EXPLORE AI

2401 PT_DS | NLP & Classification Kick-Off

September 2024

ANNALS OF TECHNOLOGY

THE PASTRY A.I. THAT LEARNED TO FIGHT CANCER

In Japan, a system designed to distinguish croissants from bear claws has turned out to be capable of a whole lot more.

By James Somers

March 18, 2021

EXPLORE AI ACADEMY

01. Announcements

02. Overview of Machine Learning

03. Classification

04. Natural Language Processing

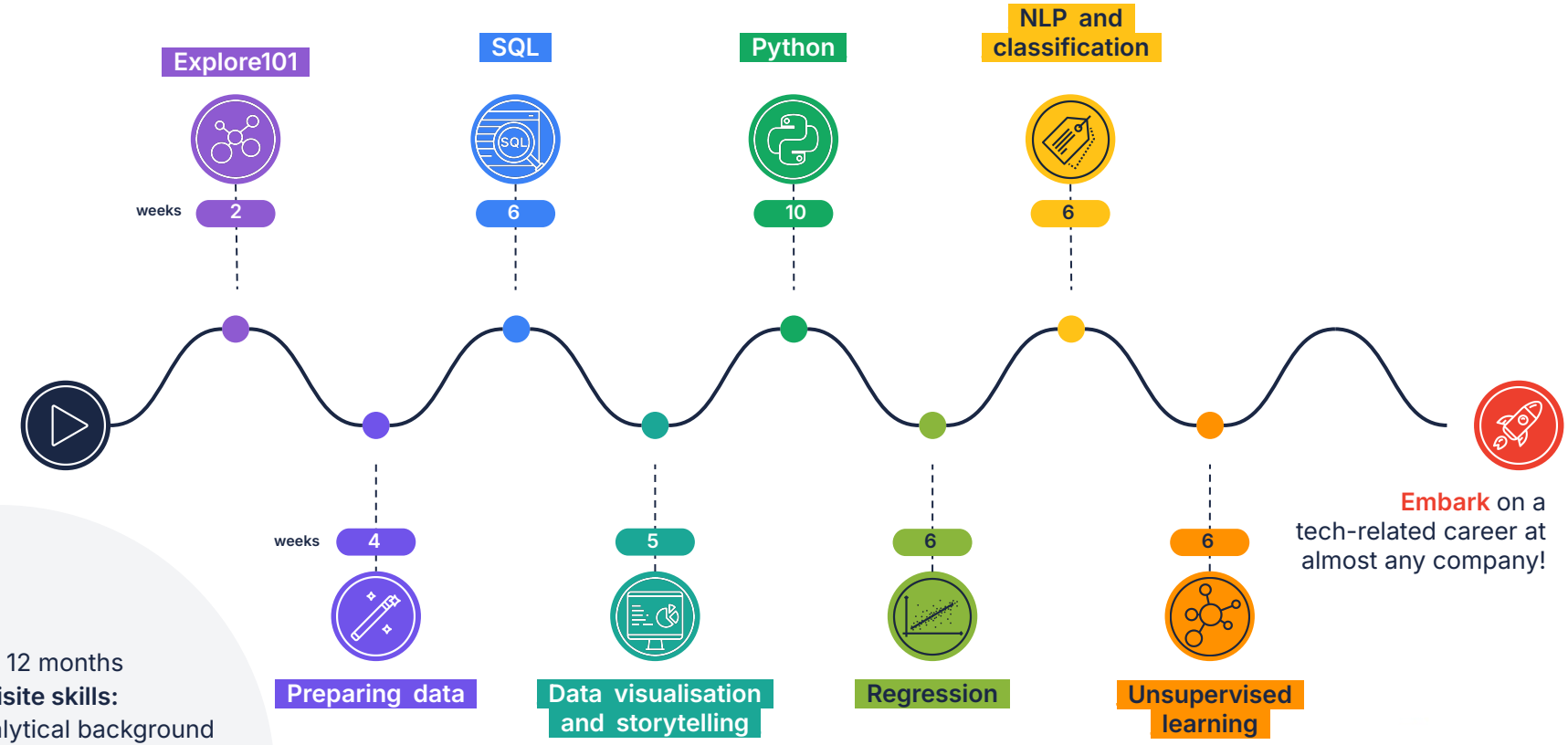
05. Project Kick-off

Announcements

Things to remember...

