Name: **Areej Imran**

Student ID: **23033267**

Introduction

What is Cross-Validation in Machine Learning?

Cross-validation is a statistical technique for estimating the accuracy of a machine learning model and generalizability. Compared to one train-test split, cross-validation splits the data into multiple subsets and trains and validates the model on different data splits. Repeated validation approximates the consistency of the model and avoids the dangers of overfitting.

Why Use Cross-Validation?

There is not much data in real-world scenarios, noisy, or unbalanced. A single train-validation split can be deceptive in performance metrics if the validation set itself is not generalizable to the actual distribution. Cross-validation addresses this by using a lot of train-validation splits so every point is trained on and tested from.

Adding K-Fold and Stratified K-Fold

The most popular method is K-Fold Cross-Validation. It divides the dataset into k almost equal-sized groups and trains on k-1 and tests on the last group in one iteration. This, however, introduces class ratio bias when used on imbalanced classification datasets. Stratified K-Fold does not have this problem as it keeps a constant ratio of classes in each group, which leads to balanced and improved performance measures.

What This Tutorial Covers

This tutorial provides an extensive comparison of the K-Fold and Stratified K-Fold techniques. We can observe how they are different from code, plots, and validation metrics, what are their most notable differences, and how they influence model validation, especially on real-world datasets with class imbalance.

Understanding K-Fold Cross-Validation

How K-Fold Splits the Dataset

K-Fold Cross-Validation is one effective method of testing the performance of a machine learning algorithm on new, unseen data. It begins by splitting the dataset into k equally sized subsamples, which are known as folds, by splitting it randomly. The validation set is one fold, and the other k-1 folds are used in training the model. This process is carried out k times, with each fold taken in turn as the validation set. Then all the performance measures (e.g., accuracy or F1-score) are averaged to provide a final representation of the model performance.

For example, for 5-Fold Cross-Validation, the data are split into five groups. Four of these are used for training and one for testing. This is repeated five times in a manner such that each data point is used once for training and testing.

Why K-Fold Allows Us to Test Generalization

One of the primary goals of machine learning is to learn models that generalize—i.e., do sensibly well on new, unobserved data. K-Fold allows us to try out a model on an astronomically large number of different splits of data, which has the effect of revealing its average performance. This avoids the risk of over-optimistic or over-pessimistic performance estimates from one split, perhaps unbalanced.

The Impact of Shuffling and Rotation of Fold

It is generally pre-shuffled in advance prior to dividing the data. Shuffling avoids any hidden pattern or ordering that will bias the folds (e.g., ordered classes). In cross-validation iteration, each fold cycles through its index—each fold is used once as the validation set and k-1 times as training. Cycling allows unbiased and complete use of the entire dataset and thus increases the robustness of model estimation.

Visual Aid: Interpreting the 5-Fold Cross-Validation Chart

The figure shows 5-Fold Cross-Validation. The data is split into five equal folds, and one row is used for one run. One of the folds reds is used as a validation set and the rest of the folds green are used as training sets for each run. It is repeated five times, and hence each fold is used only once as a validation set. Therefore, every data point is used once for testing and four times for training. This looping improves equity of evaluation and prevents overfitting. Final model performance is estimated by average over all iterations. This procedure gives stable and consistent model estimation, especially when the datasets are small. In most of our experiments, this procedure was used.

