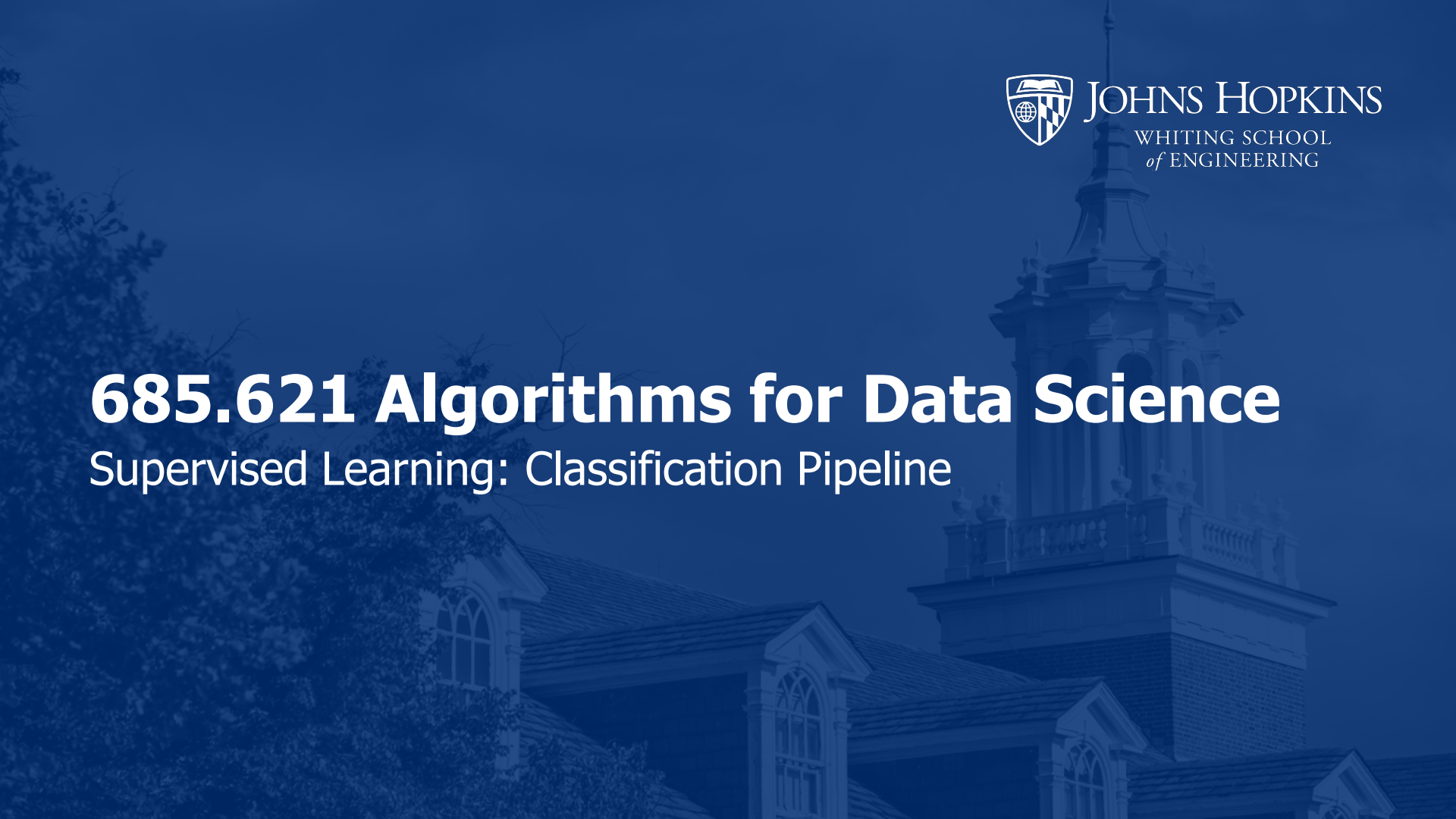
