Comparing the Performance of Various Machine Learning Algorithms

Arnab Hati
Master of Science in Data Analytics
National College of Ireland
Dulin, Ireland
x22107321@student.ncirl.ie

Abstract — **The paper focuses on achieving fruitful findings by comparing a pair of separate libraries i.e. Scikit learn and PyCaret, with several machine learning algorithms on three different datasets. The Employee Attrition dataset, the Dimond Price Prediction dataset, and the Heart Disease dataset are the three distinct datasets explored in the present article. Utilizing classification machine learning techniques like Logistic Regression, Random Forest Classification, and KNeighborsClassifier, the Employee Attrition dataset, and Heart Disease dataset are used to analyze Confusion Matrix and identify employee attrition and cardio attack based on various features. Based on a diamond's physical properties and surroundings, the Linear Regression and Random Forest Regressor are employed in the Dimond Price dataset to forecast the diamond's price. Confusion Matrix, ROC, and Area Under the Curve (AUC) For classification datasets, the ROC curve is employed, while for regression models, R2, RMSE, and MSE are used. Confusion Matrix and several metrics, including Sensitivity, Specificity, Accuracy, and Kappa, are used to evaluate the performance of the model. Machine learning models are created and predictions are formed by examining various algorithms and the aspects connected with them.**

Keywords— Supervised Machine Learning algorithms, Logistic Regression, Linear Regression, Random Forest, K Nearest Neighbors, Scikit Learn, PyCaret Confusion Matrix, ROC, Area Under ROC curve, R2 score, Mean Square Error, Root Mean Square Error.

I. INTRODUCTION

A component of artificial intelligence called machine learning makes use of data to provide systems with the capacity to autonomously improve through data analysis, enhance decision-making, and forecast effective outcomes. An algorithm used for machine learning uses sample data as input, recognizes patterns in the data, and forecasts a useful outcome. Machine learning has several uses, including determining the authenticity of spam emails, social networking features, bank fraud detection, etc. Instagram is the most excellent representation of how a machine learning algorithm operates in a social media setting. Instagram records all of our activity, including likes and comments on particular employment opportunities, conversations, the sites and posts we visit, and the total amount of time we spend using the app each day. It creates a pattern based on the activities and preferences.

The most difficult challenge in all businesses is managing the enormous amount of data that increases every second. All of the conventional methods of gathering, analyzing, and storing this tremendous quantity of data do not produce and forecast improved outcomes. For this, it is necessary to have the appropriate expertise in data analysis and data processing. This is among the most crucial justifications for why data scientists utilize machine learning tactics and procedures for higher value forecasts that may aid in enhanced decision-making and provide real-time solutions with little assistance from humans. A usual data mining methodology works on these steps: requirement gathering and understanding, data collection and preparation, data modelling, data evaluation and predicting result. For these the most common methodologies are CRISP-DM (Cross-Industry Standard Process for Data Mining) and KDD (Knowledge Discovery Databases). Different methods like Classification, Regression, Clustering etc. are used to evaluate patterns and predict interesting results from different types of datasets.

In order to get actionable insights from the models and prediction results acquired, a supervised machine learning technique is deployed on 3 distinct types of datasets in this project. As we know a lot of companies in recent times is witnessing employee attrition in the workplace due to various reasons like job dissatisfaction, less pay, hectic schedule etc. Random forest Classification and Logistic Regression algorithms are applied to the Employee Attrition dataset in order to determine the likelihood of employee attrition for a company. This assessment will help the company to self-audit its work culture in order to bring changes if needed for the betterment of the employee as well as their business. Prices of diamonds are very dynamic and it's quite a gamble to invest in it, so to make it easier this paper determines the diamond price at a particular time period in order to give us a fair assumption of our financial investment. Linear Regression and Random Forest Regressor are used for this determination. With the growing heart disease rate, it is important for us to have a thorough check-up to make sure we have a healthy heart. KNeighbors Classifier and Logistic Regression are used to determine one's heart condition and whether that person is suffering from heart disease or not.

II. RELATED WORK

In this research, both Scikit Learn and PyCaret, two separate machine learning libraries, are used to compare the performance of different machine learning algorithm designs. The datasets are analyzed and modeled, and the most accurate and successful predictions are used to evaluate the final models. To understand these ideas and how different supervised machine-learning algorithms were applied to the datasets utilized in this study, several academic articles were reviewed and analyzed.

[1] Demonstrates how the use of machine learning may be used to create models from massive amounts of data and solve hard mathematical issues, which can assist any corporate organization in overcoming its obstacles.

[2] The paper discusses the benefits and drawbacks of various algorithms, and how machine learning may be applied in many industries to cope with enormous amounts of complicated and dynamic data.

[3] explains how supervised machine learning algorithms may be used to diagnose and forecast cardiac disease at a high level. This document describes heart disease's prognosis so that it may be treated at the appropriate time because it is the cause of many other ailments.