- Assessments Today, 09 September 2024 @ 11:59PM CAT:
  - Regression Exam [MCQ]
  - Regression Project
- Introduction to NLP and Classification Webinar Thursday
- Verified Zoho Request via Email -> Requirement QCTO Accreditation

Data Science with EXPLORE AI ACADEMY



Duration: 12 months
Pre-requisite skills:
Basic analytical background
Course difficulty: Advanced

EXPLORE AI ACADEMY

01. Announcements

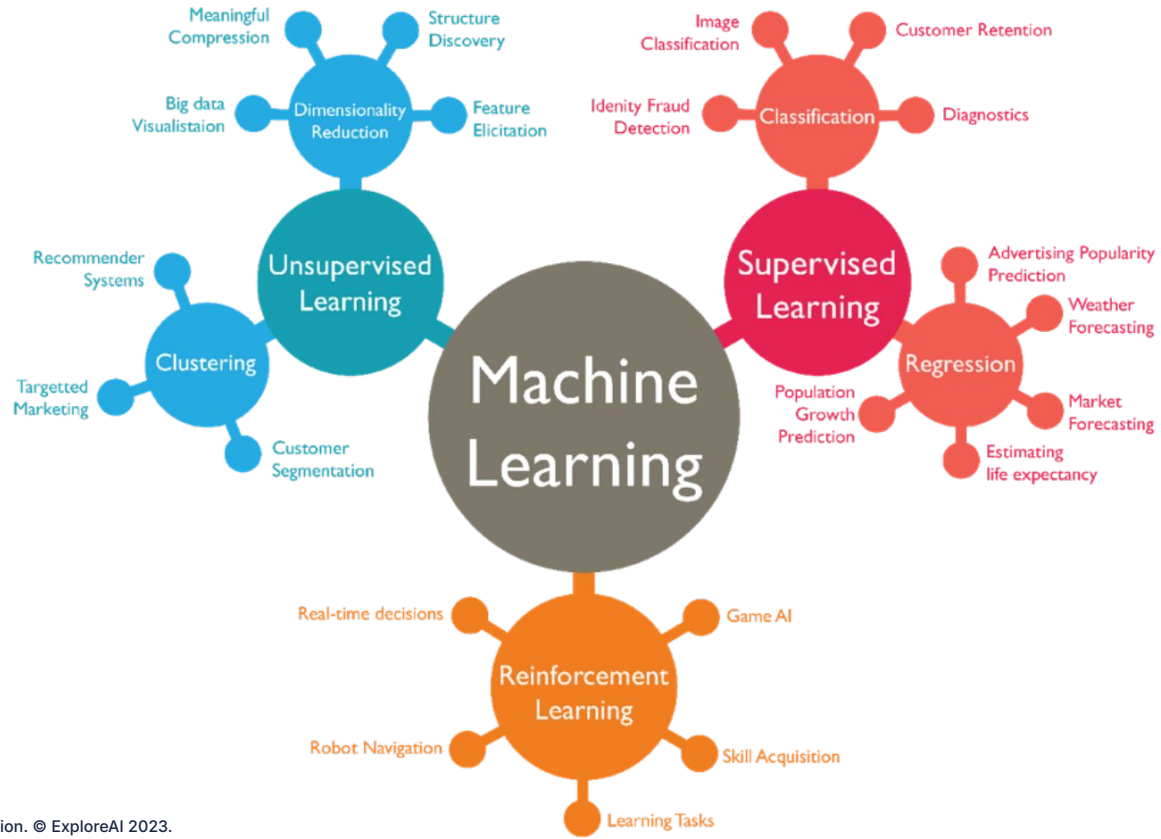
02. Overview of Machine Learning

03. Classification

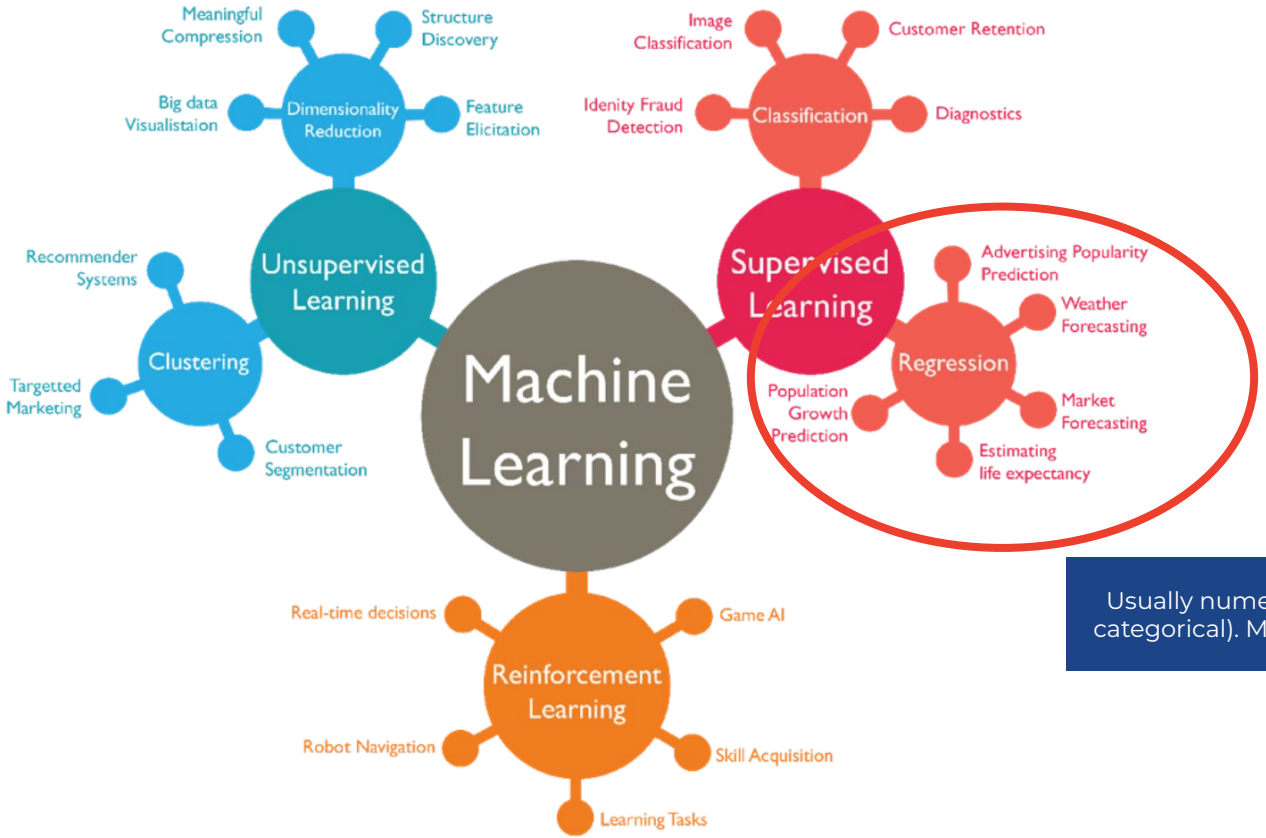
04. Natural Language Processing

05. Project Kick-off

Overview of Machine Learning



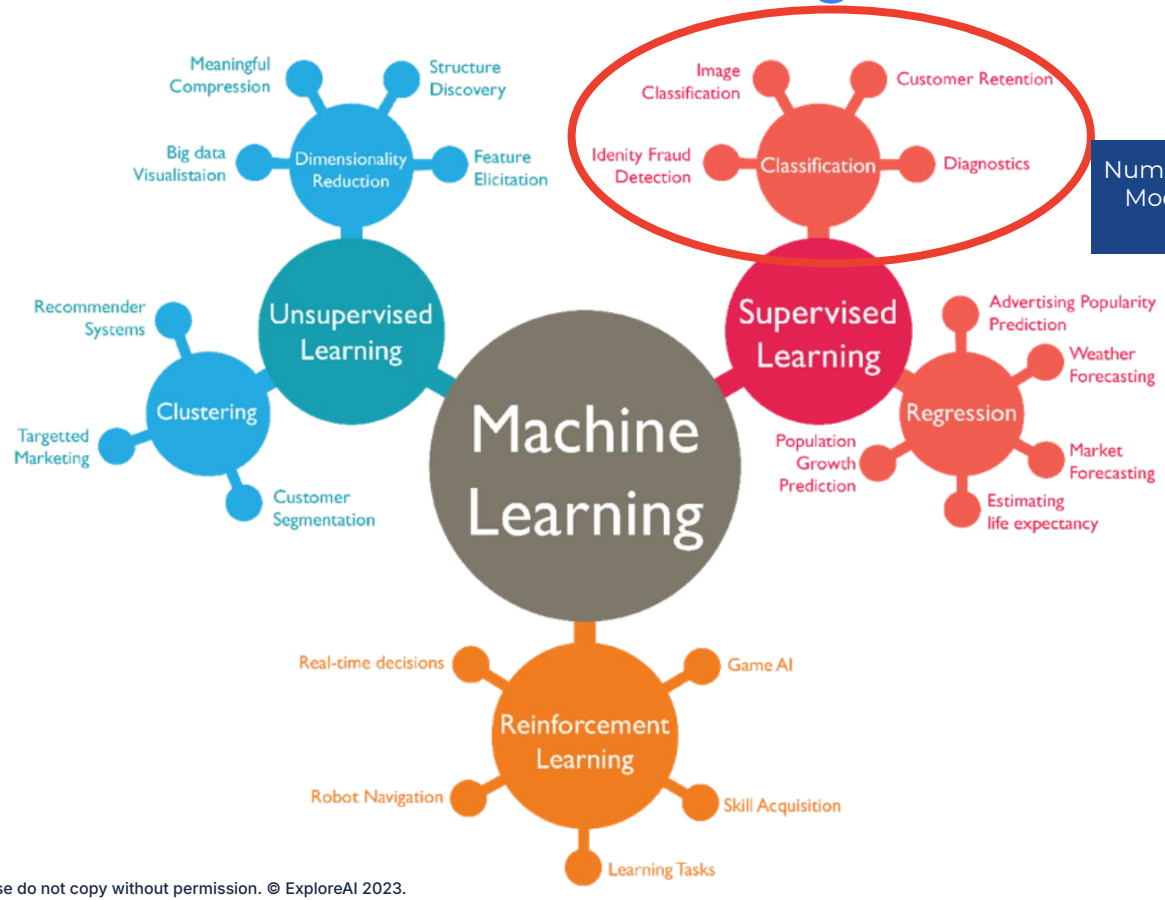
Overview of Machine Learning



Predicting the value of a continuous variable

Usually numerical independent variables (can be categorical). Model for value of dependent variable.

