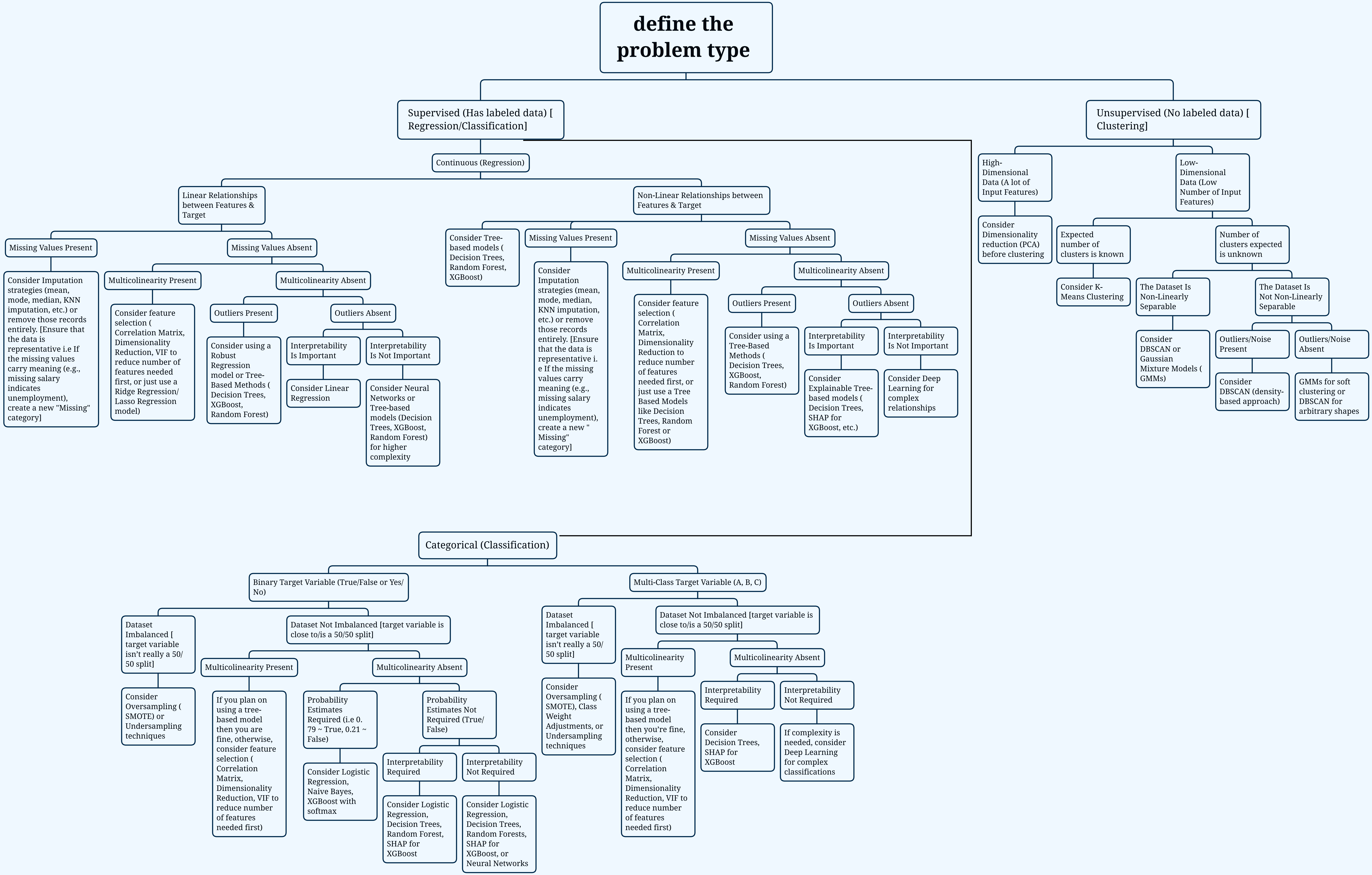
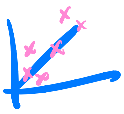

