name, and save it



Notebook Interface

- Two terms in the Menu Bar to know: Cell and Kernel
  - Cell: block or section where you write your code or text. The code you write will be executed. Cells help organize work in your workbook
  - Kernel: computational engine that runs your code. When you write the code in the cell, it processes it and returns the results.
- To run the code you can either click into the cell and hit run, or click ctrl + enter



Package: Scikit-learn

- Getting Started — scikit-learn 1.6.0 documentation
- Open Source Machine Learning Library
- Supports Supervised and Unsupervised Learning
- Included Tools
    - Model Fitting
    - Data Preprocessing
    - Model selection
    - Model Evaluation

We need to walk before we run!

- Python basics

Classification

Classification Definition

- Task of “classifying things” into categories.
- Classification is part of “Supervised Machine Learning”
  - Provide data with an input, and an output
    - For example: These flowers have x measurements, and they are classified as a y species of that flower.
  - Machine learning algorithm is trained on this labeled dataset to predict the class or category of new, unseen data
    - Y = f(x)
  - The goal is to create a classification **model** that can make future predictions

Classification Types

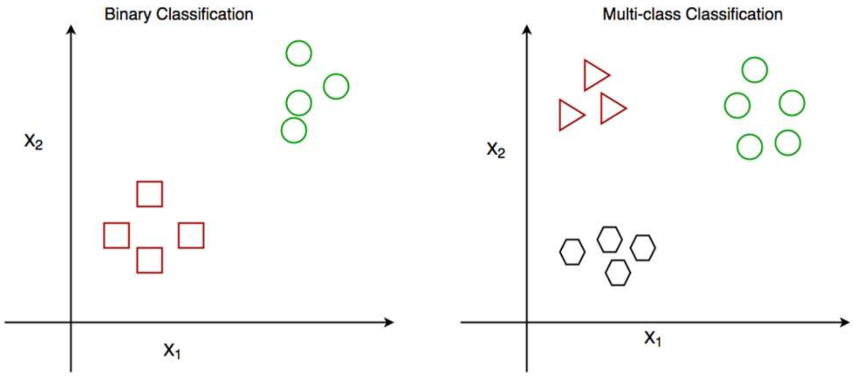
- Two main types of classification in ML
  - Binary Classification
  - Multi-class Classification

Binary Classification

- Categorizing data into two distinct classes or outcomes
  - Yes/No, True/False, Positive/Negative
- Examples:
  - Spam Detection
  - Medical Diagnosis
  - Sentiment Analysis
- Common Algorithms
  - Logistic Regression
  - Support Vector Machines
  - Naïve Bayes
  - Neural Networks

Multi-Class Classification

- Categorizing data into three or more distinct classes. Each data point can only belong to one category.
- Examples:
  - Image recognition
  - Handwritten Digit Recognition
  - Product Categorization
- Common Algorithms for Multi-Class Classification
  - Decision Trees (e.g., Random Forests)
  - K-Nearest Neightbors
  - Naive Bayes



Quick Quiz!