[4] The performance of Kneighbour and Random Forest in image identification is the main subject of this article. Additionally, a result summary is offered to the researchers at the conclusion, making the data useful for obtaining a more accurate result set while minimizing human efforts.

[5] In this paper, decision trees, random forests, and support vector machines are discussed as machine learning methods for forecasting diamond prices. In addition to case studies of research projects that have employed machine learning for diamond price prediction, it includes an overview of data sources, advantages, problems, and limits.

[6] It is discussed how decision trees, random forests, and logistic regression are used to forecast staff attrition. In addition to case studies of research investigations that have included machine learning, it gives an overview of data sources, advantages, obstacles, and limits. It offers insightful information on the use of machine learning methods.

[7] Conduct tests on real-world datasets and contrast the findings to evaluate two well-known machine learning frameworks, PyCaret and Scikit-learn. Additionally, they offer case studies of how each library was used for various machine-learning tasks. The authors come to the conclusion that Scikit-learn is more versatile and adjustable for expert users, whereas PyCaret is more user-friendly and needs less code. This essay is a useful tool for anyone trying to decide between these two well-known libraries.

[8] The CRISP-DM technique and its several phases, which include business understanding, data unstanding, data preparation, modeling, assessment, and deployment, are described in this document. It emphasizes the benefits of employing the technique, including its methodological framework, adaptability, and capacity to take into account various kinds of data and machine learning models. It is a useful resource for people considering applying the concept to machine learning applications, particularly those involving the prediction of financial values.

[9] The need for a sound assessment technique is discussed in order to guarantee that machine learning models generalize effectively to new data. It also describes the many assessment measures, including accuracy, precision, recall, F1-score, and ROC-AUC, that may be used to rate a model's effectiveness. Additionally, it covers the significance of model choice and how holdout validation, k-fold cross-validation, and nested cross-validation can affect a machine learning system's performance. Finally, it offers suggestions for choosing an algorithm. In conclusion, this post offers a helpful introduction of machine learning model evaluation, model selection, and algorithm selection.

[10] the significance of model selection in statistical learning and offers helpful advice on picking the best approach for a particular issue. It highlights the significance of comprehending the underlying assumptions and limits of each approach and issues a warning against using any approach indiscriminately without taking the context of the issue into account. In general, the article serves as a useful tool for academics and professionals working in the fields of statistical learning and machine learning.

III. DATA MINING METHODOLOGY

Data mining procedure is a systematic method for identifying significant patterns, trends, and correlations in huge datasets. Data cleansing, integration, selection, transformation, and mining are some of the processes in the process. Prior to collection and preparation, data that is inconsistent or irrelevant is eliminated. The data is then translated into a suitable format for additional analysis. In order to glean knowledge and insights from the data, data mining techniques including classification, clustering, regression, and association rule mining are used. The findings are then assessed, and the most promising patterns are chosen for further analysis and application in decision-making procedures.

Data collection, comprehending the data, preparation utilizing cleaning and transformation, data modeling, and model assessment are the five stages of a typical data mining process. Each stage emphasizes a more thorough analysis of the data and the extraction of knowledge for future forecasting. Most corporate organizations use the most well-known techniques, including CRISP-DM, SEMMA, and KDD. The processes for building models are nearly the same in KDD and SEMMA. The most widely utilized of these three approaches, CRISP-DM, follows the whole life cycle of data modeling and offers a consistent platform for project management.

For three major datasets, CRISP-DM (Cross-Industry Standard Process for Data Mining) is employed to offer greater insights. There are six stages, each with a description, the steps taken during that stage, and the results of that stage. The details of these steps are provided.

A. Business Understanding

This stage involves understanding the domain of the problem and identifying the key stakeholders who will be impacted by the project.

- Organizations can use the Employee Attrition dataset to anticipate employee departure probabilities and improve working conditions.
- Diamond dealers and purchasers may make wise judgments based on the variables that determine pricing by analyzing the Diamond Price dataset, which can offer insights into the diamond industry.
- The purpose of the heart disease dataset is to gain insights into risk factors and symptoms associated with heart disease and identify patterns and trends to predict the likelihood of heart disease in future patients.

B. Data Understanding

In the stage of data understanding, the amount and quality of the data are evaluated, missing or incomplete data is found, and the distribution of the data is understood. Techniques for

exploratory data analysis are used to view the data and understand its features.

- The 'Attrition' column is the dependent variable in the employee attrition dataset, which has 1470 columns and 35 rows.
- 'price' is the dependent column in the 53940 rows and 11 columns of the diamond price data.
- With the exception of the cardio column, all of the 13 columns in the heart data set are independent features, totaling 70000 rows.

C. Data Preparation

For data mining algorithms to produce accurate and insightful findings, this phase is crucial. It comprises tasks like data integration, data reduction, data cleansing, and data transformation. This step makes that the data is accurate, reliable, and in the right format for analysis.
The three datasets have no missing values.
After box charting two datasets—employee attrition and diamond pricing. Violin plotted the heart disease dataset, removing the outliers from the collection of data.
The dataset's labels with the most unique values and those with a single unique value were eliminated from the dataset.