define the problem type
Supervised (Has labeled data) [Regression/Classification]
Unsupervised (No labeled data) [Clustering]
Continuous (Regression)
Linear Relationships between Features & Target
Non-Linear Relationships between Features & Target
Missing Values Present
Missing Values Absent
Consider Imputation strategies (mean, mode, median, KNN imputation, etc.) or remove those records entirely. [Ensure that the data is representative i.e If the missing values carry meaning (e.g., missing salary indicates unemployment), create a new "Missing" category]
Multicolinearity Present
Multicolinearity Absent
Consider feature selection (Correlation Matrix, Dimensionality Reduction, VIF to reduce number of features needed first, or just use a Ridge Regression/ Lasso Regression model)
Outliers Present
Outliers Absent
Consider using a Robust Regression model or Tree-Based Methods (Decision Trees, XGBoost, Random Forest)
Interpretability Is Important
Interpretability Is Not Important
Consider Linear Regression
Consider Neural Networks or Tree-based models (Decision Trees, XGBoost, Random Forest) for higher complexity
Consider Tree-based models (Decision Trees, Random Forest, XGBoost)
Missing Values Present
Missing Values Absent
Consider Imputation strategies (mean, mode, median, KNN imputation, etc.) or remove those records entirely. [Ensure that the data is representative i.e If the missing values carry meaning (e.g., missing salary indicates unemployment), create a new " Missing" category]
Multicolinearity Present
Multicolinearity Absent
Consider feature selection (Correlation Matrix, Dimensionality Reduction to reduce number of features needed first, or just use a Tree Based Models like Decision Trees, Random Forest or XGBoost)
Outliers Present
Outliers Absent
Consider using a Tree-Based Methods (Decision Trees, XGBoost, Random Forest)
Interpretability Is Important
Interpretability Is Not Important
Consider Explainable Tree-based models (Decision Trees, SHAP for XGBoost, etc.)
Consider Deep Learning for complex relationships
High-Dimensional Data (A lot of Input Features)
Low-Dimensional Data (Low Number of Input Features)
Consider Dimensionality reduction (PCA) before clustering
Expected number of clusters is known
Number of clusters expected is unknown
Consider K-Means Clustering
The Dataset Is Non-Linearly Separable
The Dataset Is Not Non-Linearly Separable
Consider DBSCAN or Gaussian Mixture Models (GMMs)
Outliers/Noise Present
Outliers/Noise Absent
Consider DBSCAN (density-based approach)
GMMs for soft clustering or DBSCAN for arbitrary shapes
Categorical (Classification)
Binary Target Variable (True/False or Yes/ No)
Multi-Class Target Variable (A, B, C)
Dataset Imbalanced [target variable isn't really a 50/ 50 split]
Dataset Not Imbalanced [target variable is close to/is a 50/50 split]
Consider Oversampling (SMOTE) or Undersampling techniques
Multicolinearity Present
Multicolinearity Absent
If you plan on using a tree-based model then you are fine, otherwise, consider feature selection (Correlation Matrix, Dimensionality Reduction, VIF to reduce number of features needed first)
Probability Estimates Required (i.e 0. 79 ~ True, 0.21 ~ False)
Probability Estimates Not Required (True/ False)
Consider Logistic Regression, Naive Bayes, XGBoost with softmax
Interpretability Required
Interpretability Not Required
Consider Logistic Regression, Decision Trees, Random Forest, SHAP for XGBoost
Consider Logistic Regression, Decision Trees, Random Forests, SHAP for XGBoost, or Neural Networks
Dataset Imbalanced [target variable isn't really a 50/ 50 split]
Dataset Not Imbalanced [target variable is close to/is a 50/50 split]
Consider Oversampling (SMOTE), Class Weight Adjustments, or Undersampling techniques
Multicolinearity Present
Multicolinearity Absent
If you plan on using a tree-based model then you're fine, otherwise, consider feature selection (Correlation Matrix, Dimensionality Reduction, VIF to reduce number of features needed first)
Interpretability Required
Interpretability Not Required
Consider Decision Trees, SHAP for XGBoost
If complexity is needed, consider Deep Learning for complex classifications


