

of text with videos and images as metadata. The data sets may need filtering to remove some noise. Such noise is normally created by bad quality images or videos, unclear text and website pages and fonts on background colors,

Type of testing. The type of measurement utilized in machine learning projects impacts the testing procedure. Numerical data can be analyzed using parametric tests if the sample size is at least 30 and the central limit theorem is taken into account. When dealing with categorical data, nonparametric tests can be utilized even with a small dataset size.

Dataset's distribution sample. The sample selection process in supervised models is determined at the beginning, similar to selecting sample data from the population. This may involve a random sample selection process based on chosen labels or categories. Conversely, when testing the null hypothesis, it may be necessary to use stratified or clustered sampling. This pertains to the comparison between model selection using a single dataset versus employing many models to choose the most suitable model for estimating machine learning measures.

Unsupervised models do not utilize labels, therefore, it is advisable to employ random selection methods. In statistical testing, sample selection may involve convenient sampling, which is picking students from specific classes. In machine learning testing diagnostics, the sample distribution can be referred to as a priori, but it may also be characterized as a posterior distribution that arises from testing samples from mice or people. Choosing the threshold for converting numerical measurements into binary may lead to sampling distributions with significant bias, high variance, or a balance between the two. Data purification or filtering may be necessary in this instance, even when using visual plots to show the link between the two classifications or labels.

Using several labels and doing multiple correlation analysis on the dataset can help identify any bias or variance. These dataset checks are conducted as part of the verification process. The blog has a table that connects specific machine learning models with dataset size and its characteristics [26].

6.2 Testing related taxonomy

Single model against numerous datasets. This entails evaluating a single model using various data sets. This method involves utilizing training data, testing data, unseen data, and cross-validation data. This test will verify that the model is suitable for various datasets. This factor could be applied to various datasets representing different marketing

tactics or training sub-datasets. Multiple datasets can be organized into either batches or epochs [27]. This is carried out as part of the verification process.

Single dataset against several models. If the sample distribution is not assumed to be normal and is calculated as a posterior, it may have various properties such as clustered, stratified, linear, or other non-linear representations. This technique aims to determine which model provides the most accurate estimation for a specific dataset.

Model(s) versus model(s). This entails evaluating ensemble models against each other and against individual models. The comparison of ensemble and ensemble models using the same dataset is crucial. Comparing ensemble models to single models typically demonstrates an improvement in results. Testing might be conducted on both approaches to determine if the improvement justifies utilizing the ensemble model. Another method employed in certain studies involves utilizing a combination of two models in succession. Similar to ensemble models, this typically results in enhancements in fundamental machine learning measures, although it is not as common as in ensemble models. Another form of research involves utilizing hybrid models that combine two different models in sequence, sometimes leading to enhancements in machine learning statistics. An accuracy comparison can be made among the hybrid model paradigm, ensemble model, and individual model.

Intra-organizational testing. When a business is testing a new product, marketing estimation, or investment alternatives, it is crucial to maintain consistency in dataset and model selection to validate the results for comparison testing. The same rule applies when testing is conducted for an extended duration.

Inter-organizational testing. The case of COVID-19 exemplifies this testing approach. Organizations must be cautious when choosing datasets and models to provide accurate validation and enable effective testing and comparisons between organizations.

One-tail or two-tail testing. A two-tailed test is typically used to identify both the lower and higher values of the data. For F statistics in ANOVA, a one-tail test is utilized due to the skewed distribution of the F statistic towards the lower end.

6.3 Timing of testing

Risk factor. Three levels define risk: high, medium, and low. Medical informatics, organizational-based ML projects (i.e. investment, fraud detection, and marketing decisions), web page design strategies are examples of the three risks, respectively.

During model/dataset verification. Following the hypothesis setting and data collection and splitting, data calibration and model selection starts. Projects with higher risk require more diligent work mapping dataset with an appropriate model. With several testing hypothesis, this should take care of dealing with overfitting and underfitting. Another activity related to this process is changing the threshold to discover how that may affect the sampling distribution. A significant change, like changing the sample to clustered or stratified sample, may dictate a different model selection to improve ML project's performance.

During model validation. Once a project manager is assured that the verification phase is adequately and comprehensively processed, validation activities start. At this stage, ML project will present the finding of the testing and validating the final dataset and the selected model.

An emphasis here is necessary for unsupervised model use and that it will require validation with expert opinion unknown to the ML project.

7 Sample machine learning testing from available research

ANOVA test is carried using an online calculator [28]. The test was conducted using $\alpha = 0.10$ and the results of the tests are shown below.