True or False

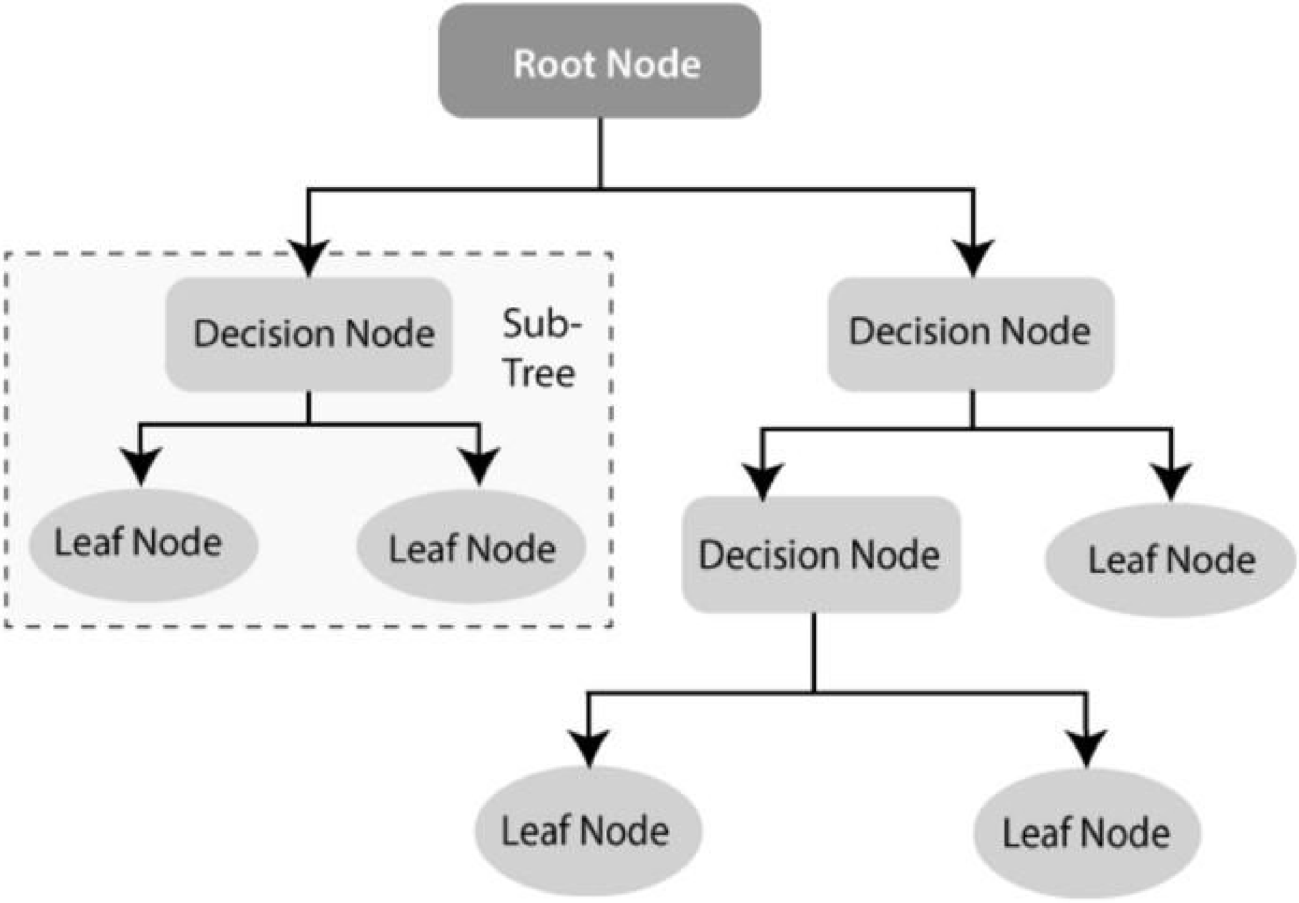
1. Classification models are used to train unlabeled data.
2. Binary classification only involves two possible classes for each data point.
3. Multi-class classification can handle situations where each data point belongs to more than one category.
4. Sentiment analysis is an example of a binary classification task.

Machine Learning Classification Techniques

Decision Trees and Random Forests

Decision Trees

- Flow-Chart structure used to make decisions or predictions
- Decision Nodes
  - Represent decisions or tests on various attributes of the data
  - Root node represents the initial decision to be made.
  - Internal nodes represent the decisions or tests on attributes
- Branches
  - Represent the outcomes of these decisions
- Leaves
  - Representing the final outcomes or decisions


Root Node
Decision Node
Sub-Tree
Decision Node
Leaf Node
Leaf Node
Decision Node
Leaf Node
Leaf Node
Leaf Node

