**QMBE 3730: Statistical Machine Learning**

**Machine Learning Life Cycle.**

The machine learning cycle is a step-by-step process that guides the development and deployment of machine learning models. It includes various steps from problem definition to model evaluation, deployment, and continuous improvement. The machine-learning life cycle is defined in detail below:

1**. Defining the problem:**

In this initial stage, the problem that needs to be solved using machine learning is defined. This involves understanding the application context, identifying the objectives, and determining which machine learning task, model, and module is appropriate to tackle the problem.

2. **Data Collection:**

Data is the basis of any machine learning project. The output of any machine learning project is only as good as your data. Relevant data needs to be collected from various sources, which might include databases, APIs, or manual data entry. The quality and quantity of data play a crucial role in the success of the project.

3. **Data Preprocessing:**

Raw data is often messy, contains missing values and variables which are not useful for the project. In this stage, the data is cleaned, transformed, and preprocessed. This includes tasks such as handling missing values, removing duplicates, creating features, and encoding categorical variables.

4. **Feature Engineering:**

Feature engineering involves selecting, creating, or transforming features to make them suitable for model training. This step can significantly improve or degrade the model's performance. Theoretical/domain/ logical knowledge is often essential here.

5. **Model Selection:**

Based on the problem type (e.g., classification, regression, clustering), you select an appropriate machine learning algorithm(s) to build the model(s). The choice of algorithm depends on factors such as the nature of the data, the complexity of the problem, the accuracy of desired outcomes, and the domain of the application.

6. **Model Training:**

In model training, the selected algorithm is trained on the training set (usually 70%-90% of the dataset). The model learns the underlying patterns, relationships, and trends in the data – model objects (coefficients, decision rules, margins). The training process involves iteratively updating the model's parameters.

7. **Model Evaluation:**

The trained model's performance is evaluated using suitable evaluation metrics. These metrics vary depending on the machine learning task. For example, classification might use accuracy, precision, recall, and F1-score, while regression might use mean squared error or R-squared.

8. **Model Tuning:**

If the model's performance is not satisfactory, hyperparameters (parameters that are set before training) can be adjusted to improve performance. This involves techniques like grid search and random search to find the best set of hyperparameters. In model tuning we usually use a part of the dataset called the validation set- which is different from the training set and the test set.

9. **Model Validation:**

The model's performance is validated using a separate validation dataset to ensure that it generalizes well to unseen data. Cross-validation techniques can be used to estimate how well the model will perform on new data.

***The parts below are often involved in the machine learning life cycle but for this class we will not cover them.***

10. **Model Deployment:**

Once the model has been trained and validated, it can be deployed to production environments where it can make predictions on new, unseen data. Deployment might involve integrating the model into a web application, mobile app, or other systems.

11. **Monitoring and Maintenance:**

After deployment, the model's performance is continually monitored. Drift in data distribution or changes in the problem domain might lead to decreased performance over time. Regular updates and retraining might be necessary to maintain model accuracy.

12. **Continuous Improvement:**

As new data becomes available and the system is used, feedback can be collected to improve the model further. This feedback loop feeds back into the beginning of the cycle, where data is collected, preprocessing is performed, and the model is retrained.

The machine learning cycle is iterative and involves going back and forth between different stages to refine and improve the model's performance and applicability to real-world problems.