The first example is on comparing 6 feature extracting methods using accuracy rate for training and testing data. The printout below indicate that all methods has comparable feature extraction accuracy ratios.

The f-ratio value is 0.63045. The p-value is .445624. The result is not significant at $p < .10$.

The second example involves comparing the MAE, and the MAE for three objects selected from a dataset available on Amazon. The results for both were the same as above, indicating the results among the three categories are comparable.

The third example compares three prior models with the model followed by the research. The test was not significant among the four models, even though the suggested model by the research have results showing improvements over all three models. However, when the author computed the average of the three previous models and run the test at the same α level, the test showed that the result is significant and that the alternative hypothesis is accepted, indicating that there is a statistically significant improvement of the new model.

These examples show how close the metrics in ML normally are. Perhaps using a higher α value may

result different conclusions. Normally, research shows minor improvement the new model or experience little variation among metrics. The reader to decide which methodology to use, and how it will impact the application in question.

8 ML hypothesis testing scenarios

8.1 Supervised models

The following lists a number of scenarios where hypothesis testing is frequently used in the context of machine learning supervised models, along with the corresponding test statistics and verified with sample references for the unpopular models:

Model Comparison:

Null Hypothesis (H0): There is no significant difference in the performance of two or more models.

Alternative Hypothesis (H1): There is a significant difference in the performance of two or more models.

Application: t-Test for comparing two models or ANOVA for comparing multiple models.

Feature Importance Testing:

Null Hypothesis (H0): The importance of a specific feature is not significantly different from zero.

Alternative Hypothesis (H1): The importance of a specific feature is significantly different from zero.

Application: t-Test for comparing the importance of individual features calculated using a variety of techniques, such as decision trees, random forests, linear models, and neural networks [27].

A/B Testing for Model Variants:

Null Hypothesis (H0): There is no significant difference in the performance of two model variants (e.g., before and after hyperparameter tuning) (AWS, January 2024).

Alternative Hypothesis (H1): There is a significant difference in the performance of two model variants.

Application: t-Test for comparing the performance of two model variants, or webpage user interface design elements.

The two models could be generated by support vector machines. When utilizing Support Vector Machines (SVMs), hypothesis testing usually entails determining the model's performance significance, contrasting models, or analyzing the effects of different hyperparameters.

Bias Testing:

Null Hypothesis (H0): The model's predictions are not biased towards specific groups or classes.

Alternative Hypothesis (H1): The model's predictions are biased towards specific groups or classes.

Application: Apply statistical tests to assess whether there is a statistically significant difference in prediction accuracy or error rates across different demographic or categorical groups. Chi-Square Test for independence or Fisher's Exact Test for 2x2 contingency tables. Chi-Square test has two limitations: it cannot be computed if one cell has a zero, and a cell must have at least 5 observations for the test to be valid. For a two-classification confusion matrix, ML metrics should be adequate to test the different hypotheses. However, with multiple criteria confusion matrix of more than 3x3, Chi-Square test can be used to support the results of the ML metrics. Fisher's exact test must be used when more than 20% of cells have expected frequencies less than 5, as applying the approximation method is insufficient in such cases [29, 30].

Model Calibration Testing:

Null Hypothesis (H0): The model's predicted probabilities are well-calibrated.

Alternative Hypothesis (H1): The model's predicted probabilities are not well-calibrated.

Application: Visual inspection of calibration curves or metrics like Brier Score. An assessment metric called the Brier score is used to determine how good a predicted probability score is. This is comparable to the mean squared error; however, it is exclusive to prediction probability scores, which have values between zero and one [31].

Overfitting/underfitting Testing:

Null Hypothesis (H0): The model does not overfit/underfit the training data.

Alternative Hypothesis (H1): The model overfit/underfit the training data.

Application: Compare performance metrics between training and validation/test datasets using paired t-Test or Wilcoxon Signed-Rank Test. The paired samples t-test and the Wilcoxon test both use the same procedure to determine the test statistic and p value. The Wilcoxon test performs the analysis using ranks assigned to the data rather than the raw data values themselves [32].

Generalization Testing:

Null Hypothesis (H0): The model does not generalize well to unseen data.

Alternative Hypothesis (H1): The model generalizes well to unseen data.

Application: Permutation Test or Bootstrap Resampling to compare model performance on the original test set vs. shuffled or resampled test sets. While bootstrap is a large-sample technique, permutation tests can be applied to small samples, though a restricted selection of significance levels can occasionally be an issue with very small samples; if

used with small samples, the results may not be very useful in many cases [33]. It may be used in the verification activities of the selected model.