D. Data Modelling

At this point, a suitable data model or machine learning technique—such as Random Forest, Linear Regression, etc.—is selected. If multiple methods are used, the model must be properly displayed and analyzed using a variety of statistics.

- On the classification datasets, Logistic Regression, SVM - Linear Kernel, and K Neighbors Classifier have been utilized.
- K Neighbors Regressor and Random Forest Regressor are the models implemented in the regression dataset.

E. Data Evaluation

evaluation stage aims to identify any issues or problems that may affect the quality of the model and its predictions. It involves checking for missing values, outliers, and other anomalies in the data, exploring the data, selecting appropriate variables, and creating a training and testing dataset.
The pre-processed data was divided into a training model and a test model at a ratio of 80:20 for model fitting. For the categorical feature to numerical conversion in the employee attrition and diamond price prediction dataset, label encoder and one-hot encoding are utilized. Standard scaler and min-max scaler are two ways to scale the dataset. The libraries used for model development and model comparison for best suited are Scikit Learn and PyCaret.

F. Deployment

The CRISP-DM methodology's deployment step, which entails putting the created model into use, is its last phase. It entails integrating with current systems and procedures, testing with actual data, and keeping track of its performance to make sure it achieves company goals.

IV. MODEL IMPLEMENTATION

A. Dataset 1: Employee Attrition Dataset

B. Dataset 2: Heart Diseases Dataset

C. Dataset 3: Dimond price prediction dataset

V. CONCLUSION

The research focuses on using several machine learning libraries to apply different machine learning methodologies and algorithms to datasets related to heart disease, employee attrition, and the Dimond Price. Different methods of assessment are used to build models and assess their performance.

When applied to the Employee Attrition dataset, Logistic Regression performed better than KNN at predicting Attrition. SciKit Learn Library's accuracy was 87%, whereas PyCaret Library's accuracy for the identical model was 87.08%. For this Dataset, both libraries had comparable results.

To assess the likelihood of having a heart attack, the Heart Disease dataset was subjected to the XGboost and Logistic Regression models, with the XGBoost model performing 75% more accurately than the Logistic model.

In the Dimond Prediction dataset, KneighborsRegressor and RandomForestRegressor were applied. The kneighborsRegressor model gave accuracy for SciKit Learn and PyCaret library 96.46% and 94.30% respectively, while Random Forest Regressor gave an accuracy of 98.12% and 97.99% respectively. Sklearn Random Forest Regressor performance is better for this dataset.

REFERENCES

[1] Cantwell, G., Han, J., Wang, X., & Zhang, J. (2019). Machine learning applications in marketing: A review. Journal of Business Research, 98, 136-150.

[2] Zhang, Y., Liu, Y., & Zhou, W. (2018). Big data-driven smart manufacturing research and applications. IEEE Transactions on Industrial Informatics, 14(4), 1537-1546. doi: 10.1109/TII.2018.2824360.

[3] R. Shahid, M. Farooq and M. S. Sarfraz, "Machine Learning Techniques for Heart Disease Diagnosis," 2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE), 2020, pp. 1-6, doi: 10.1109/ICECCE50494.2020.9218777.

[4] M. Sheykhmousa, M. Mahdianpari, H. Ghanbari, F. Mohammadimanesh, P. Ghamisi and S. Homayouni, "Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 13, pp. 6308-6325, 2020, doi: 10.1109/JSTARS.2020.3026724.

[5] A. K. Pandey and N. Kumar, "Predicting Diamond Prices using Machine Learning Techniques," 2021 3rd International Conference on Inventive Research in Computing Applications (ICIRCA), 2021, pp. 637-642, doi: 10.1109/ICIRCA51687.2021.9459337.

[6] G. G. Pavithra and S. K. Srivatsa, "Machine Learning Techniques for Employee Attrition Prediction," 2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN), 2019, pp. 1-4, doi: 10.1109/ViTECoN.2019.8705239.

[7] Y. Wang, H. Wu and J. Li, "Comparative study of PyCaret and scikit-learn in machine learning," 2021 IEEE 3rd International Conference on Advances in Computer Technology, Information Science and Communications (CTISC), 2021, pp. 68-73, doi: 10.1109/CTISC52264.2021.00022.

[8] K. S. Al-Mashaqbeh, A. Al-Ajlouni and O. Al-Khazali, "CRISP-DM for Predicting Future Digital Currencies Prices," 2019 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 2019, pp. 332-337, doi: 10.1109/CSCI48969.2019.00060.

[9] J. Wang, J. Cao, Y. Yu, and R. Wang, "Model selection and evaluation in big data analysis," IEEE Access, vol. 5, pp. 23153-23168, 2017, doi: 10.1109/ACCESS.2017.2761558.

[10] D. M. W. Powers, "Model selection: A review," in IEEE Signal Processing Magazine, vol. 28, no. 3, pp. 62-75, May 2011, doi: 10.1109/MSP.2010.939537.