OptiML: An Analysis of Various Machine Learning Algorithms' Performance on Diverse Datasets

Mridula Venkatasamy, Amy An, Sarah Huang, Julia Steed

***Abstract*— Data mining and machine learning (ML) are rising in popularity for use in many different fields, such as healthcare. However, many inexperienced users of ML do not know how best to apply machine learning techniques to their data. In this study, we aim to address this challenge by creating a Jupyter notebook that provides examples of how best to apply ML techniques to various datasets. In this study we will compare the performance of six different scikit-learn [7] machine learning classifiers on five datasets with unique characteristics. We will primarily focus on accuracy and compare our results to the work in [1]. We anticipate that our classifiers will achieve comparable performance to the results in [1]. We expect random forest to perform the best for most datasets due to its robustness. We will schedule work on this study in week long sprints. We will delegate tasks based on our team members skillsets to analyze, preprocess, normalize, and classify our datasets.**

I. Introduction

Knowledge extracted from raw data is shaped by the analysis technique. Different approaches prioritize different aspects of the data and thus lead to distinct conclusions.

Data mining specializes in identifying hidden relationships and making future predictions. Combined with machine learning classifiers, data behavior and prediction making can be validated. Data mining and machine learning can be utilized to analyze and predict data in any field such as predicting diseases. This work aims to determine which machine learning classifiers combined with universal data mining techniques are better suited for prediction making of a particular dataset.

Our study replicates previous work "Comparison of Machine Learning Algorithms in Data classification" by Hassan, et al. The authors compared different machine learning classifiers on two datasets about heart disease and hepatitis disease to determine which is more effective in analysis and disease prediction. Specifically, they compared accuracy, precision, and f measure for Logistic Regression, Decision Tree, Niven Bayes, k-Nearest Neighbors, Support Vector Machine and Random Forest classifiers [1]. At the end, they concluded random forest performed the best with 83% accuracy in the heart dataset and 85% accuracy in hepatitis disease prediction [1].

This study will apply the same machine learning classifiers on five unique datasets in terms of subject matter and number of variables. We will add to their results and verify if the Random Forest classifier remains the best machine learning classifier across different data aside from disease prediction.

With experience in machine learning and visualization, the team has the capabilities to complete this project. Two team members, Julia and Amy, specialize in implementing machine learning algorithms on different datasets. Mridula has real experience in data development and analytics from her background of database development for the University of Tennessee Men's basketball team. Sarah's research focus is in data visualization has experience in executing visualizations in jupyter notebooks. Together, this team covers the prerequisites for a project that incorporates machine learning algorithms, data analysis, and visualization.

II. Research Value

Machine learning is a rapidly growing sector of the technology industry. Companies are trying to see how machine learning algorithms can help them grow. However, selecting the right machine learning algorithm can be challenging for users who are unfamiliar with the options that are available. By using diverse datasets across multiple machine learning algorithms, we are able to provide a conclusive answer to the user based on the accuracy.

By evaluating the most accurate and impactful machine learning algorithms across diverse datasets, we are able to provide the user with conclusive results on which algorithm will deliver the best results. This saves the users the time and effort of learning and testing each algorithm, which in return will allow them to focus on other important tasks. The user will also benefit from access to a well versed guide (jupyter notebook) on the methods needed to replicate our results with their dataset.

User feedback will play an important aspect in whether or not our research was beneficial. It is not feasible to test hundreds of datasets due to the time constraint. The user feedback will play a key role in validating our findings or in highlighting an issue in our initial conclusion. The paper will include a google form, where the user can provide insights on what did or did not work for them.

III. Methodology

In this study, we plan to evaluate six popular machine learning classifiers on five unique datasets. This section provides an overview of the five datasets, the data analysis and pre-processing we plan to perform, and how we plan to train and test machine learning classifiers on our data.

A. Datasets

We sourced five datasets from Kaggle. We aimed to find diverse datasets with different characteristics in order to perform a comprehensive evaluation of our classifiers. By datasets that are distinct from one another, we can effectively

evaluate the robustness of the classifiers to a wide range of data. The five datasets are as follows:

- Spotify tracks: We classify a song's genre based on characteristics such as danceability, acousticness, energy, and tempo [2]
- Road accident survival: We classify whether or not someone survived a car accident based on whether they wore a seatbelt and/or helmet and speed of impact [3]
- Breast cancer: We classify whether the patient had a recurrence of cancer based on data such as their age, tumor size, and degree of malignancy. [4]
- Apple quality: We classify whether an apple is "good" or "bad" based on its juiciness, ripeness, crunchiness, and similar factors [5]
- NYC squirrels: We classify a squirrel's age as juvenile or adult based on its moans, tail twitches, tail flags, and similar factors [6]

B. Data Analysis

Before applying ML classifiers to our data, we will analyze our data. Our analysis will give us insight into whether we need to scale, normalize, or subsample it to give us the best possible classifiers. We will plot our data in a Jupyter notebook and apply any normalization or other preprocessing as needed before training classifiers on our data. We will also remove any unnecessary features from our datasets as needed if we don't plan to use that feature in the training process.