8.2 Unsupervised models.

The following list of scenarios, along with the corresponding test statistics, illustrates how hypothesis testing is frequently used in the context of unsupervised learning:

Clustering Quality Testing:

Null Hypothesis (H0): There is no structure or meaningful clustering in the data.

Alternative Hypothesis (H1): There is a structure or meaningful clustering in the data.

Application: Use internal validation indices (e.g., silhouette score, Davies-Bouldin index) or statistical tests to assess the quality of clusters generated by algorithms like k-means or hierarchical clustering.

Whereas the Davies-Bouldin index computation has a linear time complexity in relation to the number of clustered vectors, the Silhouette index computation, on the other hand, has a quadratic time complexity in relation to the number of vectors involved in the clustering [34].

Association Rule Significance:

Null Hypothesis (H0): There is no significant association between different features or items in the dataset.

Alternative Hypothesis (H1): There is a significant association between different features or items in the dataset.

Application: Use statistical tests (e.g., chi-square test) to assess the significance of associations discovered by association rule mining algorithms [35].

Graph Analysis Hypothesis Testing:

Null Hypothesis (H0): There is no significant structure or pattern in the graph.

Alternative Hypothesis (H1): There is a significant structure or pattern in the graph.

Application: Apply statistical tests to assess the significance of graph properties, such as clustering coefficients or centrality measures [36].

Embedding Space Evaluation:

Null Hypothesis (H0): There is no significant structure or separation in the embedding space.

Alternative Hypothesis (H1): There is a significant structure or separation in the embedding space.

Application: Apply statistical tests to assess the significance of distances or relationships between points in the embedded space generated by techniques like t-SNE or UMAP [37].

Dimensionality Reduction Assessment:

Hypotheses: Null Hypothesis (H0): The reduced representation does not capture significant information.

Alternative Hypothesis (H1): The reduced representation captures significant information.

Application: Apply dimensionality reduction (e.g., PCA) to obtain a lower-dimensional representation. Assess the distribution of data in the reduced space using Kolmogorov-Smirnov test. This applies when reducing or increasing a number of classifications to deal with over- or under- fitting [37].

Anomaly Detection Assessment:

Null Hypothesis (H0): Anomalies are not significantly different from normal instances.

Alternative Hypothesis (H1): Anomalies are significantly different from normal instances.

Application: Evaluate whether the identified anomalies are statistically significant. Use a statistical test to compare the distributions of features between normal and anomalous instances, such as a T-test, or Mann-Whitney U test [38]; [39]. An example of an anomalous case is Outliers.

9 Conclusions

The paper provided a detailed explanation of exploring machine learning hypothesis testing through taxonomies related to factors like supervised versus unsupervised model selection, testing statistics, data types, testing during verification and validation activities, and models versus datasets testing. The paper outlined hypothesis testing scenarios with H0 and H1 and suggested testing for both supervised and unsupervised models.

ML hypothesis testing is still developing compared to the well-established practice in traditional statistical theory that has been around for decades. This development began with the release of the SMV model in 2000. This research utilizes statistical testing methods like regression, correlation, and chi-square, in addition to Bayesian theory. Some of these are mostly relevant to unsupervised models. Another instance where statistical testing is used is when we compare fundamental machine learning measures such as accuracy and F1 score across different domains and projects. No machine learning testing methods based on sample and population distributions for accuracy or F1 scores exist. Future research will investigate the potential application of histogram approximation of the harmonic mean for testing F1 measures. The author aims to inspire and provide guidance to scholars

interested in undertaking machine learning-based hypothesis testing initiatives.

The study focuses solely on the fundamental structure of machine learning metrics and does not encompass projects related to websites with extensive datasets or in-depth research on genome-wide association studies (GWAS) or specialized cases like semi-supervised data. Future study will focus on testing hypotheses using Python. Furthermore, perform a future literature review to identify the obstacles and opportunities in Machine Learning hypothesis testing as they develop.

