**Elements of Data Science – F2024**

**Final Review**

This is intended as a guide and is not guaranteed to be comprehensive. Material considered fair-game for the exam is anything from class.

**Intro to ML**

* “Dimensions” of ML
  + Interpretation vs. Prediction
  + Learning Paradigms (SL,UL,etc.)
  + Regression vs. Classification
  + Binary, Multiclass, Multilabel Classification
* sklearn common functions
  + .fit()
  + .predict()
  + **.**predict\_proba()

**Machine Learning Models**

* Simple Linear Regression
  + Interpreting Coefficients of OLS
  + Colinearity
* Multiple Linear Regression
* Logistic Regression
  + Concept of Gradient Descent
* k-Nearest Neighbor
* Decision Trees
* Ensembles
  + Random Forest
  + Gradient Boosting
  + Stacking
* Perceptron/Multilayer Perceptron
* Multiclass, Multilabel and One vs. Rest Classification

**Model Evaluation**

* Generalization
  + Train/Test split
  + Stratification
* Overfitting/Underfitting
  + Bias/Variance Tradeoff
* Baseline/Dummy Models
* Tuning Hyperparameters and Model Selection
  + k-Fold Cross Validation
  + Grid Search
* Metrics for Classification
  + Accuracy/Error
  + Confusion Matrix
  + Precision
  + Recall
  + F1 Score
  + ROC Curve
  + ROC AUC
* Metrics for Regression
  + R2
  + Adjusted-R2
  + Mean Squared Error
  + Root Mean Squared Error
* Regularization
  + Ridge
  + LASSO
  + ElasticNet

**Data Cleaning**

* Dealing with Duplicates
* Dealing with Missing Data
* Dummy Variables
* Rescaling
* Dealing with Skew
* Detecting/Removing Outliers

**Feature Engineering**

* Binning
* One-Hot Encoding
* Derived Features

**Joining Datasets**

* pandas df.join() and pd.merge()
* Join Types
  + LEFT
  + RIGHT
  + INNER
  + OUTER

**Dimensionality Reduction**

* Feature Selection
  + LASSO
  + Feature Importance from Tree-Based Models
  + Univariate Tests
  + Recursive Feature Selection
* Feature Extraction
  + PCA

**Sklearn Pipelines**

* .fit\_transform() on train and .transform() on test
* GridSearch on Pipelines
* ColumnTransformer

**NLP and Topic Modeling**

* What is a corpus?
* Tokens and Tokenization
* Vocabulary
* Bag Of Words Representation
* n-grams
* Term Frequency
* Document Frequency
* Stopwords
* TfIdf
* Sentiment Analysis as Classification
* Topic Modeling with Latent Dirichlet Allocation (general concept)
  + per document topic distribution
  + per topic term distribution

**Clustering**

* k-Means
  + Within Cluster Sum of Squared Distances
* Hierarchical Agglomerative Clustering
  + linkage types
  + dendrogram representation

**Recommendation Engines**

* Content-Based Filtering
* User-Based Collaborative Filtering
* Issues
* Evaluating
  + Precision and Recall at K

**Dealing with Imbalanced Data**

* Random Undersampling majority class
* Random Oversampling minority class
* Oversample Synthetic Minority Items
  + SMOTE and ADASYN (general concepts)