# Capstone Project 🎓🥳

By: Nourah Almutlag
eng. Esraa Madhi


## About the Capstone Project ⁉️

The main objective behind this project is to use what you learned during the Bootcamp including all libraries and skills that you have gained. Moreover, to evaluate learning outcomes by applying the main concepts using related technologies such as NumPy, pandas, matplotlib, seaborn, Plotly, and scikit-learn. We aim to apply the whole LifeCycle of Data Science and to collaborate as one team on the final project.

> *Remember: In data science, mindset then toolset.*

---

# Essential Requirements 📝

In the capstone project, we will wrap up all skills that you learned in the Data Science LifeCycle in multiple phases as the following:

## Capstone (Graduation Project)

### Step 1: Defining the Problem Statement

- You have to find an interesting question or problem and try to answer this question using data science techniques.
- Discuss three project ideas with your instructors to prioritize them.
- Initial discussion deadline: June 2, 2024.
- Final deadline for idea submission: June 5, 2024.

### Step 2: Collecting Data

- Pick a suitable dataset that helps you to find reasonable answers to your questions.
- Your dataset must be real so that needs to clean and preprocess.
- The dataset should contain a minimum of 3,000 entries for machine learning algorithms, or 10,000 entries for deep learning algorithms. If it does not meet these criteria, you must obtain approval from your instructors.
- Make sure that you really understand your dataset.

### Step 3: Data Quality Checking and Remediation

### Step 4: Exploratory Data Analysis

- For these two steps, make sure to do:

a. Data Profiling: apply the 7 types of data profiling
b. Data Cleaning: handle missing values, correcting errors, and dealing with outliers.
c. Univariate Analysis &Bivariate/Multivariate Analysis: to understand their distribution and look at the relationships between variables. For your visualizations make sure to:
    - Drive meaningful insights that would help you in building models (at least 3 different charts).

## Step 5: Data Modeling

## Step 6: Evaluation

During training model and evaluation, make sure to do:

- Feature Engineering: Apply feature engineering techniques to create new features or modify existing ones to improve model performance.
- Model Training: Choose at least four different machine learning models to train on your dataset. validation set or employ cross-validation to assess model performance during training.
- Hyperparameter Tuning: Fine-tune the hyperparameters of each model to optimize performance.
- Performance Metrics: Use appropriate performance metrics to evaluate the models.
- Model Validation: Validate the model's performance using the test set to ensure that the model generalizes well to unseen data.
- Model Comparison: Compare the models based on their performance metrics to determine which model performs the best on your dataset.
- Overfitting & Underfitting Check: Ensure that models are not overfitting or underfitting by comparing training and validation performance.

## Step 7: Data Communication

Document the methodology and conclusions from your model training experience in a one-page README markdown file comprising the following sections:

- Team Members: List all individuals who contributed to the project.
- Introduction: Briefly describe the problem being addressed and the goals of the project.

- Dataset Synopsis and Origin: Provide a summary of the dataset utilized and its source.
- Model Selection: Train at least 2 different machine learning models for each task.
- Feature Engineering: Outline the steps taken to manipulate or create new features to improve model performance.
- Hyperparameter Optimization: Explain the process and methods used to fine-tune model hyperparameters.
- Performance Metric Visuals: Include charts or graphs that illustrate the performance of each model across various metrics.
- Best Model Determination: Explain the criteria for selecting the best-performing model.
- Feature and Prediction Insights: Offer an interpretation of how different features influence the model's predictions.

## Step 8: Model Deployment

Implement the deployment of our most proficient model as follows:

- Construct a FastAPI endpoint to serve the model
- Develop a Streamlit application or any ui as a demo for your project

## Step 9 : Model Performance Maintenance in Production

Not applicable

**Note:** the yellow steps means they are Optional in the project

---

# Final Deliverables 📦

Each team has to create a capstone project repository with the following files:

1. Notebook file(.ipynb).
2. Dataset file.
3. README.md file that follows the provided template

> ***Note:*** *Please, use the* ***proper Markdown format*** *for readability reasons* 🤓🙏*.*

1. Demo (ex: streamlit app)

2. Presentation of 10 minutes
3. 2 min recorded video of your idea, demo

---

# Evaluation Criteria:

1. **Content Accuracy**
- Concepts & Methods: Are data science concepts and methodologies accurately and clearly explained?
- Data & Algorithms: Are data sources, manipulations, and algorithm implementations correctly described and justified?

2. **Clarity of Presentation**
- Structure & Design: Is the presentation logically organized and visually clear?
- Technical Explanation: Are technical details, including data processing and model selection, clearly articulated?

3. **Engagement**
- Interaction & Examples: Does the presenter effectively engage the audience with relevant examples and interactions?

4. **Code Demonstration**
- Code Clarity & Innovation: Is the code clearly explained and shown to be innovative or efficiently written?
- Problem-Solving: Can the presenter adeptly answer technical questions about their code?

5. **Time Management**
- Pacing & Coverage: Is the presentation well-paced and comprehensive within the given time limit (10 min)?
- Conclusion & Q&A: Is there adequate time for a summary and audience questions?

**Project Deadline (13 Jun at 8:00 am)**

---

# Resources:

- https://pub.towardsai.net/building-industry-level-data-science-projects-a-step-by-step-guide-aeef0efb39d8
- https://www.nocode.ai/use-cases-by-industry/