**MODULE 2 - EDA**

**EDA**

1. What is an EDA and why it's needed
2. What kind of analysis can be performed as part of an EDA
3. What types of features exist
4. How to detect anomalies (in univariant case)
5. How to detect gaps in data
6. How to visualize results
7. How to calculate correlation coefficients, why they are needed
8. How to calculate co-occurrences, how to visualize them

**Data Preparation**

1. Why would you need to do a DP task
2. How to deal with missing values
3. How, when and why to scale the data
4. How to prepare categorical features
5. How to use a pipeline
6. How to combine different preprocessing procedures into one pipeline.

**Dimensionality Reduction**

1. Why would you need to do DR
2. PCA
3. SVD
4. t-SNE
5. UMAP
6. Explained variance and cumulative variance.

**MODULE 3 - STATS**

**1. Basics:**

* Random variables - categorical, ordered and continuous;
* Cumulative Distribution Function;
* Probability Density Functions.

**2. Descriptive Statistics:**

* Mean, median, mode;
* Degree of freedom;
* Standard deviation, corrected standard deviation, dispersion;
* Confidence intervals, probability interval;
* Skewness, kurtosis;
* Percentile, quantiles, quartiles, interquartile range;
* Boxplot, histogram;
* Robustness, outliers;
* Maximum Likelihood Method.

**3. Sampling methods:**

* A mixture of distrubutions;
* Random sampling;
* Stratified sampling;
* Cluster sampling;
* Optimal (Neyman) sampling.
* How to estimate statistics in the case of a mixture of distributions?

**4. Scaling data:**

* Standardization;
* Mean Normalization;
* Effects for non-normalized data to some ML algorithms.

**5. Correlations:**

* Correlation;
* Dependency (causability);
* Multicollinearity;
* Pearson's;
* Cohen's coefficient (Cohen's kappa);
* Spearman's rank;
* Kendall's rank.

**6. Hypothesis testing:**

* Null hypothesis, alternative hypothesis;
* Type 1, Type 2 errors;
* P-value, significance level;
* Pre-test assumptions;
* Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors tests, QQ-plots;
* D’Agostino’s K^2, Anderson-Darling;
* T-test, Paired T-test, Welch T-test;
* 1-way ANOVA, 2-way ANOVA;
* Bonferroni correction;
* Tukey's test;
* Chi^2 test, contingency table;

**MODULE 4 - Regression**

1. Core ML concepts:

* ML model pipeline
* Train/test split
* Cross-validation
* Gradient descent
* Loss function vs quality metric
* Model validation
* Bias-variance tradeoff
* Overfitting
* Underfitting

2. Regression tasks:

* Linear regression
* Polynomial regression
* Residual analysis

3. Algorithms:

* Ridge
* Lasso
* Elastic Net

4. Metrics:

* MAE
* MSE
* RMSE
* R
* R-squared
* Adjusted R-squared

**MODULE 5 - Classification**

1. What is Binary/Multiclass/Multilabel classification
2. Support Vector Machines
   * decision boundary
   * margin / soft margin
   * kernel
   * gamma
3. Logistic regression
   * sigmoid
   * decision boundary
4. K-nearest neighbours (logic of algorithm)
5. Quality metrics
   * confusion matrix
   * accuracy
   * precision
   * recall
   * f1-score (aka f1-measure)
   * cross-entropy
   * pr-curve
   * roc-curve
   * roc-auc
   * micro/macro-metrics

**MODULE 6 – Classification 2**

1. What is NLP
2. Text processing tools (+ RegExps)
3. Support Vector Machines
   * decision boundary
   * margin / soft margin
   * kernel
   * gamma
4. Stochastic Gradient Descent
   * idea & approach
   * variations of SGD
5. Naive Bayes
   * how Bayesian approach is different from deterministic?
   * ideas
   * why is it called 'Naive'?

**MODULE 7 – Classification 3**

* **Methods**
  + LDA and QDA
    - LDA and QDA for classification
    - Shrinkage as regularization
  + Gradient boosting
    - Concept behind boosting
    - What is a weak classifier?
    - Why do we use trees for boosting?
    - How and why do class borders produced by boosted trees differ from SVM classifier?
  + Random Forest
    - What is the difference between boosting and random forest in terms of classification error?
    - How to randomize the process of trees fitting?
* **Ensembling methods**
  + What ensembling methods are there?
* **Class imbalance**
  + What is the class imbalance? Why is it a problem?
  + How to solve the imbalance problem?
* **Classification problems**
  + How does multiclass classification differ from multilabel?
  + What metrics could be used for each of the problems?
  + How do multilabel metrics differ from multiclass in general?

**MODULE 8 – Clustering & Outliers (Unsupervised learning)**

* **Clustering**
  1. Why would you need to do a clustering;
  2. KMeans;
  3. DBSCAN;
  4. Agglomerative (hierarchical) clustering;
  5. Linkage;
  6. Dendrogram plot;
  7. Silhouette coefficient;
  8. Inertia.

* **Outlier detection**
  1. Why would you need to do OD;
  2. The difference between Mahalanobis and Euclidean distances;
  3. How to use Mahalonobis rule with SciPy and Scikit-learn;
  4. LOF;
  5. Reachability distance;
  6. Local reachability density;
  7. Isolation Forest;
  8. One-Class SVM.

**MODULE 9 – Regression advanced**

1. There is clear interpretation of Ridge and Lasso regressions:
   * It is probabilistic
   * Weights of the model are limited by some distribution
   * Regularization term can be interpreted
2. A lot of models are inferred from classification ones:
   * KNN uses averaged (possibly weighted) values of neighboring objects
   * Decision Tree averages values inside leafs
   * Ensembles average values across weak classificators
   * SVM transform itself into two divisive hyperplanes

**MODULE 10 – Time Series**

1. Data types
   * Cross-sectional data
   * Time series data
   * Panel data
2. Time series decomposition
   * Trend
   * Seasonality
   * Cyclical movements
   * Noise
   * Additive model
   * Multiplicative model
3. Time series statistics
   * Autocorrelation
   * Weakly stationary process
   * Approaches to make process stationary
4. Time series smoothing
   * MA
   * Exponential smoothing
5. ARIMA models (only general idea)
6. Validation of time series
7. How use linear regression for time series data
8. Which features could be extracted from time series

**MODULE 11 – NN & CV**

1. Architecture of MLP
   * Types of layers
   * Activation function
     + Softmax
     + ReLU / Leaky ReLU
     + Sigmoid
     + TanH
     + Linear
   * NN training
     + - Forward propagation
       - Backward propagation
       - Chain rule
       - Batch-training
2. Regularization
   * L1 / L2
   * Dropout
   * Early Stopping
3. Optimization
   * Gradient Descent and Learning Rate
   * Stochastic gradient descent
   * Mini-batch gradient descent
   * Initial Learning Rate
   * Learning Rate Schedule
   * Mini-batch Size
   * Input normalization
   * Batch normalization
   * Weights initialization
4. Computer Vision
   * Convolutions
     + Convolution (for 1 channel and multichannel picture)
     + Padding
     + Striding
     + Pooling
     + Dilated Convolution
   * Why CNN is better for images than fully connected NN
   * Data Augmentation: what is it, why it is used, common techniques

**MODULE 12 – NN & NLP**

1. What is NLP?

2. Word Embeddings

* + frequency based embedding (count vector, tf-idf, co-occurence)
  + prediction base embedding (cbow, skip-gram)
  + word2vec

3. RNN, LSTM