define the problem type
Supervised (Has labeled data) [Regression/Classification]
Continuous (Regression)
Linear Relationships between Features & Target
Missing Values Present
Consider Imputation strategies (mean, mode, median, KNN imputation, etc.) or remove those records entirely. [Ensure that the data is representative i.e If the missing values carry meaning (e.g., missing salary indicates unemployment), create a new "Missing" category]
Missing Values Absent
Multicolinearity Present
Consider feature selection (Correlation Matrix, Dimensionality Reduction, VIF to reduce number of features needed first, or just use a Ridge Regression/Lasso Regression model)
Multicolinearity Absent
Outliers Present
Consider using a Robust Regression model or Tree-Based Methods (Decision Trees, XGBoost, Random Forest)
Outliers Absent
Interpretability Is Important
Consider Linear Regression
Interpretability Is Not Important
Consider Neural Networks or Tree-based models (Decision Trees, XGBoost, Random Forest) for higher complexity
Non-Linear Relationships between Features & Target
Consider Tree-based models (Decision Trees, Random Forest, XGBoost)
Missing Values Present
Consider Imputation strategies (mean, mode, median, KNN imputation, etc.) or remove those records entirely. [Ensure that the data is representative i.e If the missing values carry meaning (e.g., missing salary indicates unemployment), create a new "Missing" category]
Missing Values Absent
Multicolinearity Present
Consider feature selection (Correlation Matrix, Dimensionality Reduction to reduce number of features needed first, or just use a Tree Based Models like Decision Trees, Random Forest or XGBoost)
Multicolinearity Absent
Outliers Present
Consider using a Tree-Based Methods (Decision Trees, XGBoost, Random Forest)
Outliers Absent
Interpretability Is Important
Consider Explainable Tree-based models (Decision Trees, SHAP for XGBoost, etc.)
Interpretability Is Not Important
Consider Deep Learning for complex relationships
Categorical (Classification)
Binary Target Variable (True/False or Yes/No)
Dataset Imbalanced [target variable isn't really a 50/50 split]
Consider Oversampling (SMOTE) or Undersampling techniques
Dataset Not Imbalanced [target variable is close to/is a 50/50 split]
Multicolinearity Present
If you plan on using a tree-based model then you are fine, otherwise, consider feature selection (Correlation Matrix, Dimensionality Reduction, VIF to reduce number of features needed first)
Multicolinearity Absent
Probability Estimates Required (i.e 0.79 ~ True, 0.21 ~ False)
Consider Logistic Regression, Naive Bayes, XGBoost with softmax
Probability Estimates Not Required (True/ False)
Interpretability Required
Consider Logistic Regression, Decision Trees, Random Forest, SHAP for XGBoost
Interpretability Not Required
Consider Logistic Regression, Decision Trees, Random Forests, SHAP for XGBoost, or Neural Networks
Multi-Class Target Variable (A, B, C)
Dataset Imbalanced [target variable isn't really a 50/50 split]
Consider Oversampling (SMOTE), Class Weight Adjustments, or Undersampling techniques
Dataset Not Imbalanced [target variable is close to/is a 50/50 split]
Multicolinearity Present
If you plan on using a tree-based model then you're fine, otherwise, consider feature selection (Correlation Matrix, Dimensionality Reduction, VIF to reduce number of features needed first)
Multicolinearity Absent
Interpretability Required
Consider Decision Trees, SHAP for XGBoost
Interpretability Not Required
If complexity is needed, consider Deep Learning for complex classifications
Unsupervised (No labeled data) [Clustering]
High-Dimensional Data (A lot of Input Features)
Consider Dimensionality reduction (PCA) before clustering
Low-Dimensional Data (Low Number of Input Features)
Expected number of clusters is known
Consider K-Means Clustering
Number of clusters expected is unknown
The Dataset Is Non-Linearly Separable
Consider DBSCAN or Gaussian Mixture Models (GMMs)
The Dataset Is Not Non-Linearly Separable
Outliers/Noise Present
Consider DBSCAN (density-based approach)
Outliers/Noise Absent
GMMs for soft clustering or DBSCAN for arbitrary shapes

MACHINE LEARNING MODELS

Regression Models

1. Linear Regression



Strengths:

- Simple and easy to interpret
- Computationally efficient
- Works well when relationships between features and target are linear

Weaknesses:

- Assumes linearity between features and target
- Sensitive to outliers
- Affected by multicollinearity

Assumptions:

- **Linearity:** The relationship between features and the target is linear (e.g., salary increases proportionally with years of experience)
- **Independence:** Observations are independent (e.g., predicting house prices, each observation should be a different house)
- **Homoscedasticity:** Constant variance of residuals (e.g., residuals should not increase with target values)
- **Normality:** Residuals should be normally distributed

Expected Data Processing:

- Requires **feature scaling** (Standardization or Normalization)
- Handles missing values poorly (imputation required)
- Cannot handle categorical features directly (One-Hot Encoding needed)

2. Ridge Regression (L2 Regularization) & Lasso Regression (L1 Regularization)

Strengths:

- Addresses multicollinearity
- Can perform feature selection (Lasso)
- More robust to overfitting compared to standard Linear Regression

Weaknesses:

- Requires hyperparameter tuning for optimal performance
- Lasso Regression may eliminate important features if lambda is too high

Assumptions:

- Same as Linear Regression

Expected Data Processing:

- Requires **feature scaling**
- Requires **feature selection** if dataset is high-dimensional (especially for Lasso)
- Categorical variables must be **encoded**

3. Robust Regression

Strengths:

- Works well with outliers
- More robust than standard Linear Regression

Weaknesses:

- Can be less efficient on clean datasets
- Interpretability may be slightly reduced