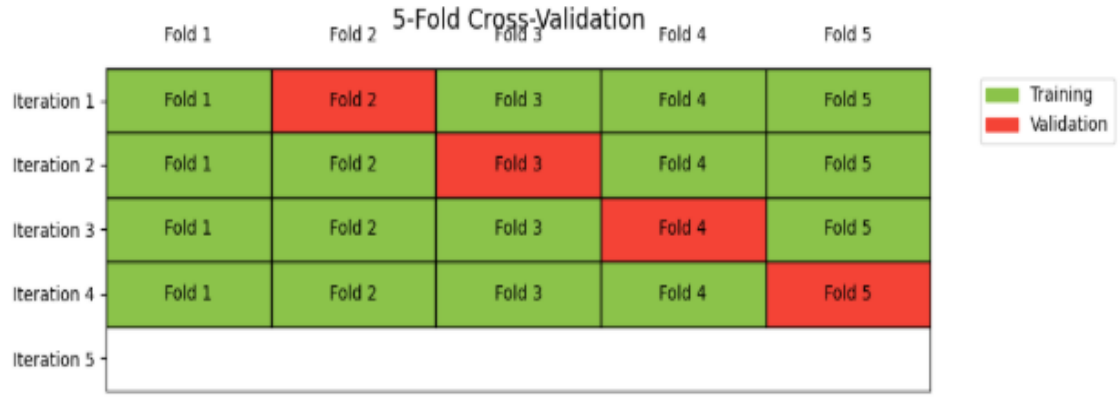

Fig1: visualization of 5-fold cross-validation process

The Issue of Class Imbalance

In the majority of real-world classification tasks, datasets are imbalanced and one class contains significantly larger instances than other classes. For instance, in a medical dataset, 95% of patients are healthy (negative class) while only 5% are ill (positive class). The imbalance renders model training and model evaluation a critical issue. If the model is trained and tested on non-

representative data of the true class distribution, the performance metrics—particularly on the minority class—can be misleading.

Why Regular K-Fold Fails

K-Fold Cross-Validation divides data into k equally sized chunks (folds) and iterates through which fold to validate on. It does so randomly, though, with no regard for class labels. For imbalanced datasets, that random split might leave folds that are extremely unbalanced in their class distribution, or even completely empty of minority class samples. This leads to imbalanced performance measures, especially for measures like precision, recall, and F1-score, which are class imbalance sensitive.

Formally, a class distribution of a dataset is given by:

$$\boldsymbol{P(c_i) = \frac{n_i}{N}}$$

Where:

- $\boldsymbol{P(c_i)}$ is the proportion (or probability) of class Ci
- ni is the number of samples in class Ci
- N is the total number of samples in the dataset

What is Stratified K-Fold?

Stratified K-Fold Cross-Validation succeeds the standard K-Fold through the preservation of the initial class balance in all folds. Each training and validating set gets an equilibrated set of classes to the general set. It is particularly important in binary and multi-classification where accuracy across each class is crucial.

Under Stratified K-Fold, in each fold Fi, the percentage of classes gets very close to the overall structure:

$$\boldsymbol{P_{f_j}(c_i) \approx P(c_i)}$$

Where:

- $\boldsymbol{P_{f_j}(c_i)}$ is the proportion of class Ci in fold Fi
- The symbol ≈ indicates that the proportion in the fold is approximately equal ti that of the dull dataset

This restriction renders each fold a small replica of the original information, thus measurement metrics are more accurate and consistent.

Benefits of Stratification

Using Stratified K-Fold provides better performance scores, lower fold variance, and better generalization estimates. It keeps minority classes adequately represented in training and validation in order to avoid artificially high accuracy by class dominance.

These are the exact same reasons that libraries like Scikit-learn apply Stratified K-Fold as the default strategy for classification models. It enhances model validation, enhances fairness, and enhances overall robustness in measurement of performance.

Description of Block Grid Visualization

This block grid plot shows how class samples are distributed into validation folds in Regular K-Fold and Stratified K-Fold cross-validation. One row per fold, one block per row per sample, colored by class: green for Class 0, red for Class 1. In Regular K-Fold (left), there is class variability between folds—some rows have very few or no red blocks, indicating imbalance. In Stratified K-Fold (right), every row is having a balanced ratio of class colors. This is beautifully showing us how stratification is making every fold a representation of the entire dataset, making the evaluation fairer and more believable.

