Machine Learning Terminology & Jargon – PDF Notes

1. Data & Features

- Dataset – Collection of data used for ML.
- Feature (X) – Input variables.
- Target / Label (y) – Output variable to predict.
- Observation / Sample – One row of data.
- Feature Engineering – Creating new useful features.
- Feature Scaling – Standardization, Normalization.
- Categorical Features – Non-numeric values.
- Encoding – One-Hot, Label Encoding.

2. Training & Evaluation

- Train Set – Used to train model.
- Validation Set – Used for tuning.
- Test Set – Final unseen evaluation.
- Train-Test Split – Dividing dataset.
- Cross-Validation – Multiple splits for robustness.

3. Learning Types

- Supervised Learning – Labeled data.
- Unsupervised Learning – No labels.
- Semi-Supervised Learning – Few labels.
- Reinforcement Learning – Reward-based learning.

4. Classification Metrics

- Confusion Matrix – TP, FP, TN, FN.
- Accuracy – Overall correctness.
- Precision – Correct positive predictions.
- Recall – Captured actual positives.
- F1-Score – Balance of precision & recall.
- ROC Curve – TPR vs FPR.
- AUC – Area under ROC curve.

5. Regression Metrics

- Residual – Actual - Predicted.
- MSE – Mean Squared Error.
- RMSE – Root MSE.
- MAE – Mean Absolute Error.
- R^2 Score – Variance explained.

6. Bias & Variance

- Underfitting – High bias.
- Overfitting – High variance.
- Bias – Error from assumptions.
- Variance – Sensitivity to data.
- Bias-Variance Tradeoff – Balance.

7. Parameters & Hyperparameters

- Parameters – Learned by model.
- Hyperparameters – Set by user.
- Hyperparameter Tuning – GridSearch, RandomSearch.

8. Optimization

- Loss Function – Measures error.
- Cost Function – Avg loss.
- Gradient Descent – Optimization method.
- Learning Rate – Step size.

9. Regularization

- L1 (Lasso) – Feature selection.
- L2 (Ridge) – Shrinks weights.
- Elastic Net – L1 + L2.

10. Tree-Based Models

- Root Node – First split.
- Leaf Node – Final prediction.
- Gini Impurity – Node impurity.
- Entropy – Impurity measure.
- Information Gain – Entropy reduction.
- Pruning – Prevent overfitting.

11. Ensemble Learning

- Bagging – Bootstrap aggregation.
- Boosting – Sequential learning.
- Random Forest – Ensemble of trees.

12. Distance & Dimensionality

- KNN – Nearest neighbors.
- Distance Metrics – Euclidean, Manhattan, Cosine.
- Curse of Dimensionality – High-dim problems.
- PCA – Dimensionality reduction.
- Explained Variance – Info retained.

13. ML in Production

- Inference – Making predictions.
- Model Drift – Data changes.
- Data Leakage – Using future info.
- Pipeline – Preprocessing + model.
- Baseline Model – Simple comparison model.
- Explainability – SHAP, LIME.