Terminology

This document's purpose is to get a clear vision on what the meaning of the terminology that is used for the project.

Term	Description
Accuracy	Percentage of correct predictions made by the model. *TP + TN / all observations (N)*
Agora	Online dark web marketplace operating on the Tor network.
Alternative hypothesis (H_1)	The observed phenomenon is the result of a non-random cause
Artificial intelligence	The Simulation of human intelligence by machines, such as abstraction, learning or problem solving.
Back propagation	A learning algorithm used by neural networks to compute a gradient descent with respect to weights.
Bag of words	The bag-of-words model is a way of representing text data when modeling text with machine learning algorithms. Any information about the order or structure of words in the document is discarded.
Baseline	The result of a very basic model/solution.
Baseline testing	The validation of documents and specifications in comparison to the baseline.
Batch Size	Total number of training examples present in a single batch.
Bayes' theorem	A theorem that describes the probability of an event, based on prior knowledge of conditions that might be related to the event.

Bias	The simplifying assumptions made by a model to make the target function easier to learn.
Bias-variance trade-off	The property of a set of predictive models whereby models with a lower bias in parameter estimation have a higher variance of the parameter estimates across samples, and vice versa.
BibTeX	A reference management software for formatting lists of references, often used in combination with LaTeX.
Bi-variate analysis	The simultaneous analysis of two variables (attributes). It explores the concept of relationship between two variables, whether there exists an association and the strength of this association, or whether there are differences between two variables and the significance of these differences.
Chi squared test	A test intended to test how likely it is that an observed distribution is due to chance.
Classification	Find a function that optimally estimates to which class a data point belongs.
Classification Threshold	The lowest probability value at which we are comfortable asserting a positive classification. *E.g. If the predicted probability of being diabetic is > 50%, return True, otherwise return False.*
Clustering	Unsupervised grouping of data into buckets.
Codebook	A document describing the contents, structure, and layout of a data collection.
Confusion matrix	A table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.

Convergence	A state reached during the training of a model when the loss changes very little between each iteration.
Convex function	A real-valued function defined on an n-dimensional interval is called convex if the line segment between any two points on the graph of the function lies above or on the graph.
Convolutional neural network	A Deep Learning algorithm which can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other.
Dark web	A collection of websites that exist on an encrypted network that cannot be found using traditional search engines or by the usage of browsers.
Data munging	The initial process of refining raw data into content or formats better-suited for consumption by downstream systems and users.
Data preprocessing	A data mining technique that involves transforming raw data into an understandable format.
Dataset	A collection of related sets of information that is composed of separate elements but can be manipulated as a unit by a computer.
Decision boundary	The region of a problem space in which the output label of a classifier is ambiguous. Maximize $P(y = 1\|X_1)$ and $P(y = 0\|X_0)$
Decision trees	A flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes).

Deduction	A top-down approach to answering questions or solving problems. A logic technique that starts with a theory and tests that theory with observations to derive a conclusion. *E.g. We suspect X, but we need to test our hypothesis before coming to any conclusions.*
Deep Learning	A subset of machine learning in artificial intelligence (AI) that has networks capable of learning unsupervised from data that is unstructured or unlabeled.
Deep web	Parts of the World Wide Web whose contents are not indexed by standard web search-engines.
Deep web	The part of the internet whose contents are not indexed by standard web search-engines.
Doc2Vec	An unsupervised algorithm to generate vectors for sentence/paragraphs/documents.
Docker	A tool designed to make it easier to create, deploy, and run applications by using containers.
Effectiveness	The degree to which a learned model is successful in producing a desired result.
Efficiency	Computation time, memory use, etc.
Ensemble Learning	This approach allows improvement of machine learning results by combining several models for the production of better predictive performance compared to a single model.
Epoch	When an entire dataset is passed forward and backward through the neural network only once.
Evaluation metric	Metrics used to measure the quality of the statistical or machine learning model.

Extrapolation	Making predictions outside the range of a dataset. *E.g. My dog barks, so all dogs must bark. In machine learning we often run into trouble when we extrapolate outside the range of our training data.*
F1-score	A weighted average of the precision and recall.
False negative	When actual class is yes but predicted class in no.
False positive	When actual class is no and predicted class is yes.
Fasttext	A library for learning of word embeddings and text classification created by Facebook's AI Research lab.
Feature	With respect to a dataset, a feature represents an attribute and value combination.
Feature scaling	A technique to standardize the independent features present in the data in a fixed range.
Generalization	How well the concepts learned by a machine learning model generalizes to specific examples or data not yet seen by the model.
Gradient descent	It is an iterative optimization algorithm used in machine learning to find the best results (minima of a curve). The algorithm is iterative means that we need to get the results multiple times to get the most optimal result. The iterative quality of the gradient descent helps a under-fitted graph to make the graph fit optimally to the data. Parameters: learning rate α
Hold out	Splitting up the dataset into a ‘train’ and ‘test’ set.