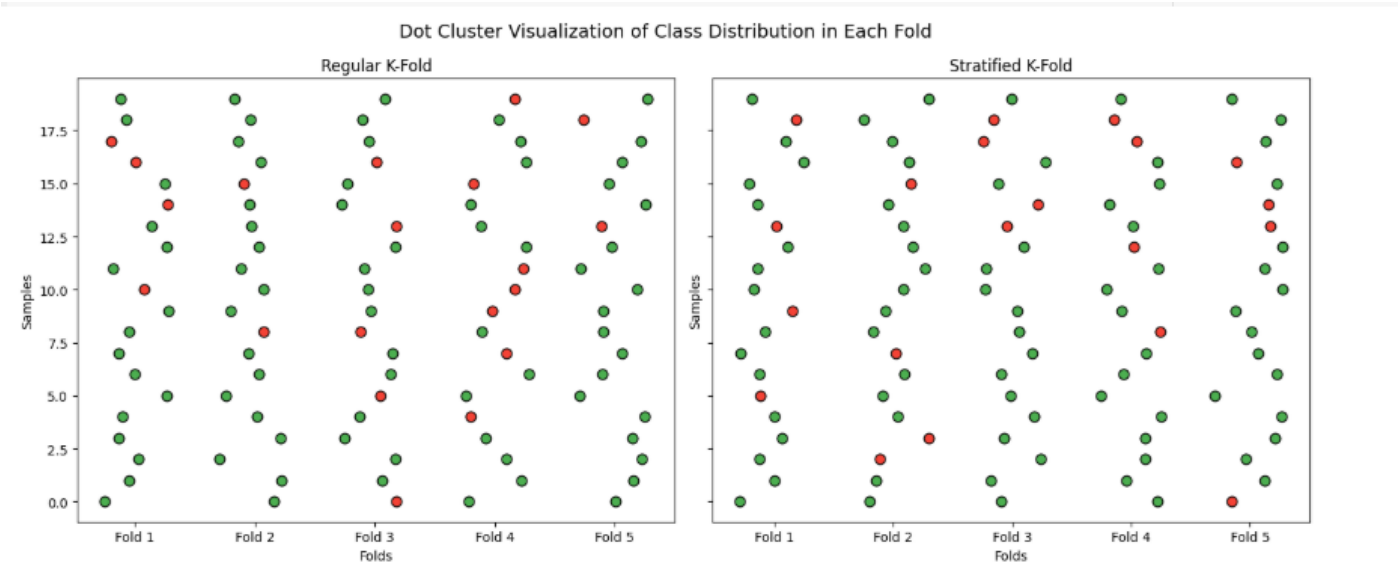

Fig2: Dot cluster comparison of class distribution across folds.

Description of the Breast Cancer Dataset

The Breast Cancer Wisconsin dataset is among the most popularly used binary classification datasets on which one can predict whether a tumor is malignant or benign given features extracted through digitizing images of fine needle aspirates (FNA) of breast masses. It has 569 examples consisting of 30 numeric attributes like radius, texture, perimeter, area, and smoothness. The target variable is a little imbalanced with roughly 63% benign and 37% malignant cases. The data set is therefore best suited for experimentation with the effect of cross-validation techniques on model performance, especially concerning the effect of class distribution in K-Fold and Stratified K-Fold validation.

Results and Comparison: K-Fold vs Stratified K-Fold

Overview

The subsequent part compares the results of K-Fold and Stratified K-Fold cross-validation on the Breast Cancer dataset with a logistic regression classifier. The comparison is performed on the basis of three graphical outputs: fold-wise accuracy, confusion matrices, and validation class composition. Both techniques are compared on the grounds of classification performance, fold-wise consistency, and fairness of class representation.

Fold-Wise Accuracy Comparison

Both K-Fold and Stratified K-Fold in the provided bar chart possess good accuracy for all five folds, generally between 0.91 and 0.98. K-Fold has a bit more variation in performance, particularly for Fold 2 and Fold 4, where accuracy is much higher or lower than the average. The variation in accuracy reflects inconsistency in data distribution across the folds. Stratified K-Fold, on the other hand, possesses more stable accuracy for all the folds. The relative evenness of the heights of the green bars indicates that stratification produces more balanced performance through the preservation of class proportions across splits.

