**Machine Learning Project**

For the third module of the Data Science bootcamp, you will have to develop a machine learning model with the data obtained in the Data Analysis stage (or with new data).

Data: Originates from the EDA project, whenever it makes sense. If not, search on Kaggle (or other sources), APIs, or through Web scraping for a topic that interests you. The data volume must not be less than 1,000 observations.

Deadline for submission and presentation: **March 1, 2024**

**Objective**

The objective of this project is to create a predictive Machine Learning model using the data obtained in the Data Analysis stage or by exploring new datasets.

You need to develop a Machine Learning workflow: Data collection, Cleaning, EDA, Feature Engineering, Testing Various Models, Interpretation of Variables, and Business Impact.

You will need to apply knowledge learned so far in programming, data cleaning, visualization, and modeling. You have the opportunity to demonstrate your evolution and the work done during the time offered (we do not want weekend projects).

**Evaluation**

- Presentation (50%) between 10 and 15 minutes maximum! where you will present your case.

- Code (50%) where 30% will be evaluated from the project\_resume.ipynb. This great effort has to be shared, and you will need to create a repository on GitHub to add your personal projects. If the project consists of sensitive data or you do not want to publish it, it can be kept in a private repository (consult teachers).

**Deliverables**

It will consist of uploading to Google Classroom a Github\_rep.txt with the repository of your GitHub with the developed code.

The code must follow the following folder structure:

1. src/: here will go all the code

2. src/utils: all the modules and auxiliary functions created for the development of the project.

3. src/data/raw: the raw datasets of data necessary to start with the analysis and the model. It's possible you won't be able to upload them to GitHub if the files are too heavy, omit in that case.

4. src/data/processed: clean data that are used in the training of the model.

5. src/notebooks: notebooks for cleaning, processing, EDA, and Machine Learning models.

6. src/notebooks/project\_resume.ipynb: a clean notebook where you summarize the steps you have followed in the project and serves as a report of the project.

7. src/train.py: model training in a script, instead of a notebook. This script has to train the model and save it as new\_model.

8. src/model: save here your already trained models ready for production. The chosen model will be named as my\_model.

**GitHub:** you need to create a README with a description of your project. Brief. If you already have a repo of projects, add it and update the README. The purpose of this README is to quickly understand what your project is about and what technologies you have used to carry it out, it is nothing more than a showcase of your work. You must document your work well to finish the bootcamp with a good portfolio. In [this link](https://github.com/alexhuang1117/Data-Science-Portfolio) you have an example of how to have your projects documented on your GitHub. You will find in the markdown notebooks of the class everything you need to create a decent README.

The correct implementation of the code will be evaluated on the same day, so you will have to be very strict with the structure and conditions that are asked in the statement.

**Presentation**

The presentation becomes more complex compared to the EDA because it involves discussing something much more technical. Start as in the EDA, by presenting the case, and develop the ML solution for your data. You will need to answer some key questions in your presentation:

- What problem or need are we going to solve? Can we solve it with ML?

- What solution does your ML model provide?

- Which models have you tested?

- What results and conclusions have you drawn?

- What were the variables of greatest impact?

- What decisions or actions does your model enable you to carry out? What are the business implications?

Presentation formats? No Notebooks, no reports in PDF or HTML. The format should be a presentation in PowerPoint, Prezi, or whatever platform you are most comfortable with, perfectly combined with a dashboard developed using one of the tools learned during the course (Tableau, Streamlit, Plotly...).

Tips: If you use PowerPoint... it will support you during the presentation, here the protagonist is your speech. Please, don't make the presentation a set of disjointed points, try to follow a thread, as if it were a story. Sell yourself, highlight the strengths of your work over the weaknesses. Interact with the audience. If you're going to talk about business, can you translate it into numbers? What impact will your analytics have? Show some specific data (simple statistics) to justify your business case or to close your presentation. If you have chosen a theme that aligns with your tastes/background, take the opportunity at the beginning to share this and empathize more with the audience.

