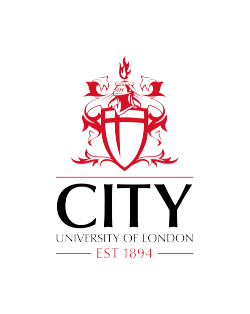
****

**Glass Identification Given Oxide Content: A Comparative Study Of Random Forests And Decision Trees**

**Supplementary Material**

Ayliah Fani

City, University of London Department of Computer Science

# Glossary

|  |  |
| --- | --- |
| **Accuracy** | The fraction of predictions that a classification model got right. In multi-class classification, accuracy is defined as follows: |
| **Class** | One set of enumerated target values for a label. |
| **Classification model** | A type of machine learning model for distinguishing among two or more discrete classes. |
| **Confusion matrix** | An NxN table that summarizes how successful a classification model's predictions were; that is, the correlation between the label and the model's classification. One axis of a confusion matrix is the label that the model predicted, and the other axis is the actual label. N represents the number of classes. |
| **Continuous feature** | A floating-point feature with an infinite range of possible values. |
| **Cross-validation** | A mechanism for estimating how well a model will generalize to new data by testing the model against one or more non-overlapping data subsets withheld from the training set. |
| **Dataset** | A collection of examples |
| **Decision tree** | A model represented as a sequence of branching statements. |
| **Discrete feature** | A **feature** with a finite set of possible values. |
| **Ensemble** | A merger of the predictions of multiple **models**. |
| **Example** | Row of a dataset, containing one or more features/predictors and a response/class |
| **Feature** | An input variable used in making predictions |
| **Holdout data** | Examples intentionally not used ("held out") during training. The validation dataset and test dataset are examples of holdout data. Holdout data helps evaluate your model's ability to generalize to data other than the data it was trained on. The loss on the holdout set provides a better estimate of the loss on an unseen dataset than does the loss on the training set. |
| **Hyperparameter** | The "knobs" that you tweak during successive runs of training a model. |
| **Label** | In supervised learning, the "answer" or "result" portion of an example. Each example in a labelled dataset consists of one or more features and a label. |
| **loss** | A measure of how far a model's predictions are from its label. Or, to phrase it more pessimistically, a measure of how bad the model is. To determine this value, a model must define a loss function. |
| **Loss curve** | A graph of loss as a function of training iterations. The loss curve can help you determine when your model is **converging**, **overfitting**, or **underfitting**. |
| **Machine Learning** | A program or system that builds (trains) a predictive model from input data. The system uses the learned model to make useful predictions from new (never-before-seen) data drawn from the same distribution as the one used to train the model. Machine learning also refers to the field of study concerned with these programs or systems. |
| **Model** | The representation of what a machine learning system has learned from the training data. |
| **Multi-class Classification** | Classification problems that distinguish among more than two classes. |
| **Overfitting** | Creating a model that matches the training data so closely that the model fails to make correct predictions on new data. |
| **Precision** | A metric for classification models. Precision identifies the frequency with which a model was correct when predicting the positive class. That is: |
| **Prediction** | A model’s output when provided with an input example |
| **Random Forest** | An ensemble approach to finding the decision tree that best fits the training data by creating many decision trees and then determining the "average" one. The "random" part of the term refers to building each of the decision trees from a random selection of features; the "forest" refers to the set of decision trees. |
| **Supervised Machine Learning** | Training a model from input data and its corresponding labels. |
| **Test Set** | The subset of the dataset that you use to test your model after the model has gone through initial vetting by the validation set. |
| **Training** | The process of determining the ideal parameters comprising a model |
| **Training set** | The subset of the dataset used to train a model |
| **Validation** | A process used, as part of training, to evaluate the quality of a machine learning model using the validation set. Because the validation set is disjoint from the training set, validation helps ensure that the model’s performance generalizes beyond the training set. |
| **Validation set** | A subset of the dataset—disjoint from the training set—used in validation. |

Definitions taken from Google Developer’s Machine Learning Glossary [1]

# Intermediate Results

# Implementation Details

* intermediate results including any negative results (5%),
* implementation details including a brief description of main implementation choices (5%).

# Bibliography

[1] ‘Machine Learning Glossary’, *Google Developers*. https://developers.google.com/machine-learning/glossary (accessed Dec. 08, 2020).