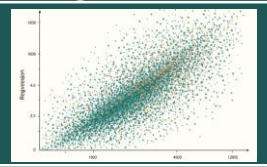
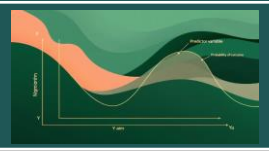
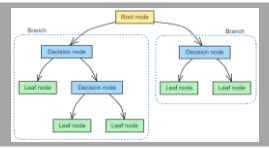
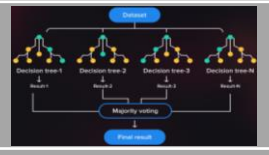
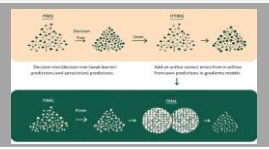
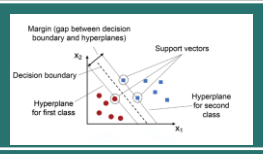
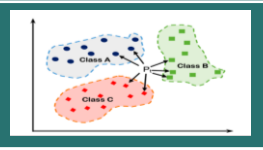
DAY6 – MACHINE LEARNING ALGORITHMS CHEAT SHEET

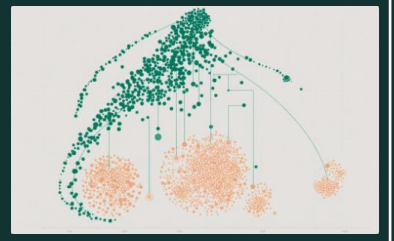
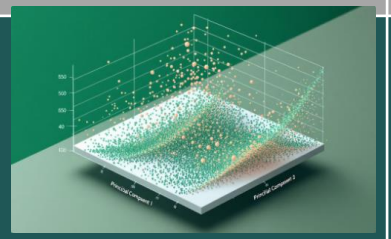
AI Advent Calendar 2024

AI Transformers

SUPERVISED MODELS

	Algorithm	Description	Applications	Advantages	Disadvantages
Regression Models	Linear Regression 	Models a linear relationship between inputs and a continuous numerical output variable.	- Stock Price Prediction - Housing Price Prediction - Customer Lifetime Value Prediction	- Simple and fast to train - Interpretable results through output coefficients - Explainable method	- Assumes linearity - Sensitive to outliers - May underfit small, high-dimensional data
	Logistic Regression 	Models a linear relationship between inputs and a categorical output (e.g., 1 or 0).	- Predicting Credit Risk Score - Customer Churn Prediction - Multi-class Predictions	- Interpretable and explainable - Regularization reduces overfitting - Handles multi-class predictions	- Assumes linearity - Can overfit with small, high-dimensional data
Regression And Classification Models	Decision Tree 	Models decisions based on feature rules; used for classification or regression.	- Customer Churn Prediction - Credit Score Modeling - Disease Prediction	- Explainable and interpretable - Handles missing values	- Prone to overfitting - Sensitive to outliers
	Random Forest 	Combines outputs of multiple decision trees for better accuracy.	- Credit Score Modeling - Predicting Housing Prices	- Reduces overfitting - Higher accuracy than single models	- High training complexity - Less interpretable than single decision trees
	Gradient Boosting Regression 	Boosts weak models to form a strong predictive ensemble.	- Predicting Car Emissions - Ride-Hailing Fare Prediction	- Better accuracy than basic regression - Handles multicollinearity and non-linear relationships	- Sensitive to outliers - Computationally expensive
	XGBoost 	Efficient and flexible gradient boosting algorithm for classification or regression.	- Churn Prediction - Insurance Claims Processing	- Highly accurate - Captures complex non-linear relationships	- Hyperparameter tuning is complex - Poor performance on sparse datasets
Classification Models	SVM (Support Vector Machine) 	A classifier that uses a linear decision boundary but can handle non-linear classification using kernel tricks to map data into high-dimensional spaces.	- Image Classification - Text Categorization - Bioinformatics	- Handles high-dimensional data well - Works when variables > samples - Flexible kernel functions	- Sensitive to overfitting - Requires careful kernel selection - Computationally expensive - Poor with noisy/overlapping data
	K Nearest Neighbours 	Predicts a label based on the closest data points in a given space; ideal for non-parametric learning where decision boundaries are irregular.	- Fraud Detection - Recommendation Systems - Medical Diagnosis	- Effective for irregular boundaries - No assumptions about data distribution	- Sensitive to noisy/missing data - Computationally expensive since it requires accessing all data points

UNSUPERVISED MODELS

	Algorithm	Description	Applications	Advantages	Disadvantages
Clustering Algorithms	K-Means 	Determines K clusters based on Euclidean distances; widely used for clustering tasks.	- Customer Segmentation - Recommendation Systems	- Scales to large datasets - Simple to implement and interpret - Creates tight clusters	- Requires predefining the number of clusters - Struggles with varying cluster sizes/densities
	Hierarchical Clustering 	"Bottom-up" approach, merging closest data points iteratively into clusters.	- Fraud Detection - Document Clustering	- No need to predefine the number of clusters - Provides an informative dendrogram	- May not yield the best clustering - High complexity; unsuitable for large datasets
Association Rule Mining: Apriori Algorithm	Apriori Algorithm 	A rule-based approach that identifies the most frequent itemsets in a dataset using prior knowledge of frequent itemset properties.	- Product Placements - Recommendation Engines - Promotion Optimization	- Intuitive and interpretable results - Exhaustively finds all rules based on confidence and support	- Generates many uninteresting or overlapping itemsets - Computationally and memory-intensive
Dimensionality Reduction	PCA (Principal Component Analysis) 	PCA reduces dimensionality by extracting key features while minimizing information loss, using linear functions.	- Image Compression - Noise Reduction - Preprocessing for Machine Learning Models	- Explainable & interpretable results - Maps new data points into existing PCA space - Helps reduce overfitting - Removes correlated features	- Sensitive to outliers - Requires data standardization