**Project Steps**

1. Get the data

2. Define your Machine Learning problem: classification/regression, supervised/unsupervised, time series, images, text...

3. Exploratory: obtain all the statistics and graphs you need to understand your dataset well.

4. Clean the data: duplicates, missings, outliers, useless columns...

5. Feature engineering: transformation and creation of new variables.

6. Test various models

7. Analyze the results using a metric appropriate to your problem.

8. Interpret the results and understand the model's outputs.

9. Next steps. Can the model be enriched with other tests or data?

10. Must include Random Forest Model.

11. Bonus track (optional): set up a dashboard/report to show how your model learns, what hyperparameters it uses, and the results obtained.

**Example Dataset: Heart Disease UCI**

**Dataset Overview:**

The Heart Disease UCI dataset is a popular dataset used in machine learning to predict the presence of heart disease in patients. It contains a set of attributes related to heart health, including:

Age: The age of the patient.

Sex: The gender of the patient (male/female).

Chest Pain Type: Type of chest pain experienced by the patient.

Resting Blood Pressure: Resting blood pressure (in mm Hg).

Serum Cholestoral: Serum cholesterol in mg/dl.

Fasting Blood Sugar: If fasting blood sugar > 120 mg/dl.

Resting Electrocardiographic Results: Resting electrocardiographic measurement.

Maximum Heart Rate Achieved: Maximum heart rate achieved.

Exercise Induced Angina: Exercise-induced chest pain.

ST Depression: ST depression induced by exercise relative to rest.

Slope of the Peak Exercise ST Segment: The slope of the peak exercise ST segment.

Number of Major Vessels Colored by Fluoroscopy: Number of major vessels colored by fluoroscopy (0-3).

Thal: A blood disorder called thalassemia.

The dataset typically includes a target variable that indicates the presence of heart disease in the patient.

**Machine Learning Question:**

Can we predict the presence of heart disease in a patient based on their medical attributes?

**Approach with Machine Learning:**

To answer this question using machine learning, you would typically follow these steps:

Data Preprocessing: Clean the dataset by handling missing values, encoding categorical variables, and normalizing/standardizing the numerical variables.

Exploratory Data Analysis (EDA): Analyze the data to understand the distribution of various features, the relationship between them, and their correlation with the target variable.

Feature Selection: Identify the most relevant features that contribute to the prediction of heart disease.

Model Selection and Training: Choose appropriate machine learning models for classification (e.g., Logistic Regression, Decision Trees, Random Forest, Support Vector Machines, Neural Networks) and train them using the prepared dataset.

Model Evaluation: Evaluate the models' performance using appropriate metrics (e.g., accuracy, precision, recall, F1 score, ROC-AUC score) through techniques like cross-validation.

Model Tuning and Optimization: Fine-tune the models to improve performance, using techniques like grid search or random search for hyperparameter optimization.

Prediction and Interpretation: Use the best-performing model to make predictions on unseen data and interpret the results to understand the factors contributing to heart disease.

This approach can help healthcare providers to identify individuals at higher risk of heart disease early and potentially guide preventive measures or further diagnostics.

For your machine learning project on predicting obesity, answering the question of the problem or need you're going to solve involves specifying the exact objective of your project and determining the feasibility of solving it with machine learning. Here's a structured approach to formulating your answer:

Identifying the Problem or Need

Objective: Clearly define what you aim to predict or classify regarding obesity. Are you looking to predict the likelihood of an individual becoming obese based on current lifestyle and demographic data? Or are you aiming to identify key factors that contribute most significantly to obesity?

Significance: Explain why this prediction is important. This could be for the purpose of preventative health measures, to aid in the development of personalized treatment plans, or to contribute to public health policies aimed at reducing obesity rates.

Target Audience: Identify who will benefit from these predictions. This could be healthcare providers, individuals at risk of obesity, policymakers, or public health organizations.