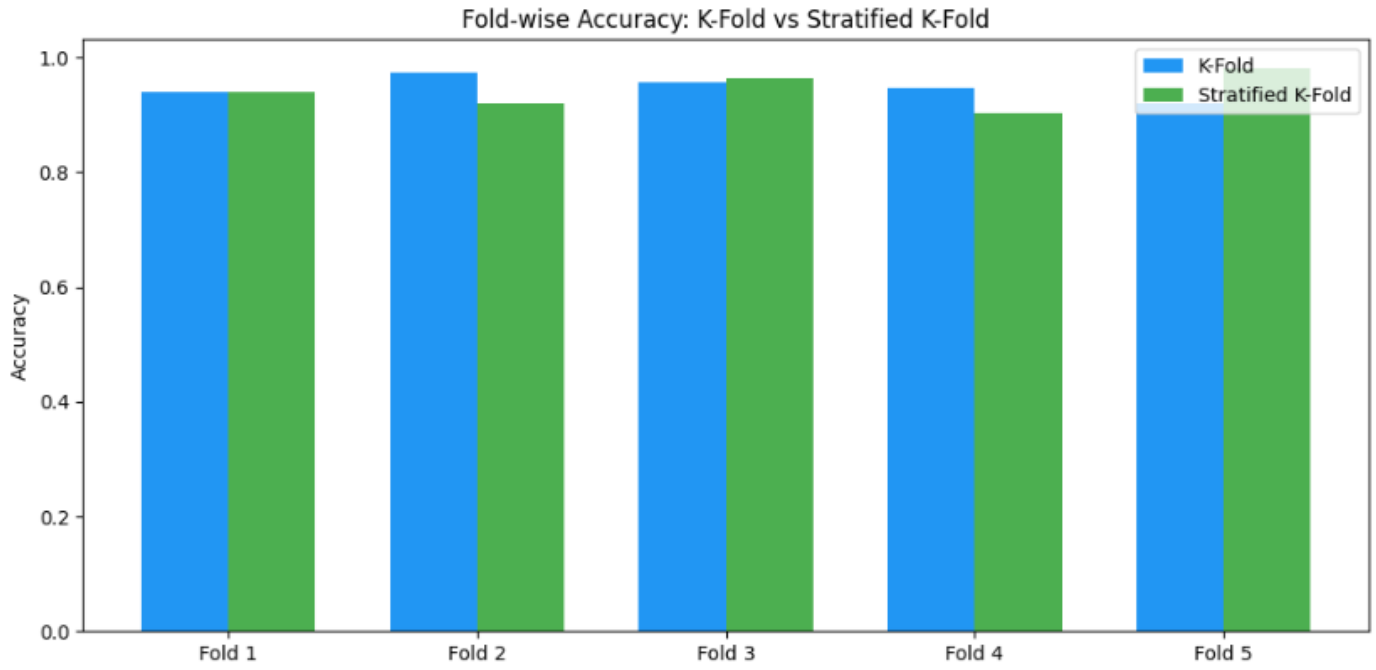

Fig1: Fold-wise accuracy comparison between K-Fold and Stratified K-Fold Cross- validation

Analysis of Confusion Matrix

K-Fold – Fold 1 and Fold 2

In Fold 1 of K-Fold, the classifier was correct on 37 malign and 70 benign cases and misclassified 2 malign as benign and 5 benign as malign. The few errors still constitute a high accuracy rating. In Fold 2, K-Fold provided even better results: 1 malign and 2 benign samples were misclassified. Again, though, this good performance can be partly attributed to an unbalanced fold with a good split of easily classifiable samples.

Stratified K-Fold – Fold 1 and Fold 2

The classifier in Stratified K-Fold Fold 1 accurately identified 38 of the malignant and 69 of the benign cases and mislabeled 5 malignant and 2 benign samples. This is almost as good in performance as K-Fold Fold 1, but most significant is the internal balance of the fold. In Fold 2, the model did slightly worse: 8 cancer instances were predicted as benign, although again the number of samples overall was roughly similar to Fold 2 in K-Fold iteration. This imbalance is a result of balanced class distribution—Stratified K-Fold preserves each fold having a balanced ratio of the minority class, which may be more stringent on the classifier but yields more representative results.