References

- [1] QuestionPro, "Sample Size Determination: Definition, Formula, and Example." Dec. 2023, <https://www.questionpro.com/blog/determining-sample-size>.
- [2] LinkedIn, "What distinguishes a stratified sample from a cluster sample?," 2023, <https://www.linkedin.com/advice/1/what-distinguishes-stratified-sample-from-cluster-skills-statistics-y4jtf>.
- [3] Forms.app, Aug. 2023, <https://forms.app/en/blog/guide-to-measurement-scales>.
- [4] IBM, August, 2021, <https://www.ibm.com/docs/en/db2woc?topic=nonparametric-usage>.
- [5] Wikipedia, Oct. 2020, https://en.wikipedia.org/wiki/Timeline_of_machine_learning.
- [6] B. Mieth, M. Kloft, J. Rodríguez, et al. "Combining Multiple Hypothesis Testing with Machine Learning Increases the Statistical Power of Genome-wide Association Studies," *Sci Rep.*, vol. 6, pp. 36671, 2016, doi: [org/10.1038/srep36671](https://doi.org/10.1038/srep36671).
- [7] A. Alwosheel, S. van Cranenburgh and G. Caspar, "Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis," *Journal of Choice Modelling*, vol. 18, pp. 167-182, Sept. 2018, doi: [org/10.1016/j.jocm.2018.07.002](https://doi.org/10.1016/j.jocm.2018.07.002).
- [8] Vinayak, N., Ahmad, S. (2023). Sample Size Estimation for Effective Modelling of Classification Problems in Machine Learning. In: I. Woungang, S. K. Dhurandher, K. K. Pattanaik, A. Verma and P. Verma (eds), *Advanced Network Technologies and Intelligent Computing. ANTIC 2022. Communications in Computer and Information Science*, vol 1798. Springer, Cham, pp. 365–378. doi: [org/10.1007/978-3-031-28183-9_26](https://doi.org/10.1007/978-3-031-28183-9_26).
- [9] J. J. Li and X. Tong, "Statistical Hypothesis Testing versus Machine Learning Binary Classification: Distinctions and Guidelines,"

- Patterns. vol. 1, no. 7, 2020, doi.org/10.1016/j.patter.2020.100115.
- [10] R. Fazai, M. Mansouri, K. Abodayeh, M. Trabelsi, H. Nounou and M. Nounou (2019), "Machine Learning-Based Statistical Hypothesis Testing for Fault Detection," 4th Conference on Control and Fault Tolerant Systems (SysTol), Casablanca, Morocco, pp. 38-43, 2019, doi:10.1109/SYSTOL.2019.8864776.
- [11] A. K. Prajapati and U. K. Singh, "An Optimal Solution to the Overfitting and Underfitting Problem of Healthcare Machine Learning Models," Journal of Systems Engineering and Information Technology (JOSEIT), vol. 2, no. 2, pp. 77-84, 2023, doi:10.29207/joseit.v2i2.5460.
- [12] A. Biswal, "Bagging in Machine Learning: Step to Perform And Its Advantages," Aug. 2023, <https://www.simplilearn.com/tutorials/machine-learning-tutorial/bagging-in-machine-learning>.
- [13] P. Huilgol, "Accuracy vs. F1-Score," Analytics Vidhya, Aug. 2019, <https://medium.com/analytics-vidhya/accuracy-vs-f1-score-6258237beca2>.
- [14] A. Bhandari, "Understanding & Interpreting Confusion Matrix in Machine Learning," Dec. 2023, <https://www.analyticsvidhya.com/blog/2020/04/confusion-matrix-machine-learning>.
- [15] Future-diagnostics, Dec. 2023, <https://www.future-diagnostics.com/blog/sensitivity-and-specificity>.
- [16] Kaggle. 2024, <https://www.kaggle.com/datasets/groupLens/movielens-20m-dataset>.
- [17] V. Singh, "Difference Between Type 1 and Type 2 Error," 2023, <https://www.shiksha.com/online-courses/articles/difference-between-type-1-and-type-2-error>.
- [18] A. Gustafsen, "The Confusion Matrix in Hypothesis Testing," Towards Data Science, Mar 2022, <https://towardsdatascience.com/the-confusion-matrix-explained-part-1-513c6f659c1>.
- [19] H. Erdogmus, "Bayesian Hypothesis Testing Illustrated: An Introduction for Software Engineering Researchers," ACM Computing Surveys. vol. 55, no. 6, pp. 1-28, Article No. 119, 2022, doi:pdf/10.1145/3533383
- [20] B. O.Tayo, "Bayes' Theorem in Plain English: Simplest explanation of Bayes' Theorem," Medium, 2021, <https://benjaminobi.medium.com/bayes-theorem-in-plain-english-eeb142710475>.
- [21] ChatGPT 3.5, Dec. 2023.
- [22] Upgrad, Dec. 2023, <https://www.upgrad.com/blog/types-of-data>.
- [23] V. R. Joseph, "Optimal ratio for data splitting," 2022, doi: org/10. 1002 /sam. 11583.
- [24] Geegsforsgeegs, Dec. 2023, <https://www.geegsforsgeegs.org/underfitting-and-overfitting-in-machine-learning>.
- [25] R. Nickerson, U. Varshney and J. Muntermann, "A method for taxonomy development and its application in information systems," Eur J Inf Syst., vol. 22, pp. 336–359, doi.org/10.1057/ejis. 2012.26.
- [26] Manika, "Your 101 Guide to Model Selection In Machine Learning," ProjectPro, Oct. 2023, <https://www.projectpro.io/article/model-selection-in-machine-learning/824>.
- [27] J. Brownlee, "Difference Between a Batch and an Epoch in a Neural Network," 2022, <https://machinelearningmastery.com/difference-between-a-batch-and-an-epoch>.
- [28] socscistatistics.com, 2024, <https://www.socscistatistics.com/tests/anova/default2.aspx>.
- [29] N. Azaria, "Feature Importance: 7 Methods and a Quick Tutorial," Aporia, Jan. 2024, <https://www.aporia.com/learn/feature-importance>.
- [30] H-Y. Kim, "Statistical notes for clinical researchers: Chi-squared test and Fisher's exact test," Restor Dent Endod, vol. 42, no. 2, pp. 152–155. 2017, published online 2017 Mar 30. doi: 10.5395/rde.2017.42.2.152
- [31] S. Dash, "Brier Score – How to measure accuracy of probabilistic predictions," Jan. 2024, <https://www.machinelearningplus.com/statistics/brier-score>.
- [32] F. Chumney, "Paired samples t test and the Wilcoxon signed ranks test," Westga.edu, Jan. 2024, https://www.westga.edu/academics/research/vrc/assets/docs/PairedSamplesTtestWilcoxonSignedRanksTest_TRANSCRIPT.pdf.
- [33] W. Xue and P. Adkins, "How to leverage permutation tests and bootstrap tests for baselining your Machine Learning models," Data Science at Microsoft, Aug. 2023, <https://medium.com/data-science-at-microsoft/how-to-leverage-permutation-tests-and-bootstrap-tests-for-baselining-your-machine-learning-models-f1010bf22e71>
- [34] K. Amrulloh, T. Pudjiantoro, P. Sabrina and A. Hadiana, "COMPARISON BETWEEN DAVIES-BOULDIN INDEX AND SILHOUETTE COEFFICIENT