Homoscedasticity	The variance around the regression line is the same for all values of the predictor variable (X).
Hyperparameters	Hyperparameters are higher-level properties of a model such as how fast it can learn (learning rate) or complexity of a model. The depth of trees in a Decision Tree or number of hidden layers in a Neural Networks are examples of hyper parameters.
Imputation	Replacing the missing values with an estimate.
Instance	A data point, row, or sample in a dataset.
Intercept	Point where line hits an axis.
Iterations	Iterations is the number of batches needed to complete one epoch.
K-fold Cross Validation	Spread the test/train sets over multiple K-folds of the total available data and then conduct experiments with those folds.
K-nearest neighbours	A simple, supervised machine learning algorithm that can be used to solve both classification and regression problems.
Kurtosis	A measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution.
Latent Dirichlet allocation (LDA)	A “generative probabilistic model” of a collection of composites made up of parts. Its uses include Natural Language Processing and topic modelling, among others.
LaTeX	A document preparation system.
Learning curve	A plot that shows time or experience on the x-axis and learning or improvement on the y-axis.

Learning rate	The size of the update steps to take during optimization loops like Gradient Descent. - With a high learning rate we can cover more ground each step, but we risk overshooting the lowest point since the slope of the hill is constantly changing. - With a very low learning rate, we can confidently move in the direction of the negative gradient since we are recalculating it so frequently. A low learning rate is more precise, but calculating the gradient is time-consuming, so it will take us a very long time to get to the bottom.
Least Squares Cost Function	Refers to the formula used as a measure of how well the computer generated line fits the data. Used in regression analysis.
Lemmatization	The process of grouping together the inflected forms of a word so they can be analyzed as a single item, identified by the word's lemma, or dictionary form.
Linear regression	A linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables).
Local minima	A point in the domain of a function that evaluates to a greater value at every other point in a neighborhood around the local minimum (a neighborhood in this case can correspond to a “ball” around the minimum) than the local minimum itself.
Logistic regression	A statistical model that in its basic form uses a logistic function to model a binary dependent variable.
Long short-term memory network (LSTM)	A modified version of recurrent neural networks, which makes it easier to remember past data in memory.
Loss/cost function	Description of how much error/difference there is between the estimated values and the real values (low is better than high).

Machine Learning	Scientific study of algorithms and statistical models that computers use to perform a specific task without using explicit instruction, relying on learned patterns and inference functions instead.
Mean	Given by the total of the values of the samples divided by the number of samples, also ‘average’.
Mean Absolute Error (MAE)	The average of all absolute errors.
Mean Squared Error (MSE)	A measure of how close a fitted line is to data points.
Median	Sort the values and take the middle value.
Memorization	The process of committing something to memory.
Mode	Represents the most common value in a data set.
Multivariate regression	A method used to measure the degree at which more than one independent variable (predictors) and more than one dependent variable (responses), are linearly related.
Naïve Bayes	A probabilistic machine learning model that’s used for classification task. The crux of the classifier is based on the Bayes theorem.
NaN	Short for ‘Not a number’.
Natural Language Processing (NLP)	The application of computational techniques to the analysis and synthesis of natural language and speech.
Natural Language Tool Kit (NLTK)	A suite of libraries and programs for symbolic and statistical natural language processing.
Neural Network	A computer program that mimics the brains functions.

N-gram	A contiguous sequence of n items from a given sample of text or speech.
Noise	The irrelevant information or randomness in a dataset.
Non-convergence	The infinite sequence of the partial sums of the series does not have a finite limit.
Normal equation	An analytical approach to Linear Regression with a Least Square Cost Function. We can directly find out the value of θ without using Gradient Descent.
Normality test	Used to determine if a data set is well-modeled by a normal distribution and to compute how likely it is for a random variable underlying the data set to be normally distributed.
NULL hypothesis (H_0)	The observed phenomenon is a result of chance.
NumPy	A library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.
Occam's razor	The law of parsimony (Occam's razor) defines that when there are multiple possibilities to something, the simplest option (with the least number of changes/guesses/assumptions) is most likely to be the best.
Oscillation	movement back and forth in a regular rhythm, happens when a model never meets the optimum and keeps bouncing out.
Outlier	An observation that deviates significantly from other observations in the dataset.
Overfitting	Model that fits the training data too well and generalizes bad. Reason for this is high variance.