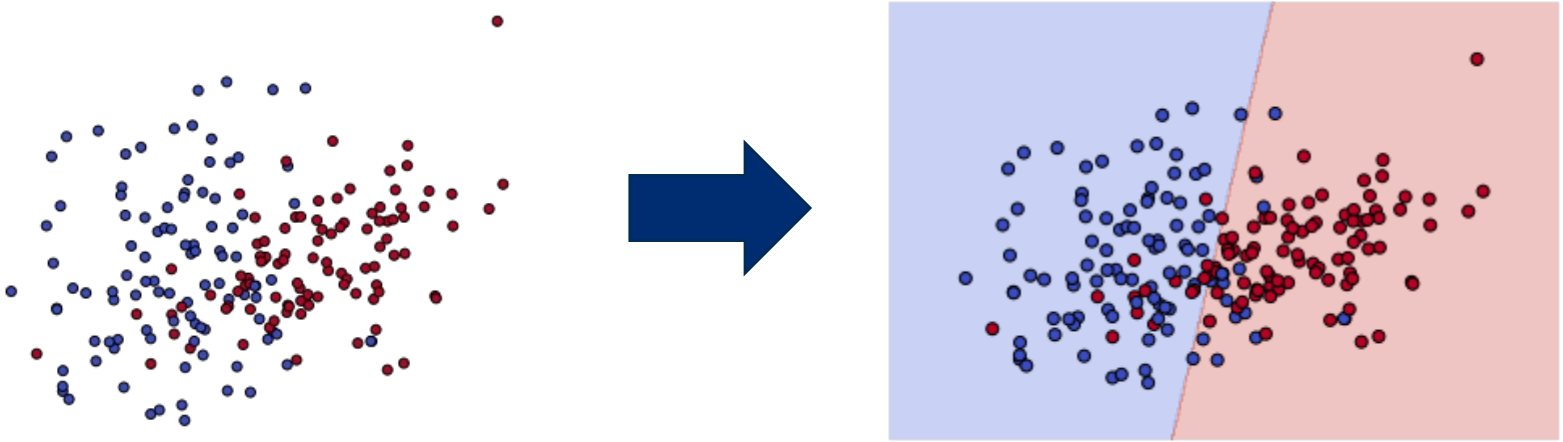