C. Data Classification

Lastly, we will take our preprocessed datasets and apply scikit-learn [7] methods for six unique classifiers to the five datasets. The classifiers we will use are as follows:

- K-Nearest Neighbors
- Support Vector Machine
- Naive Bayes
- Random Forest
- Logistic Regression
- Decision Tree

We will use scikit-learn's built-in classifier training methods for each of our classifiers. We will validate our results using scikit-learn's built-in validation and testing methods. We will compute accuracy, precision, recall, and F-score for each dataset and classifier and compare our results with plots and tables.

IV. Schedule and Project Management

A. Schedule

Our timeline is shown in Figure 1. May 6th was the last day of class. For the first week, we will find and select the dataset, and start looking into the dataset. Then, we will select the machine learning algorithms that we will be using. Furthermore, we will implement those selected machine learning algorithms for the next couple weeks. We have set it to be at least two weeks since training and testing machine learning algorithms is very time consuming. Then, we will check the performance score and make improvement on those performance through hyperparameter tuning, if necessary.

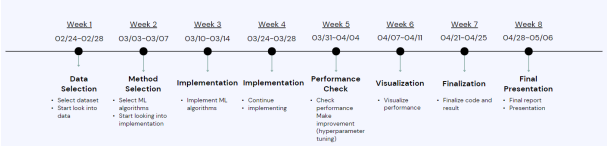

Fig. 1. Timeline

After having all the results, we will visualize the result and finalize the code for the final report and presentation by the last day of class, which is May 6th.

B. Risks

Our biggest potential risk is to find the data and preprocess it for the machine learning algorithm. Since different machine learning algorithms require input formatting, datasets need to be rescaled. Varying data formatting is the first obstacle that our team will be facing. The other obstacle that we will face is the training time. Training and testing using machine learning algorithms take a long time, having higher time complexity. This might be worsen if we choose to use the larger datasets.

C. Resources

For the datasets, we will be finding them on Kaggle. For now, we have not decided on which dataset that we will use, but we will decide by next week, following our timeline. We will also reference the paper [1] that we found to possibly replicate their work first.

V. Team

Julia has used scikit-learn classifiers in the past. She has worked on various projects such as a Bible genre predictor using Word2Vec, a rock climbing skill level predictor, and an distracted driving image classification dataset. She has worked with metrics such as accuracy and f-score and evaluated ML methods on various types of datasets with diverse characteristics. She also has experience with data analysis and preprocessing for ML.

Sarah has past experience in comparing machine learning approaches such as a class project with a Scooby-Doo watchability dataset. While her focus is not machine learning classifiers, her research specialty is in data visualization. She also has experience in Python and jupyter notebooks.

Seoyoung (Amy) An has experience with both machine learning and visualization. She has worked on a research project to visualize the neural network structures and how its accuracy changes through the training process. She also has worked on projects using different ML methods to find the best starting word for Wordle, and has some experience for ML.

Mridula has experience with data development, business intelligence, and computer vision. Her background includes working on database development for the University of Tennessee Men's basketball team, understanding user needs and requirements as an integration engineer for Osa Commerce, and working on pose estimation to derive sports analytics.

In order to leverage each team member's unique skills, Julia and Amy will focus on the application of ML classifiers to each dataset, using their prior experience with ML methods. Sarah and Mridula will focus on data analysis and visualization, leveraging their skills from previous work with data.

VI. References

References

[1] C. A. Ul Hassan, M. S. Khan and M. A. Shah, "Comparison of Machine Learning Algorithms in Data classification," 2018 24th International Conference on Automation and Computing (ICAC), Newcastle Upon Tyne, UK, 2018, pp. 1-6, doi: 10.23919/IConAC.2018.8748995. keywords: Diseases;Kidney;Heart;Data mining;Support vector machines;Machine learning;Classification algorithms;Data Mining;Machine Learning;Data Analysis

[2] M. Pandya, "Spotify Tracks Dataset." Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/maharshipandya/-spotify-tracks-dataset

[3] H. Sarder, "Road Accident Survival Dataset." Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/himelsarder/road-accident-survival-dataset

[4] B. Tandon, "Breast Cancer Data." Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/bhumikatandon/breastcancerdata

[5] A. Nelgiriyewithana, "Apple Quality." Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/nelgiriyewithana/apple-quality

[6] D. Weir, "NYC 2018 Squirrel Census." Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/dominoweir/nyc-2018-squirrel-census

[7] "scikit-learn," scikit, https://scikit-learn.org/stable/ (accessed Feb. 13, 2025).