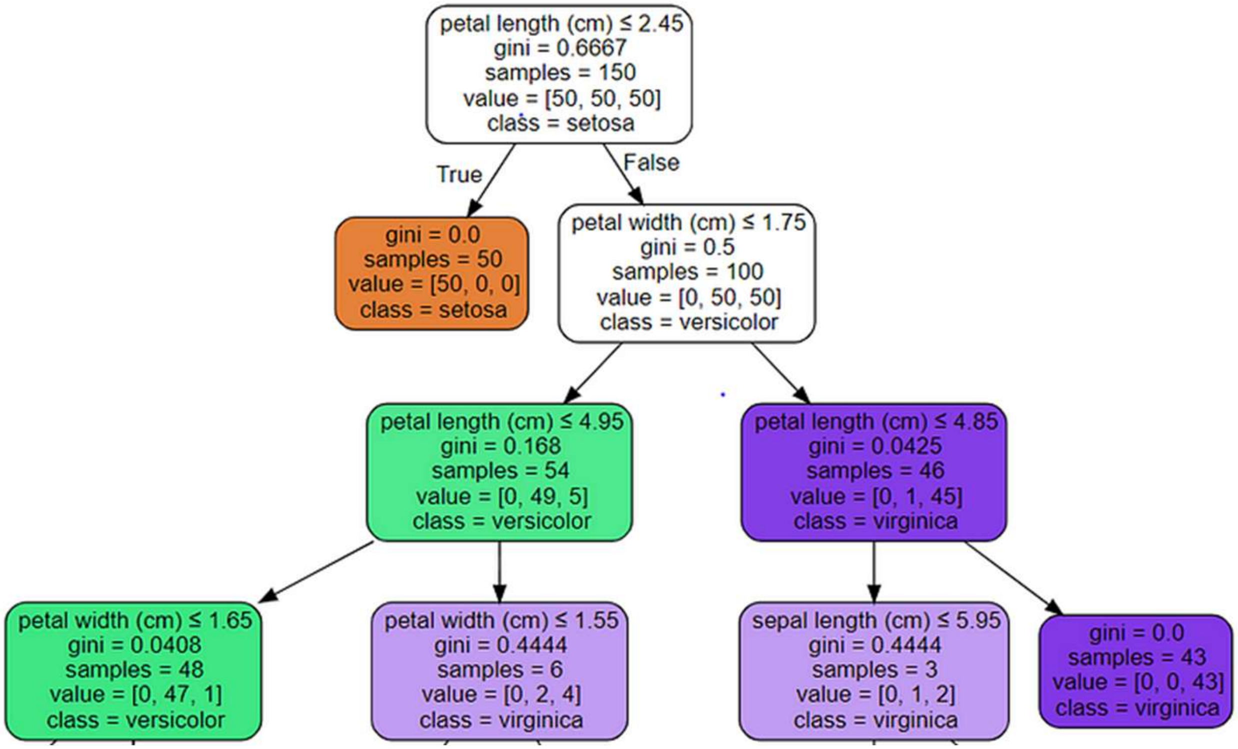
Branch Splitting

- Determines the branch path (Yes/No) based on statistical metrics
  - Gini Impurity
    - How often a randomly chosen element would be incorrectly classified
  - Entropy
    - Degree of disorder or uncertainty

Metrics for Splitting Branches

- Gini Impurity
  - Likelihood that a classification of an incorrect classification if it was randomly classified according to the distribution of the classes
  - $Gini = 1 - \sum_{i=1}^{k} {p_i}^2$, where p_i is the probability of class i
- Entropy
  - Amount of uncertainty or impurity in the dataset
  - $Entropy = \sum_{i=1}^{k} p_i log_2(p_i)$, where p_i is the probability of class i

- Information Gain (IG)
  - Informs us how valuable the split in branches is
  - $Information\ Gain = Entropy_{total} - Entropy_{split}$
  - $Information\ Gain = Gini_{parent} - Gini_{split}$


petal length (cm) ≤ 2.45
gini = 0.6667
samples = 150
value = [50, 50, 50]
class = setosa
True
False
gini = 0.0
samples = 50
value = [50, 0, 0]
class = setosa
petal width (cm) ≤ 1.75
gini = 0.5
samples = 100
value = [0, 50, 50]
class = versicolor
petal length (cm) ≤ 4.95
gini = 0.168
samples = 54
value = [0, 49, 5]
class = versicolor
petal length (cm) ≤ 4.85
gini = 0.0425
samples = 46
value = [0, 1, 45]
class = virginica
petal width (cm) ≤ 1.65
gini = 0.0408
samples = 48
value = [0, 47, 1]
class = versicolor
petal width (cm) ≤ 1.55
gini = 0.4444
samples = 6
value = [0, 2, 4]
class = virginica
sepal length (cm) ≤ 5.95
gini = 0.4444
samples = 3
value = [0, 1, 2]
class = virginica
gini = 0.0
samples = 43
value = [0, 0, 43]
class = virginica