685.621 Algorithms for Data Science

Supervised Learning: Classification Pipeline

The Classification Pipeline



Step 1: Preparing the Data for Classification

Structured

- CSV, TSV, Databases

Unstructured

- Text, Images, Video

Preprocessing Tasks

- Handling **missing values** (imputation, removal)
- Encoding **categorical values** (Label encoding, One-Hot Encoding)
- Scaling **numerical features** (Standardization, Normalization)
- Balancing **imbalanced datasets** (SMOTE)

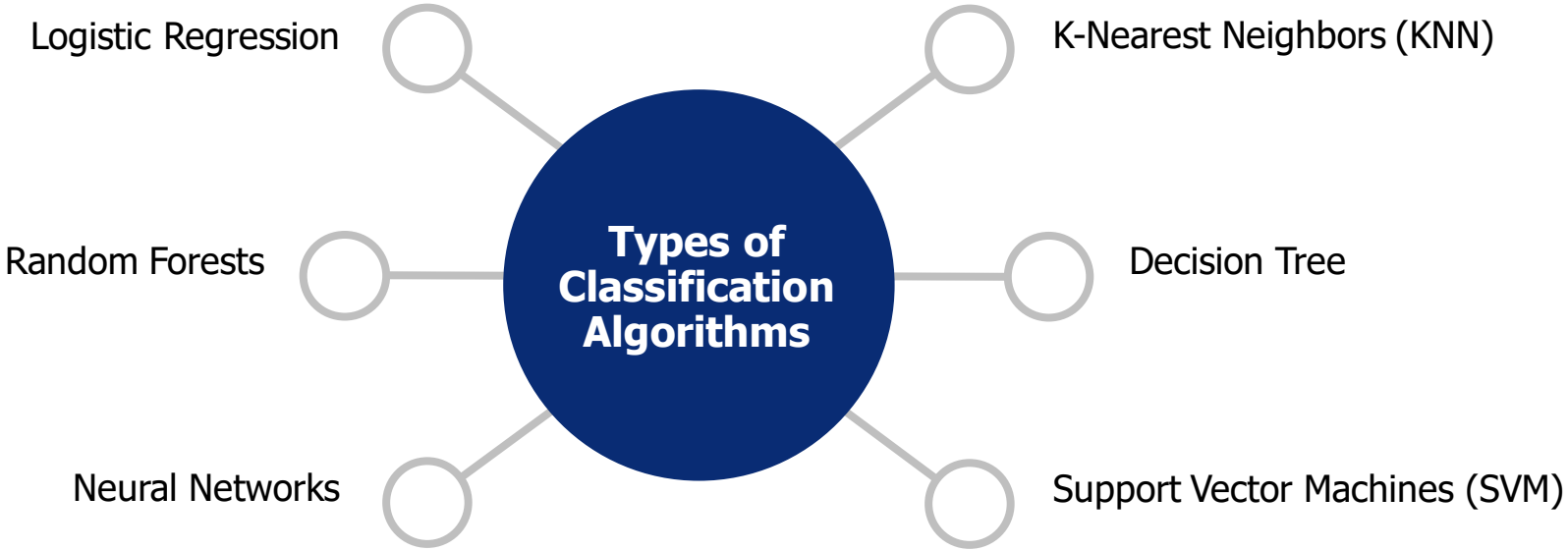
Step 2: Extract Meaningful Features

Feature Engineering – Transforming raw data into useful input features.

- **Types of Features**:
  - **Numerical Features** (e.g., Age, Salary)
  - **Categorical Features** (e.g., Gender, Product Category)
  - **Derived Features** (e.g., Total Purchase / Year, BMI)

- **Dimensionality Reduction**:
  - **Principal Component Analysis** (PCA)
  - **Feature Selection** (Eigenvalue Decomposition, Feature Importance, Fisher's Linear Discriminant Ratio)