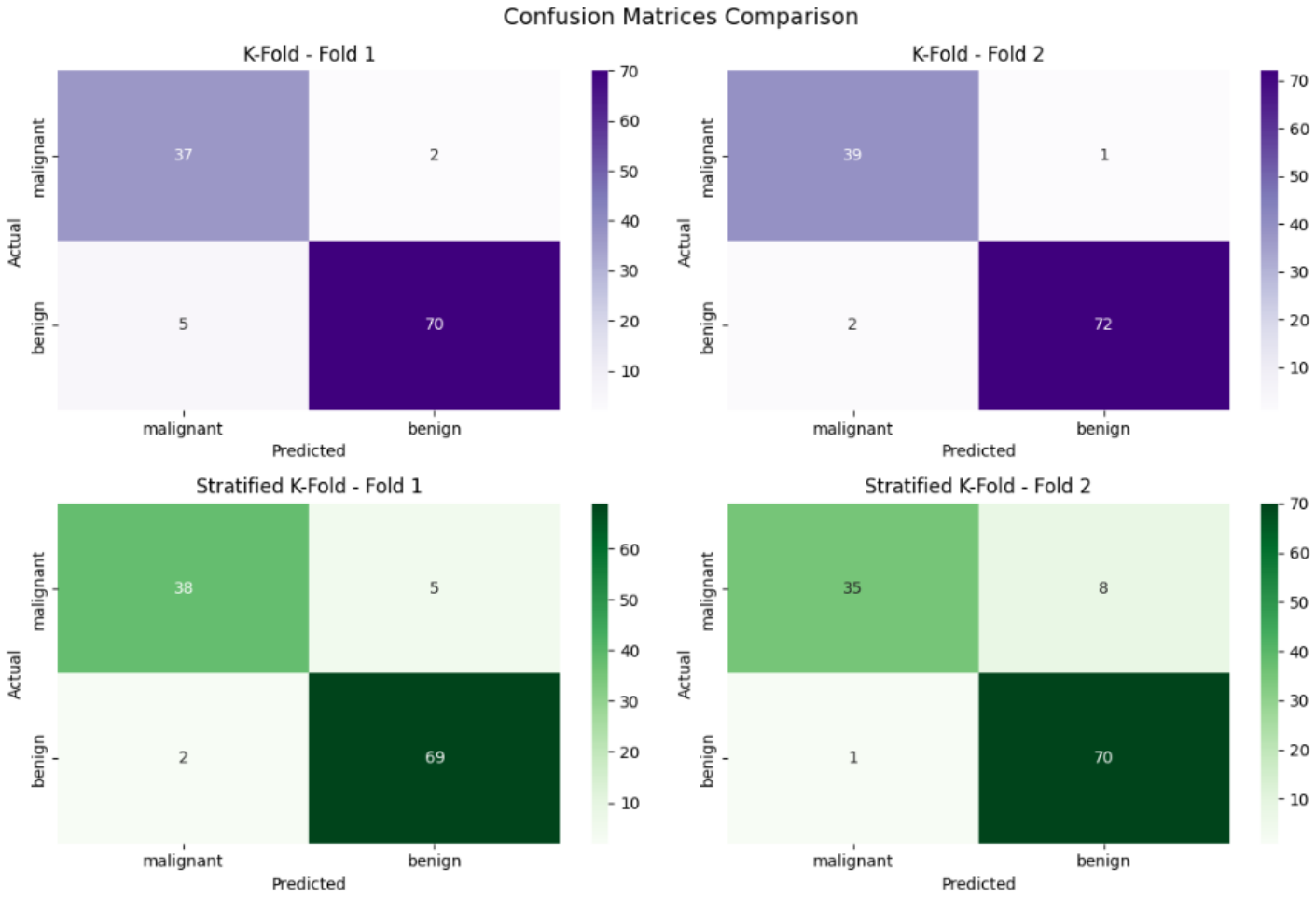

Fig2: Confusion matrices for Fold 1 and Fold 2 under K-Fold (top) and Stratified K-Fold (bottom)

Class Composition per Fold

The dot plots offer an intuitive and clear way to measure class representation by fold. Each row in the K-Fold plot represents a fold, while each dot represents a sample. The red (benign) and green (malignant) dot skewing indicates unbalanced class occurrence across the validation sets. There are a few instances with high red dot concentration, which may distort the classifier to appear more accurate than it actually is as it would tend to bias towards the majority class during testing.

On the other hand, the Stratified K-Fold dot plot shows a proper mix of green and red dots in each of the five folds. This graphic ensures that there is an approximating class ratio for each fold to the full dataset. Such a properly balancing mix forces the classifier to make more generalization to both classes, especially to the minority class (malignant), a requirement in the case of medical diagnosis.