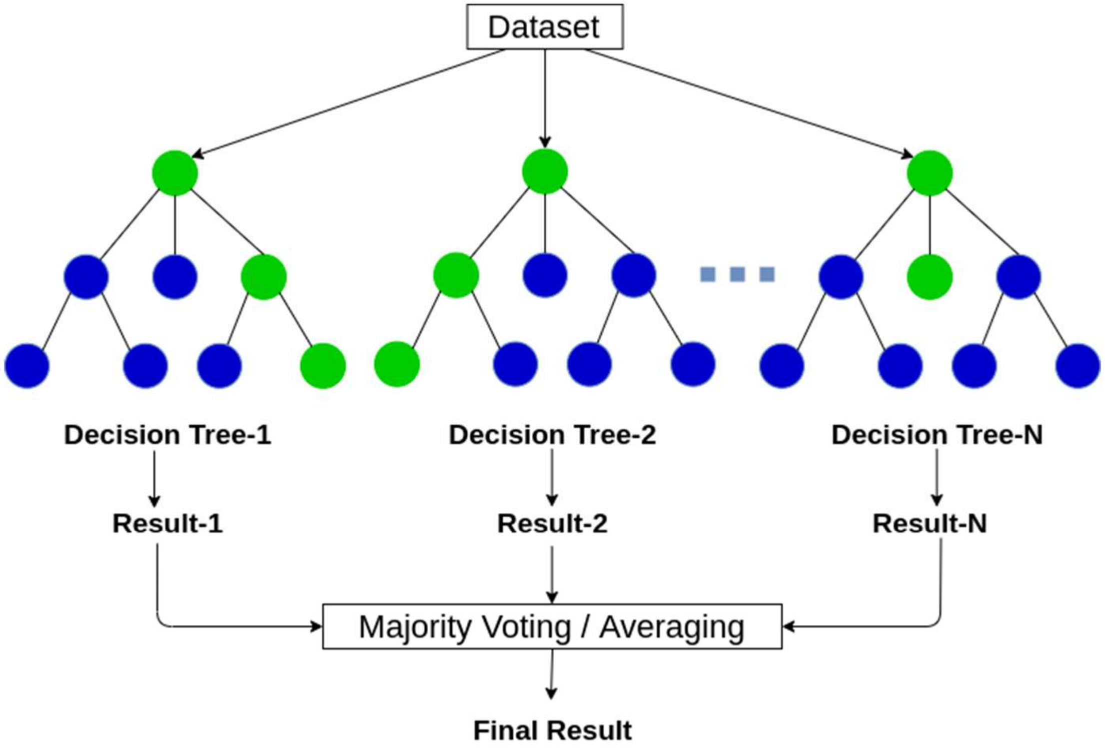
Pros and Cons of a Decision Tree

- Pros
  - Easy to understand
  - Handles categorical and numeric data
- Cons
  - Prone to overfitting
  - Sensitive to small changes

Random Forest

- Ensemble learning method primarily used for classification and regression tasks.
- The theory is that combining many weak models (individual decision trees) can produce a more robust and accurate model.
- Bootstrapping
  - Creating multiple subsets of the training data by selecting data points with replacement. (Some data points may appear in more than one sample)
- Building multiple decision trees
  - Uses random selection of features (encourages diversity among trees)
  - Each Tree is grown to a maximum depth without pruning.
- Combining Results
  - Final prediction is the made by the majority vote for classification tasks.
  - For regression tasks, the prediction is the average of all tree predictions.

Random Forest



Quick Quiz!

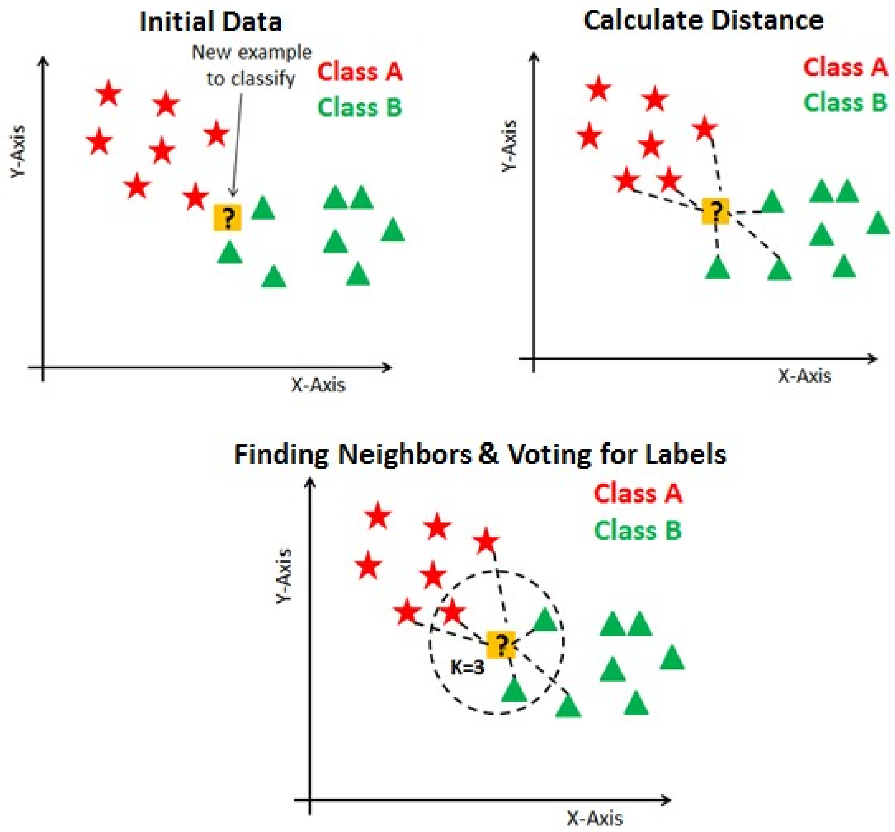
1. Explain the concept of bootstrapping in a Random Forest.
2. How does a Random Forest make predictions for regression tasks?
3. What role does Gini Impurity play in decision tree branch splitting?