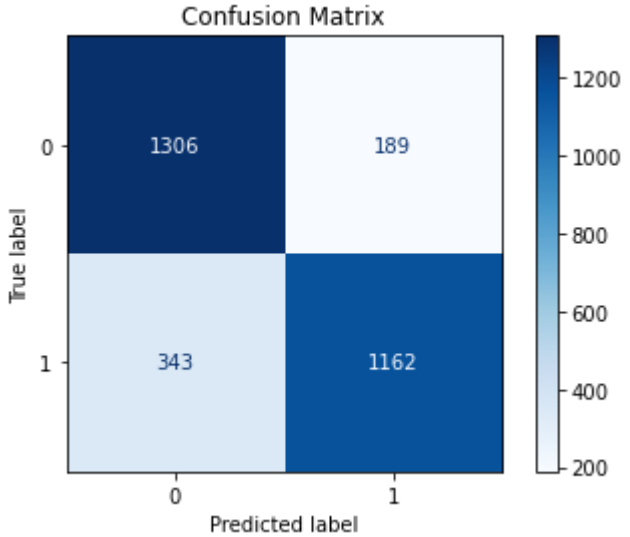
Step 3: Choosing the Right Algorithm

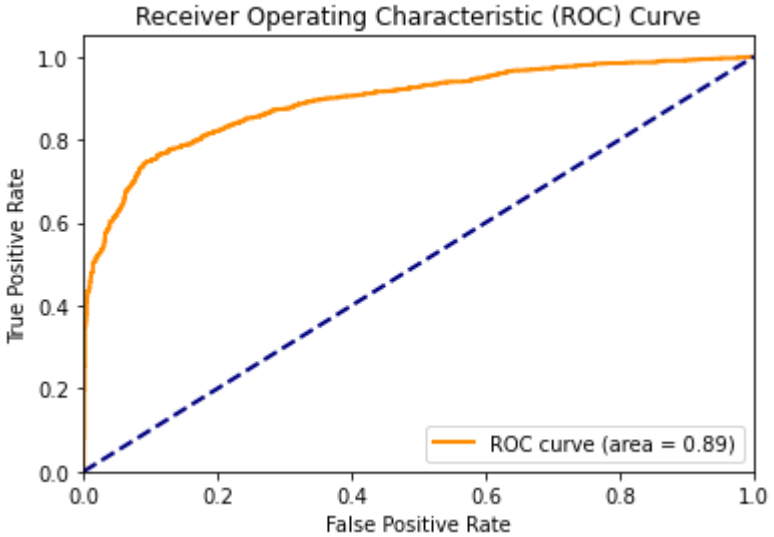


Step 4: Training & Validating the Model

K-FOLD CROSS VALIDATION	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Experiment 1	TRAIN	TRAIN	TRAIN	TRAIN	**TEST**
Experiment 2	TRAIN	TRAIN	TRAIN	**TEST**	TRAIN
Experiment 3	TRAIN	TRAIN	**TEST**	TRAIN	TRAIN
Experiment 4	TRAIN	**TEST**	TRAIN	TRAIN	TRAIN
Experiment 5	**TEST**	TRAIN	TRAIN	TRAIN	TRAIN

Step 5: Evaluation Model Performance

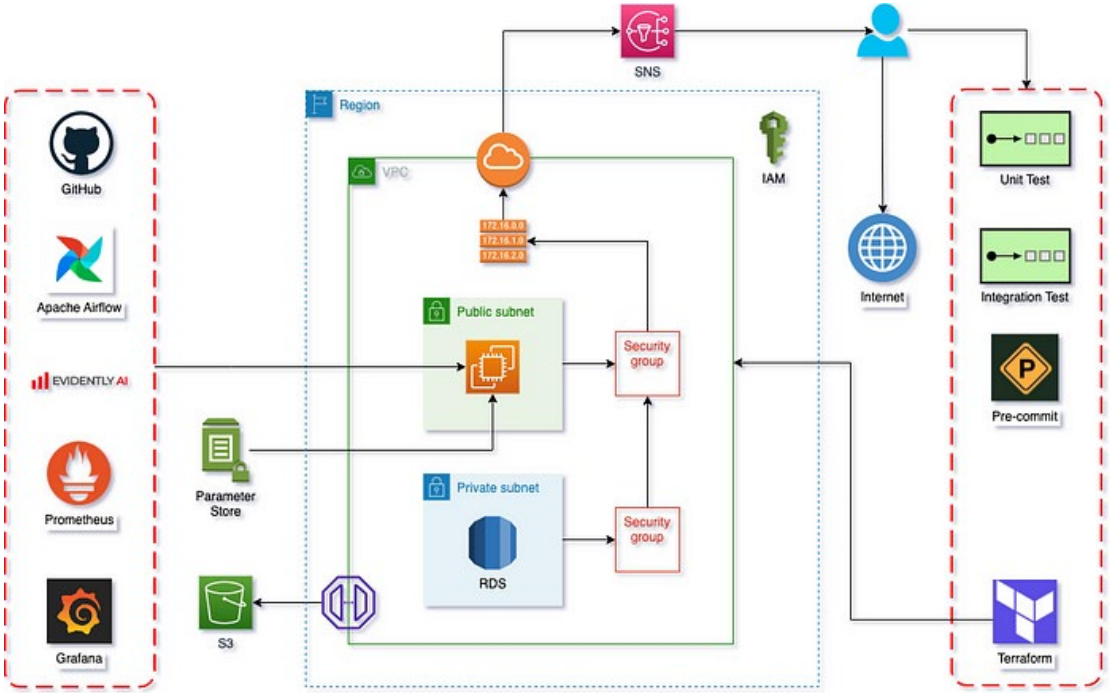




Step 6: Making Predictions & Deployment

Deployment

- Production-ready system
- Deploy Models
- Monitor Model Performance



How to Put an ML Model into Production