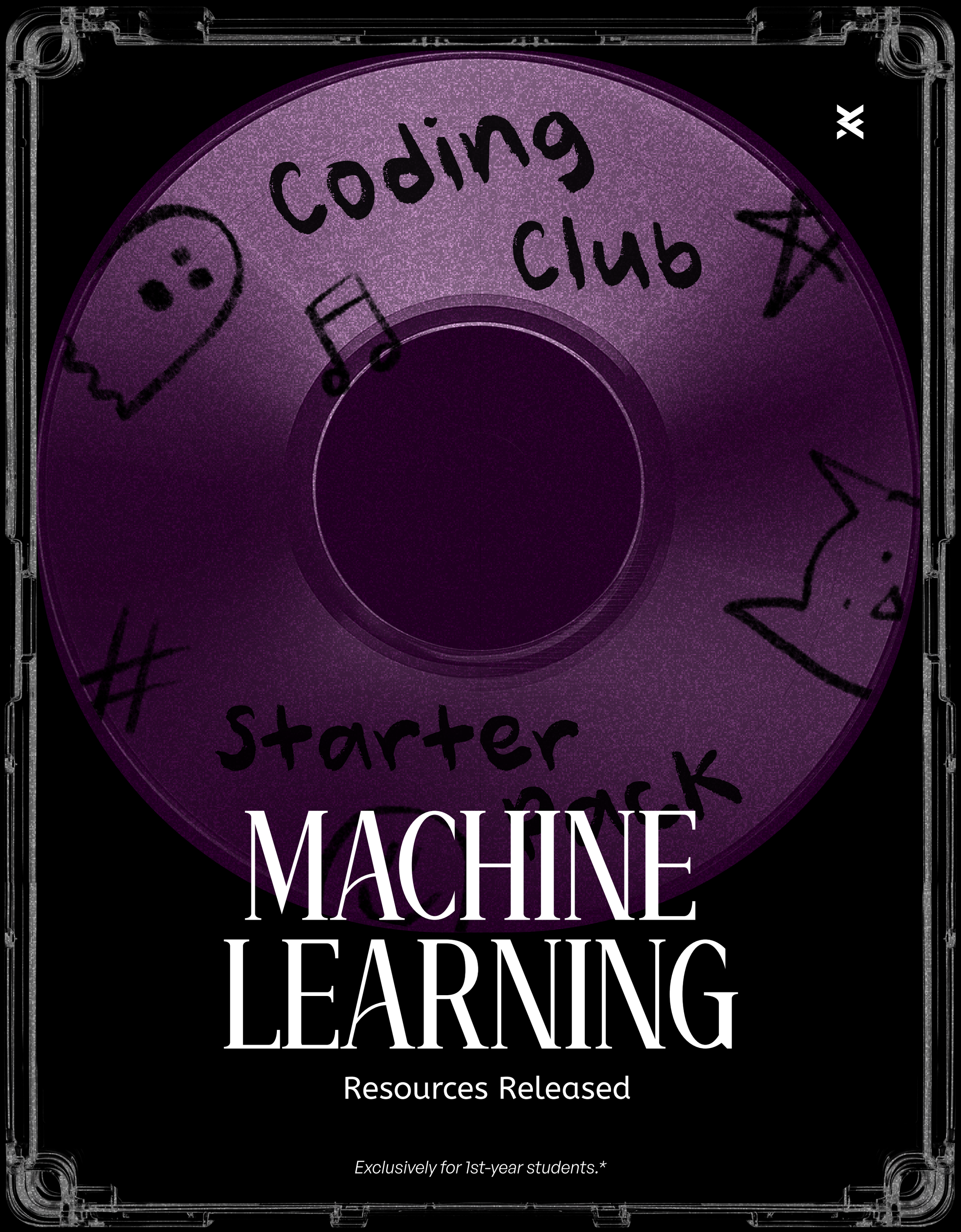
Coding Club
IIT Guwahati

CODING WEEK




# MACHINE LEARNING

## Resources Released

*Exclusively for 1st-year students.**

# What is Machine Learning?

Machine Learning (ML) is a dynamic and rapidly growing subfield of Artificial Intelligence (AI) that focuses on enabling computers to learn from data and make informed decisions without being explicitly programmed. It allows systems to improve their performance over time by recognizing patterns and drawing insights from vast amounts of information.