Assumptions:

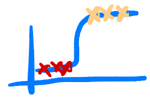
- Similar to Linear Regression but **relaxes homoscedasticity assumption**

Expected Data Processing:

- Handles missing values poorly (requires imputation)
- Requires **feature scaling**
- Categorical variables must be **encoded**

Classification Models

4. Logistic Regression



Strengths:

- Simple, interpretable, and computationally efficient
- Outputs probability estimates

Weaknesses:

- Assumes linear decision boundaries
- Can struggle with imbalanced datasets

Assumptions:

- **Linearity in log-odds:** Relationship between independent variables and log-odds of the target is linear
- **Independence of observations**

Expected Data Processing:

- Requires **feature scaling**
- Handles missing values poorly (requires imputation)
- Categorical variables must be **encoded (One-Hot Encoding)**

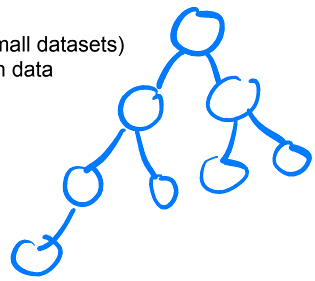
5. Decision Trees

Strengths:

- Handles both numerical and categorical features natively
- Captures non-linear relationships well
- No need for feature scaling or encoding
- Robust to class imbalance (doesn't require SMOTE/undersampling)

Weaknesses:

- Prone to overfitting (especially on small datasets)
- Can be sensitive to small changes in data



Assumptions:

- No strict statistical assumptions

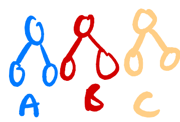
Expected Data Processing:

- Works well with missing data
- No need for feature scaling
- Handles categorical variables **natively**

6. Random Forest

Strengths:

- Reduces overfitting compared to Decision Trees
- Handles missing values well
- Works well on high-dimensional data
- No need for feature scaling or encoding



Weaknesses:

- Computationally expensive for large datasets
- Less interpretable than single Decision Trees

Assumptions:

- No strict assumptions, works well on diverse datasets

Expected Data Processing:

- Handles missing data well
- No need for feature scaling or encoding

Clustering Models

9. K-Means Clustering

Strengths:

- Simple and fast for clustering
- Scales well with large datasets

Weaknesses:

- Assumes clusters are spherical
- Sensitive to outliers and cluster initialization

Assumptions:

- **Clusters are of equal variance**
- **Euclidean distance is meaningful**

Expected Data Processing:

- Requires **feature scaling (Standardization)**
- Handles missing values poorly (requires imputation)
- Does not handle categorical variables well (requires encoding)

10. DBSCAN (Density-Based Clustering)

Strengths:

- Detects clusters of arbitrary shape
- Handles noise and outliers well

Weaknesses:

- Struggles with varying cluster density
- Sensitive to hyperparameter tuning (eps, min_samples)

Assumptions:

- **Clusters are dense regions of data**

Expected Data Processing:

- Requires **feature scaling (Standardization)**
- Handles missing values poorly (requires imputation)
- Does not handle categorical variables well (requires encoding)

11. Principal Component Analysis (PCA)

Strengths:

- Reduces dimensionality while preserving variance
- Speeds up training for high-dimensional datasets

Weaknesses:

- Hard to interpret transformed features
- Assumes linear relationships in data

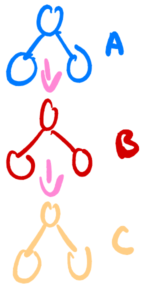
Assumptions:

- **Data is centered (mean = 0)**
- **Principal components capture most variance**

Expected Data Processing:

- Requires **feature scaling (Standardization)**
- Handles missing values poorly (requires imputation)
- Does not handle categorical variables well (requires encoding)

7. XGBoost



Strengths:

- Efficient and optimized for speed
- Handles missing values internally
- Works well with structured/tabular data

Weaknesses:

- Requires parameter tuning
- Less interpretable than simpler models

Assumptions:

- No strict statistical assumptions

Expected Data Processing:

- Works well with missing values
- No need for feature scaling

Handles categorical variables well with **Label Encoding**

8. Support Vector Machines (SVMs)

Strengths:

- Works well in high-dimensional spaces
- Effective for non-linearly separable data with kernel tricks

Weaknesses:

- Computationally expensive for large datasets
- Requires careful tuning of hyperparameters

Assumptions:

- **Linearly separable data (for basic SVMs)**

Expected Data Processing:

- Requires **feature scaling**
- Handles missing values poorly (requires imputation)
- Categorical variables must be **encoded**