- EVALUATION METHODS IN RETAIL STORE SALES TRANSACTION DATA CLUSTERIZATION USING K-MEDOIDS ALGORITHM,” In 3rd South American International Conference on Industrial Engineering and Operations Management, May 2022, doi:org/10.46254/SA03.20220384.
- [35] I. Ismiguzel, “A Guide to Association Rule Mining” Towards Data Science, 2023, <https://towardsdatascience.com/a-guide-to-association-rule-mining-96c42968ba6>.
- [36] F. Xia, et al, “Graph Learning: A Survey,” IEEE Transactions on Artificial Intelligence, vol. 2, no. 02, pp. 109-127, 2021, doi:10.1109/TAI.2021.3076021
- [37] A. Acharya, “The Full Guide to Embeddings in Machine Learning,” Encord, May 2023, <https://encord.com/blog/embeddings-machine-learning>.
- [38] A. B. Nassif, M. A. Talib, Q. Nasir and F. M. Dakalbab, “Machine Learning for Anomaly Detection: A Systematic Review,” IEEE Access, vol. 9, pp. 78658-78700. 2021, doi:10.1109/ACCESS.2021.3083060.
- [39] C. Quiroz-Vázquez, Dec. 2023, <https://www.ibm.com/blog/anomaly-detection-machine-learning>.
- [40] P. Braca, L. M. Millefiori, A. Aub De Maio, S.A.ry, Marano and P. Willett, “Statistical Hypothesis Testing Based on Machine Learning: Large Deviations Analysis,” IEEE Open Journal of Signal Processing, vol. 3, pp. 464-495, 2022, doi:10.1109/OJSP.2022.3232284.
- [41] K. Sechidis, B. Calvo and G. Brown. “Statistical Hypothesis Testing in Positive Unlabelled Data.” In: I. Calders, F. Esposito, E. Hüllermeier, R. Meo (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2014. Lecture Notes in Computer Science, vol. 8726, Springer, Berlin, Heidelberg, 2014, doi:org/10.1007/978-3-662-44845-8_5.