ML for Innovation Research — Quick Reference Card

PhD/DBA Seminar — Prof. J. Osterrieder

Discovery Techniques (Unsupervised)

PCA / UMAP — Dimensionality reduction
Use: Visualize high-dimensional data in 2D
Output: Scatter plots showing structure
Key metric: Variance explained (PCA)
`sklearn.decomposition.PCA`

K-Means Clustering — Group discovery
Use: Find natural segments in data
Output: Cluster labels, profiles
Key metric: Silhouette score (–1 to +1)
`sklearn.cluster.KMeans`

VADER Sentiment — Text polarity
Use: Score text sentiment (–1 to +1)
Output: Compound score per document
Key metric: Distribution of scores
`nltk.sentiment.vader`

LDA Topic Modeling — Theme discovery
Use: Find latent topics in text corpora
Output: Topic-word distributions
Key metric: Perplexity, coherence score
`sklearn.decomposition.LatentDirichletAllocation`

Prediction Techniques (Supervised)

Random Forest — Ensemble classifier
Use: Predict outcomes, rank features
Output: Predictions, feature importance
Key metrics: Accuracy, AUC, F1-score
`sklearn.ensemble.RandomForestClassifier`

Logistic Regression — Interpretable baseline
Use: Binary classification with coefficients
Output: Probabilities, coefficients, p-values
Key metrics: Accuracy, AUC, coefficients
`sklearn.linear_model.LogisticRegression`

Generation Techniques (GenAI)

Structured Output — LLM data coding
Use: Code qualitative data at scale
Output: JSON-structured variables
Key metric: Agreement with human coding
Tool: OpenAI/Anthropic API with JSON mode

Validation Checklist

1. Train/test split or cross-validation
2. Multiple methods compared
3. Feature importance interpreted
4. Limitations discussed honestly
5. Results connected to theory
6. Code and data available
7. AI use disclosed

Key Python Libraries

`pandas`	Data manipulation
`scikit-learn`	ML algorithms
`matplotlib`	Visualization
`seaborn`	Statistical plots
`nltk`	Text processing
`gensim`	Topic modeling
`umap-learn`	UMAP projection

Decision Framework

1. Start with your **research question**
2. Discovery question → unsupervised
3. Prediction question → supervised
4. Have text data → add NLP
5. Need scale → consider GenAI
6. Always validate and compare methods