Overview of Machine Learning



Predicting the value of a categorical variable

Numerical and/or categorical independent variables. Model for category, or probability of belonging to category.

EXPLORE AI ACADEMY

01. Announcements

02. Overview of Machine Learning

03. Classification

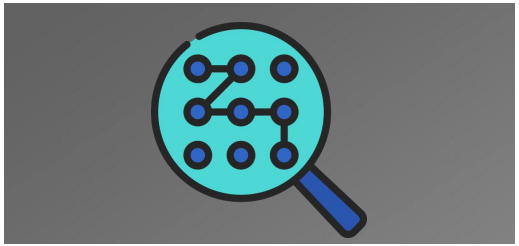
04. Natural Language Processing

05. Project Kick-off

What
Is Classification?

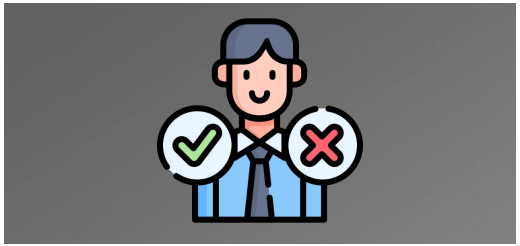
- Division of Machine Learning
- Supervised Modeling
- Categorisation of Data

Why is it Needed



Pattern Recognition

Making sense of complex datasets. Data is noisy finding patterns that we cannot perceive using maths



Decision Making

Making informed decisions by predicting the class labels of new data points

Automated Systems

The process of labeling data, which is crucial for tasks like spam detection, sentiment analysis, and medical diagnosis.

EXPLORE AI ACADEMY

01. Announcements

02. Overview of Machine Learning

03. Classification

04. Natural Language Processing

05. Project Kick-off

What is NLP

Goal: Deep Understanding

- Language
- Context
- Meaning

Reality: Shallow matching

- Needs Robustness
- Fundamental limitations

EXPLORE AI ACADEMY

01. Announcements

02. Overview of Machine Learning

03. Classification

04. Natural Language Processing

05. Project Kick-off

Classification Project

Project Overview

- **The aim of this project is to analyse a dataset that provides a comprehensive collection of news articles** across various domains, including *Business, Technology, Sports, Education,* and *Entertainment.*

- This **end-to-end project covers the entire workflow,** including data loading, preprocessing, model training, evaluation, and final deployment.