Paired T-test	Compares two means that are from the same individual, object, or related units.
Pipeline	The overall step by step process towards obtaining, cleaning, visualizing, modeling, and interpreting data within a business or group.
Polynomial regression	A form of regression analysis in which the relationship between the independent variable x and the dependent variable y is modelled as an nth degree polynomial in x.
Precision	Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. Ability of a classification model to return only relevant instances. *TP / total predicted positive (TP + FP)*
Random Forest	A data construct applied to machine learning that develops large numbers of random decision trees analyzing sets of variables. This type of algorithm helps to enhance the ways that technologies analyze complex data.
Recall	Recall is the ratio of correctly predicted positive observations to the all observations in actual class. Ability of a classification model to identify all relevant instances. *TP / total actual positive (TP + FN)*
Recurring neural network (RNN)	A generalization of feedforward neural network that has an internal memory. The network remembers the past and it's decisions are influenced by what it has learnt from the past.
Regression	A statistical measurement to determine the strength of the relationship between one dependent variable (y) and a series of other changing variables (X).

Regression Imputation	Learn a model (using the other features available within the df) and replace the missing values with an estimate (e.g. average, median, mean, etc.).
Regularization	The process of adding information in order to solve an ill-posed problem or to prevent overfitting.
Residual error	The difference between a group of values observed and their arithmetical mean.
Root Mean Square Error (RMSE)	The standard deviation of the residuals (prediction errors), the measure of how well a regression line fits the data points.
Sanity check	A basic test to quickly evaluate whether a claim or the result of a calculation can possibly be true.
Scaling out	Adding more components in parallel to spread out a load.
Scaling up	Making a component bigger or faster so that it can handle more load.
Sigmoid	A mathematical function having a characteristic "S"-shaped curve or sigmoid curve.
Skewness	A measure of symmetry, or more precisely, the lack of symmetry.
Slope	Steepness of the learning rate/line.
Softmax	A function that takes as input a vector of K real numbers, and normalizes it into a probability distribution consisting of K probabilities proportional to the exponentials of the input numbers.
Statistical Significance	The likelihood that a relationship between two or more variables is caused by something other than chance.

Stemming	Removing the suffix from a word and reduce it to its root word.
Stochastic gradient descent	An iterative method for optimizing an objective function with suitable smoothness properties.
Student's T-test	A method of testing hypotheses about the mean of a small sample drawn from a normally distributed population when the population standard deviation is unknown.
Supervised Machine Learning	The machine learning task of learning a function that maps an input to an output based on example input-output pairs.
Support vector machine (SVM)	Supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis.
T-distributed stochastic neighboring embeddings **(T-SNE/TSNE)**	An algorithm for dimensionality reduction that is well-suited to visualizing high-dimensional data.
Tensors	The mathematical term for multidimensional arrays.
TF-IDF	Term frequency inversed document frequency.
Theta (q)	The weights assigned to a particular feature.
Tor network	A free and open-source software for enabling anonymous communication.
True negative	These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no.
True positive	These are the correctly predicted positive values which means that the value of actual class is yes and the value of predicted class is also yes.

Underfitting	A model that can neither model the training data nor generalize to new data. Reason for this is a high bias.
Univariate analysis	The simplest form of analyzing data. “Uni” means “one”, so in other words your data has only one variable. It doesn’t deal with causes or relationships (unlike regression) and it’s major purpose is to describe; it takes data, summarizes that data and finds patterns in the data.
Unsupervised Machine Learning	A type of self-organized Hebbian learning that helps find previously unknown patterns in data set without pre-existing labels.
Validation	The process of evaluating a trained model on test data set.
Variable	Independent variables (also referred to as Features) are the input for a process that is being analyzes. Dependent variables are the output of the process.
Variance	The amount that the estimate of the target function will change if different training data was used.
Vectorization	The process of converting NLP text into numbers.
Wilcoxon signed rank test	A non-parametric statistical hypothesis test used to compare two related samples, matched samples, or repeated measurements on a single sample to assess whether their population mean ranks differ (i.e. it is a paired difference test).
Word Embedding	The collective name for a set of language modeling and feature learning techniques in NLP.
Word2Vec	A group of related models that are used to produce word embeddings. These models are shallow, two-layer neural networks that are trained to reconstruct linguistic contexts of words.