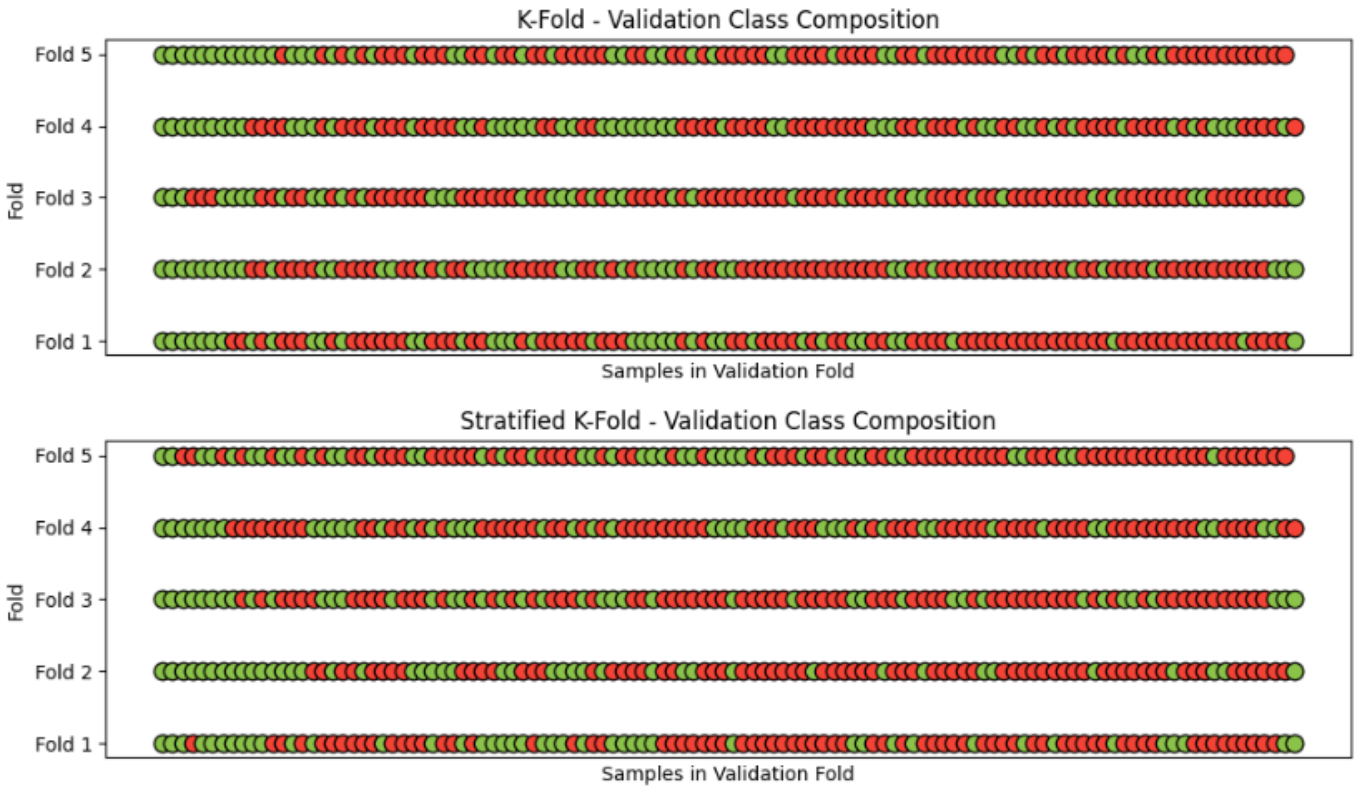

Fig3: Dot visualization of validation class composition across folds.

Conclusion

This paper emphasizes the significance of cross-validation approach in model evaluation, particularly where datasets with imbalance like Breast Cancer classification are present. Despite both K-Fold and Stratified K-Fold having high accuracy as much as they did, Stratified K-Fold provided more credible results by maintaining class distribution across all folds. Visual inspections in the way of confusion matrices and class composition plots confirmed that Stratified K-Fold maintains more balanced evaluation, especially for the minority class. In critical domains such as medical diagnosis, where misclassifying the minority class can be extremely harmful, Stratified K-Fold is the stronger and more stable validation technique.

References

1. Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th International Joint Conference on Artificial Intelligence, pp.1137–1143.
2. Refaeilzadeh, P., Tang, L. and Liu, H., 2009. Cross-validation. In: Liu, L. and Özsu, M.T. (eds), Encyclopedia of Database Systems. Boston, MA: Springer.
3. Japkowicz, N. and Shah, M., 2011. Evaluating Learning Algorithms: A Classification Perspective. Cambridge: Cambridge University Press.
4. Efron, B. and Tibshirani, R., 1997. Improvements on cross-validation: the.632+ bootstrap method. Journal of the American Statistical Association, 92(438), pp.548–560.
5. Browne, M.W., 2000. Cross-validation methods. Journal of Mathematical Psychology, 44(1), pp.108–132.
6. Bengio, Y. and Grandvalet, Y., 2004. No unbiased estimator of the variance of K-fold cross-validation. Journal of Machine Learning Research, 5, pp.1089–1105.
7. Krstajic, D., Buturovic, L.J., Leahy, D.E. and Thomas, S., 2014. Cross-validation pitfalls when selecting and assessing regression and classification models. Journal of Cheminformatics, 6(1), pp.1–15.
8. Varma, S. and Simon, R., 2006. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics, 7(1), p.91.
9. Han, J., Kamber, M. and Pei, J., 2011. Data Mining: Concepts and Techniques. 3rd ed. Amsterdam: Elsevier.
10. Forman, G. and Scholz, M., 2010. Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. ACM SIGKDD Explorations Newsletter, 12(1), pp.49–57.