- The **app should allows users to input new articles and receive category predictions**, demonstrating the model's practical utility. Hosting the model in a Streamlit app online illustrates its **real-world application** and **value.**

Communication and Project Management

- Optimize team **communication** for all discussions, collaborations, and updates.
- Use **email for formal communications**, such as reporting issues or team member statuses, and send facilitators the names of the Team Lead, Project Manager, and GitHub Manager.
- One team member should be designated as the **Project Manager to create and manage the Trello board**.
- For more information on using Trello, watch this video.

Project Overview

GitHub|Git and Notebook

- **Download the project repository as a zipped folder**, which can be found here.
- **Create a private repository and upload the contents of the downloaded folder.** Ensure all teammates and assigned facilitator are added as collaborators. The facilitators' GitHub usernames can be found here.
- The GitHub repository will require a **README file**.
- Include all the **packages used in a requirements.txt file in your GitHub repository, and add instructions in the README** on how to recreate the environment using Anaconda. Helpful links can be found here and here.
- Exporting your conda environment:

```
conda activate <env>
conda install pip
#get list of packages and pipe to txt file
pip list --format=freeze > requirements.txt
```

MLOps

- **MLOps** is a set of practices that helps manage and improve the process of **building, deploying, and maintaining machine learning models in real-world applications.**
- **MLflow, an MLOps tool, helps track hyperparameter tuning** by logging and comparing different model configurations.
- By using MLflow in your MLOps workflow, you can easily **identify and select the best-performing model based on logged metrics.**

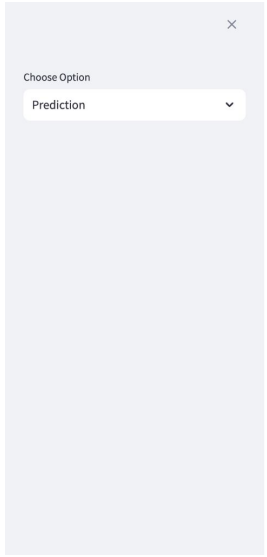
Streamlit App

- Build **your own app** using Streamlit's open source framework. Include a **minimum of three models** into your app.
- The app could be outlined with pages/sections such as your team page, project overview, EDA, and more. Aim to **create a user-friendly interface.**

Project Overview

Streamlit App

Streamlit



Project Overview

Presentation

- Your final task will be to create a presentation **slide deck using Google Slides or Canva.**
- **The slide deck could include an introduction, insights, models, a demo of your Streamlit application, and a conclusion, along with other elements that help tell your story.**

Project Timeline

Classification _ Week 1

- Submit Team Form / Team Preferences
- Get to know your Teammates and Facilitator, and ways of work.

Classification _ Week 2

- Understand the Dataset.
- Elect a team lead, project manager (manages trello), and share with your Facilitator.
- Set up all other collaborative and development tools required for the project.

Project Timeline - Continued

Classification _ Week 3

- Clean your dataset.
- Perform Exploratory Data Analysis on the Dataset.
- Begin working on the slide deck.

Classification _ Week 4

- Perform train-test splits.
- Train with a **minimum of 3 models.**
- Explore and decide on Model Evaluation Metrics.
- Compare and contrast models performance.
- Add all relevant information about the model to the slide deck (Model comparisons, evaluations, and others).

Project Timelines - Continued

Classification _ Week 5

- Continuation of Week 4 Deliverables

Classification _ Week 6

- Complete Deliverables → Send Completion Email, with all resources/deliverables included.

Submit by Deadline:

Monday, 21 October 2024 @ 11:59 PM CAT

Please note the Classification Exam [MCQ] is due the same day, plan your time carefully!

Important links

Please find below, important links:

- Project Repository: Click here
- Facilitator Github Usernames: Click here
- Managing Environments: Click here
- Creating environments from requirements.txt using "*conda create* ": Click here
- Jupyter notebook markdown cheatsheet: Click here
- Sign-up to Trello: Click here
- Video on how to set-up your Trello board: click here
- MLFlow Guide: Click here
- Streamlit Guide: Click here