Feasibility of Solving with Machine Learning

Data Availability: Assess whether there is sufficient, relevant data to train a machine learning model. This includes considering the types of data needed (e.g., dietary habits, physical activity levels, genetic information, socio-economic status) and whether such data can be ethically and legally obtained and used.

Data Quality: Evaluate the quality of the data available. Machine learning models require clean, well-structured data for effective training. Consider if the data needs preprocessing or cleaning.

Appropriate ML Techniques: Consider which machine learning techniques might be most suitable for the type of data and the problem at hand. As discussed, techniques like XGBoost, SVM, and Decision Trees each have their strengths and limitations. The choice of technique will depend on the specifics of your data and your project goals.

Model Evaluation: Plan how you will evaluate the effectiveness of your machine learning model. This includes choosing appropriate metrics (accuracy, precision, recall, F1 score, etc.) for the type of prediction you're making and considering how to interpret these metrics in the context of your project.

Ethical Considerations: Address any ethical considerations involved in predicting obesity, such as ensuring the privacy and security of personal data, avoiding bias in your model, and considering the implications of false positives or false negatives.

By thoroughly addressing these points, you can formulate a comprehensive answer to the question of what problem or need you are going to solve with your machine learning project and how feasible it is to solve it with ML. It's important to demonstrate that your approach is data-driven, ethically responsible, and methodologically sound.

**What solution does your ML model provide?**

When answering the question about what solution your machine learning (ML) model provides in the context of predicting obesity, you should focus on the practical outcomes, benefits, and potential applications of your model. This involves detailing how your model addresses the problem you've identified, the kind of results it can generate, and how these results can be utilized by the target audience or to achieve your project's objectives. Here's how you might structure your answer:

**Defining the Solution**

**Prediction Capability:** Explain what your ML model is capable of predicting or classifying. For instance, it might predict an individual's risk of developing obesity based on various predictors like dietary habits, physical activity, genetic predispositions, and socio-demographic factors. Be specific about the type of prediction (binary classification, multi-class classification, risk scoring, etc.).

**Insight Generation:** Detail how your model can provide insights into the factors contributing to obesity. This could include identifying which features (e.g., lack of exercise, high-calorie diets, socioeconomic status) are most predictive of obesity, offering valuable information for targeted interventions.

**Decision Support:** Discuss how the model can support decision-making processes. For healthcare providers, it could be used to identify patients at high risk of obesity for early intervention. For individuals, it might offer personalized recommendations to reduce their risk. For policymakers, the model's insights could inform public health strategies and resource allocation.

**Highlighting the Benefits**

**Preventative Measures:** Emphasize the model's potential in enabling preventative healthcare strategies. Early identification of high-risk individuals can lead to timely and targeted interventions, potentially reducing the incidence and severity of obesity.

**Personalized Healthcare:** Illustrate how the model supports personalized healthcare by identifying specific risk factors for individuals, enabling tailored advice and treatments that are more likely to be effective.

**Resource Optimization:** Explain how the model can help healthcare providers and policymakers optimize resources by focusing efforts on high-risk groups or regions, making healthcare delivery more efficient.

**Public Health Impact:** Highlight the potential broader impact on public health, such as contributing to a decrease in obesity rates, reducing the burden of obesity-related diseases, and improving population health outcomes.

**Practical Applications**

**Healthcare Settings:** Your model could be integrated into electronic health record systems to automatically flag patients at high risk of obesity, prompting healthcare providers to discuss lifestyle changes or interventions.

**Public Health Programs:** Insights from your model could guide the development of targeted public health campaigns, focusing on the most impactful factors contributing to obesity.

**Personal Health Apps:** The model could be used in personal health and fitness apps to provide users with personalized risk assessments and recommendations for reducing their obesity risk.

In summarizing the solution your ML model provides, it's important to link back to the specific problem or need you identified earlier and show how your model addresses this in a tangible, impactful way. This demonstrates not only the technical effectiveness of your model but also its real-world applicability and potential to make a positive difference in the field of obesity prevention and management.