Nearest Neighbors Algorithm

- Also known as k-Nearest Neighbors (k-NN)
- Simple learning algorithm for both classification and regression.
- Finds the k nearest data points to a query point by making predictions based on their characteristics. You can define what k is.
- For classification, it assigns the most frequent class label among these neighbors, and for regression it averages the k nearest neighbors.

Distance Metric for k-NN Algorithm

- Most commonly used is the **Euclidean Distance** to measure a straight-line distance between points in a Euclidean Space (2D, 3D, or higher dimensions)
- Derived from Pythagorean theorem, and is widely used in machine learning, computer vision and other fields
- Euclidean Distance
  - $d(p, q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$


Initial Data
New example to classify
Class A
Class B
?
Y-Axis
X-Axis
Calculate Distance
Class A
Class B
?
Y-Axis
X-Axis
Finding Neighbors & Voting for Labels
Class A
Class B
?
K=3
Y-Axis
X-Axis

Typical Machine Learning Pipeline

Preprocessing Step

- Transforms or cleans data

Final Predictor

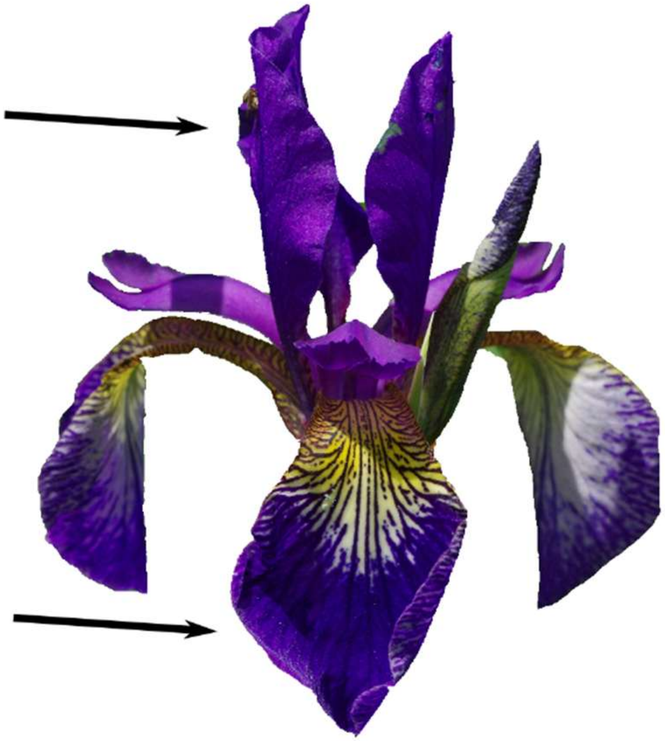
- Prediction of target values

Model Evaluation

- Test to validate the accuracy of the model

Iris Dataset

- Famous dataset used in machine learning and statistics for classification tasks.
- Contains measurements of iris flowers from three different species.
  - Setosa
  - Versicolor
  - Virginica
- We will use k-NN to classify iris flowers based on their measurements.
- 150 samples with 4 features each
  - Sepal Length
  - Sepal Width
  - Petal Length
  - Petal Width



Petal

Sepal

Resources

- Data Science vs Data Analytics - GeeksforGeeks
- Python Ml Decision Tree - Complete Guide – MrExamples
- Random Forest: A Complete Guide for Machine Learning | Built In
- load_iris — scikit-learn 1.6.0 documentation
- 7.1. Toy datasets — scikit-learn 1.6.0 documentation
- The Iris Dataset