**Course Title:** Data Science and Machine Learning Course

**Project Title:** End-to-End Machine Learning Project - From Data Collection to Model Deployment

This project also will contains a topic-wise curated list of Machine Learning and Data Science tutorials.

**1. Introduction**

This project is undertaken as a part of the **Data Science and Machine Learning** **course** taught by IIT institute to demonstrate the practical application of concepts and tools learned throughout the curriculum. The aim is to work on a complete data science pipeline: from understanding a real-world problem, collecting, and analysing data, building, and validating models, to communicating the insights and (optionally) deploying the model.

**2. Objectives**

The main objectives of the project are:

* To identify a real-world problem that can be addressed using data.
* To apply data science techniques for data preprocessing, analysis, and visualization.
* To build and evaluate multiple machine learning models.
* To interpret results and provide data-driven insights.
* To optionally deploy the best-performing model using a web framework.

**3. Problem Statement**

[Choose a specific problem related to your area of interest – this could vary based on the dataset you select.]  
*Example:*  
With the rapid growth of online retail, customer churn is a major issue for e-commerce businesses. This project focuses on predicting customer churn using historical purchase data and customer behaviour metrics, aiming to help businesses retain valuable customers.

**4. Scope of the Project**

* Use of publicly available datasets or real-time data sources.
* Full-cycle application of the CRISP-DM methodology: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment.
* Use of multiple machine learning algorithms for comparison.
* Emphasis on visualization and interpretation of results.

**5. Methodology**

The project will follow a systematic approach:

1. **Data Collection:**  
   Using open-source datasets from platforms like Kaggle, UCI, or APIs.
2. **Data Cleaning & Preprocessing:**  
   Handling missing values, outliers, feature encoding, and scaling.
3. **Exploratory Data Analysis (EDA):**  
   Understanding patterns and relationships in the data through statistical summaries and visualizations.
4. **Model Building:**  
   Implementing algorithms such as Logistic Regression, Decision Tree, Random Forest, SVM, etc.
5. **Model Evaluation:**  
   Using metrics like accuracy, precision, recall, F1-score, and ROC-AUC.
6. **Model Deployment (Optional):**  
   Building a simple interface using Flask or Streamlit to allow real-time predictions.

**6. Tools & Technologies**

* **Language:** Python
* **Libraries:** Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, TensorFlow/Keras (optional)
* **IDE:** Jupyter Notebook / Google Colab / VS Code
* **Version Control:** Git and GitHub
* **Deployment (Optional):** Flask or Streamlit

**7. Expected Outcomes**

* A clean and structured dataset ready for modelling.
* Visualizations and analytical reports that highlight key insights.
* Trained and validated machine learning models with performance metrics.
* A final report or presentation summarizing the project.
* Optional: A web application for real-time predictions.

**8. Conclusion**

This project serves as a comprehensive demonstration of the skills developed in the **Data Science and Machine Learning** course. It reflects the ability to approach a real-world problem scientifically, apply appropriate tools and techniques, and effectively communicate findings and solutions.