|  |  |
| --- | --- |
| סוגי אלגוריתם |  |
| Supervised |  |
|  | Classification |
|  | Regression |
| Unsupervised |  |
|  | Clustering |
| Reinforcement |  |

|  |
| --- |
| תצוגת הנתונים |
| Features Matrix |
| Target Array |

|  |
| --- |
| Estimator |
| Estimator |
| Transformer |
| Predictors |

|  |
| --- |
| תהליכים |
| בחירת מודל |
| בחירת hyper parameters של המודל |
| סידור הנתונים במבנה שיתאים למודל |
| יישום המודל על הנתונים – fit |
| בחירת מדד למדידת ביצועי המודל – כמו ה Mse |

|  |  |  |  |
| --- | --- | --- | --- |
| Feature engineering |  |  |  |
| Feature selection |  |  |  |
|  | VarianceThreshold – הסרת מאפיינים עם שונות נמוכה |  |  |
|  | בחינות סטטיסטיות:   * SelectKBest * SelectPercentile * SelectFpr/SelectFdr | Score functions עבור הבחינות :  רגרסיה:   * f\_regression * mutual\_info\_regression   סיווג:   * Chi\_2 * F\_classif * Mutual\_info\_classif |  |
| Feature extraction | יצירת מאפיין חדש מתוך המאפיינים הקיימים |  |  |
|  |  |  |  |
| Data Preprocessing |  |  |  |
|  | Changing the representation of the data |  |  |
|  |  | LabelEncoder |  |
|  |  | OneHotEncoder |  |
|  |  |  |  |
|  | Imputation of missing data |  |  |
|  |  | SimpleImputer | Mean  Median  Most\_frequent  constant |
|  |  |  |  |
|  | Discretization |  |  |
|  |  | KBinsDiscretizer | Strategy:   * Uniform * Quantile * Kmeans   Encode:   * onehot * Onehot-dense * ordinal |
|  |  |  |  |
|  | Feature scaling |  |  |
|  |  | MInMaxScaler |  |
|  |  | StandardScaler |  |
|  |  |  |  |
| Text Pre Processing  NLTK library |  |  |  |
|  | Removing stopwords  nltk.corpus.stopwords.words('english') |  |  |
|  | Removing special characters  re.sub(<regex>) |  |  |
|  | lower casing  doc.lower.strip()  מסיר רווחים מההתחלה והסוף |  |  |
|  | Text tokenization  בעזרת הפיצול אפשר לבדוק ולהסיר את ה stop words |  |  |
|  | * Contraction expansion * Correcting spellings * Stemming חותך את הסיומות של המילים * Lemmatization מוציא שורש |  |  |
|  | Bag of words / Bag of N-Gram  CountVectorizer Class | רוחב המטריצה הוא כגודל כמות המילים  Get\_feature\_names() – מציג את הפלט כ data frame  N – gram יכול להיות bi-gram, tri-gram...  ngram\_range – מציין את הטווח הנדרש |  |
|  | TF-IDF Model |  |  |

|  |  |
| --- | --- |
| TestSet |  |
| train\_test\_split |  |
| StratifiedKFold |  |

|  |  |
| --- | --- |
| Pipeline |  |
| make\_pipeline |  |
| ColumnTransformer | יכול לאחד מספר pipelines, וכך לאפשר טרנספורמציה שונה לכל עמודה |
| CustomTransformer | Implements from TransformerMixin or BaseEstimator (added set & get params methods) |

|  |  |
| --- | --- |
| Validation |  |
| cross\_val\_score function |  |
|  | cv – ניתן לשלוט על ה split  StratifiedKFold - Stratified sampling  Score='neg\_mean\_square\_error' |

|  |  |  |
| --- | --- | --- |
| Fine Tuning The Model |  |  |
| GridSearchCV |  |  |
|  | estimator  param\_grid  scoring  cv  refit | אפשר לשמור את המודל הנבחר |
| RandomizedSearchCV |  |  |

|  |
| --- |
| Evaluating The Test Set |
| final\_prediction = Final\_model.predict(test) |
| acuuracy\_score(y\_test,final\_prediction) |

|  |
| --- |
| The Bias-Variance Tradeoff |
| Underfit the data <-> the model has high bias  Overfit the data <-> the model has high variance |
| במילים אחרות:  למודלים עם high bias , ביצועי המודל על ה validation set דומים לאלו שעל ה training set  למודלים עם שונות גבוהה, ביצועי המודל על ה validation set רחוקים מאד מאלו שעל ה training set |
| Validation curve |
| המטרה למצוא את האיזון בין ה Bias ל variance |

|  |
| --- |
| The Bias-Variance Tradeoff |
| Underfit the data <-> the model has high bias  Overfit the data <-> the model has high variance |
| במילים אחרות:  למודלים עם high bias , ביצועי המודל על ה validation set דומים לאלו שעל ה training set  למודלים עם שונות גבוהה, ביצועי המודל על ה validation set רחוקים מאד מאלו שעל ה training set |
| Validation curve |
| המטרה למצוא את האיזון בין ה Bias ל variance |

|  |
| --- |
| מודלים מסוג linear Regression |
|  |
|  |
|  |
|  |