Today, ML is at the core of many cutting-edge technologies — from image and speech recognition to language translation, recommendation systems, financial modelling, and autonomous vehicles. Its impact is widespread across industries including healthcare, finance, e-commerce, entertainment, and robotics.

As a foundational pillar of modern AI, machine learning offers a wealth of opportunities for innovation and research. Whether you aim to build intelligent systems, analyze complex data, or develop real-world applications, ML provides the tools and techniques to bring those ideas to life.

## Python

Python is the most widely used language in the field of data science and machine learning due to its simplicity, readability, and strong community support. It offers a vast collection of libraries like NumPy, Pandas, and scikit-learn that make it easy to manipulate data, train models, and build applications. If you're new to ML, mastering Python is the first step.

## Numpy

NumPy (Numerical Python) is a core library for scientific computing in Python. It provides support for large, multi-dimensional arrays and matrices, along with a collection of mathematical functions to operate on them efficiently. Most ML workflows involve numerical computations, making NumPy a foundational tool for model development and data manipulation.

## Pandas

Pandas simplifies data analysis by offering powerful, user-friendly data structures like Series and DataFrames. It allows you to load, clean, transform, and explore datasets with ease—tasks that are essential for building reliable ML models. Mastering Pandas helps you handle real-world data more efficiently

## Data Visualization

Visualizing your data helps you uncover hidden trends, relationships, and anomalies. Libraries like Matplotlib and Seaborn allow you to create plots and charts that make your data more understandable. Data visualization is also crucial for presenting findings and building interpretable ML pipelines.

## Data Preprocessing

Before training any model, the raw data must be cleaned and prepared. This involves handling missing values, encoding categorical features, scaling numerical values, and selecting relevant features. Proper Preprocessing ensures that models learn effectively and perform well on unseen data.

## Linear Regression

Linear regression is one of the simplest and most commonly used algorithms in supervised learning. It models the linear relationship between one or more input features and a continuous target variable. It forms the basis for more complex regression and time series models, making it an ideal starting point.

## Logistic Regression

Logistic regression is used for binary classification problems (e.g., spam or not spam). Despite its name, it's a classification algorithm that uses a logistic function to estimate the probability of a class. It’s interpretable, easy to implement, and works well for baseline models.

## Decision Trees and Random Forests

Decision Trees split data into subsets based on feature values, creating a flowchart like structure that is easy to understand and interpret. Random Forest is an ensemble of decision trees that improves prediction accuracy and reduces overfitting. These models are great for both classification and regression tasks and are widely used in industry.  You can explore this section for the implementation: Decision Trees | Random Forest

# Optional Resources

## Naive Bayes:

Naive Bayes is a probabilistic classifier based on Bayes' Theorem, with the assumption that features are conditionally independent. It’s particularly effective for text classification tasks like spam detection and sentiment analysis, due to its simplicity and performance on highdimensional data.

## K-Nearest Neighbors:

KNN is an intuitive, non-parametric algorithm used for both classification and regression. It works by comparing new data points to the ‘k’ closest training examples and predicting the majority class (or average in regression). It’s easy to implement and understand, making it a great choice for beginners.

## Evaluation Metrics & Hyperparameters:

Evaluating model performance requires the right metrics mean squared error for Regression, or accuracy, precision, recall, and F1-score for Classification. Additionally, tuning hyperparameters (like learning rate or tree depth) can significantly improve model performance. Tools like GridSearchCV help automate this process.

## Scikit-learn:

Scikit-learn is the most popular machine learning library in Python. It provides tools for every stage of an ML workflow, from preprocessing and model training to evaluation and deployment using a consistent and simple API.

Join the Whatsapp Group

Have Any Doubts:
Contact Us
Prakhar +91 9305415358
Shirshendu +91 9674397828
Manthan +91 9512529995
Arkadeb +91 9476256774
Sarvesh +91 7262006199
Pushpendra +91 7351816883
Arnav +91 7982390939
Naveen +91 8074523143
Nilay +91 9340251202