Evaluating the CO_2 Emission from Gasoline-Powered Light-Duty Vehicles

1st Han Chau
Dept. of Mathematical Sciences
Stevens Institute of Technology
Hoboken, NJ, U.S.
hchau@stevens.edu

2nd Nihar Dugade
Dept. of Mathematical Sciences
Stevens Institute of Technology
Hoboken, NJ, U.S.
ndugage@stevens.edu

3rd Randy Duong
Dept. of Electrical and Computer Engineering
Stevens Institute of Technology
Hoboken, NJ, U.S.
rduong@stevens.edu

***Abstract*—Carbon dioxide (CO_2) is a greenhouse gas that contributes to global warming by trapping heat in the Earth's atmosphere. This can lead to negative impacts on both the environment and human society, such as rising sea levels, severe weather events, loss of biodiversity, and food and water insecurity. [1] Vehicles burning fossil fuels are one of the major sources of CO_2 emissions, which is a serious problem. This project aims to evaluate whether gasoline-fueled light-duty vehicles in Canada exceed the threshold of 251 grams of CO_2 emissions per kilometer driven, which is the current regulatory limit. The goal is to provide customers with information that can help them make better choices. To achieve this, the project uses machine learning algorithms, such as Logistic Regression,Random Forest and linear SVC, on a dataset of CO_2 emissions from light-duty vehicles in Canada from 2000 to 2022. By implementing and optimizing these models, we aim to gain insights into the compliance of gasoline-fueled vehicles with emission regulations**

I. Introduction

Climate change has a range of effects on the environment, society, and economy. The global rise in temperature has caused hotter temperatures, particularly in the Arctic, where temperatures have increased at least twice as fast as the global average. Climate change also leads to more severe storms, including tropical storms, which can cause deaths and huge economic losses. In addition, droughts are becoming more common, which leads to water scarcity for many people. Climate change also has significant impacts on biodiversity. The world is losing species at a rate 1,000 times greater than at any other time in recorded human history, which puts one million species at risk of becoming extinct within the next few decades. Changes in the climate and extreme weather events have caused a global rise in hunger and poor nutrition, affecting crops, livestock, and fisheries. Heat stress can also affect water and grasslands for grazing, leading to declining crop yields and affecting livestock. Climate change poses a significant health threat to humanity, causing air pollution, diseases, extreme weather events, forced displacement, mental health pressures, hunger, and poor nutrition. Environmental factors take the lives of around 13 million people every year. [2]

The increasing concern over environmental degradation and climate change has prompted the need to evaluate the carbon footprint of various industries and sectors, including the transportation sector. In particular, light-duty vehicles that use gasoline fuel for retail sale in Canada have been identified as significant contributors to carbon dioxide (CO_2) emissions, which are known to be a major greenhouse gas. In this study, we aim to evaluate whether these vehicles have exceeded the threshold of CO_2 emissions per kilometer driven from 2000 to 2022 by utilizing machine learning models and algorithms to address this problem. Our approach involves creating, implementing, and evaluating the effectiveness of several machine learning models, and optimizing them to achieve accurate and reliable results. By comparing different models and determining the best one to use, we aim to gain insights into the compliance of gasoline-fueled vehicles in Canada with emission regulations and identify areas for potential improvement in reducing CO_2 emissions from light-duty vehicles.

II. Related Work

The combustion of fossil fuels has caused a significant increase in carbon dioxide emissions since the industrial revolution. Most of the world's greenhouse gas emissions come from a relatively small number of countries. China, the United States, and the nations that make up the European Union are the three largest emitters on an absolute basis. Per capita greenhouse gas emissions are highest in the United States and Russia. [3]

Machine learning (ML) has become an important tool in modern society, helping to make various tasks easier and more efficient in fields such as medicine, business, science, environment, banking, and more. One of the most well-known examples of machine learning is Facebook's recommendation engine, which powers its news feed. [4]

Saleh et al (2016) suggested a solution for global warming by introducing a Support Vector Machine (SVM) model that predicts carbon dioxide (CO_2) emission expenditure. The model takes into account input variables such as electrical energy and burning coal, which directly affect CO_2 emissions. A trial and error method was employed to improve the prediction accuracy of the model by minimizing the error. The findings reveal that the SVM model was optimized with

a C parameter of 0.1 and an Epsilon value of 0. [5]

Hong, Tae-Hoon and colleagues employed Statistical Analysis in 2018 to forecast the CO_2 emission of Concrete using techniques such as Mann-Whitney test, ANOVA, Shapiro-Wilk W test, and Regression analysis. The outcome was a regression model that could predict the CO_2 emission of concrete based on the C/T ratio and slump value, taking into account its strength. Additionally, the validation of the proposed regression model's prediction performance indicated a significantly low error rate. [6]

In 2018, Pooja Kadam and colleagues utilized a supervised machine learning regression approach to forecast CO_2 emissions, and they evaluated the results using the Root Mean Square Error (RMSE) method. They concluded that a lower RMSE value corresponds to higher accuracy. However, the dataset appeared to follow a curve, and in this paper, linear regression was used to train the model, which may not be optimized for this scenario. [7]

As the environment is becoming a major concern, researchers have used machine learning algorithms to build models that can verify global warming and identify the factors contributing to it. For example, Harvey Zheng (2018) used data collected over the past 800,000 years to conclude that random forest is the best algorithm among others (such as lasso and support vector regression) to predict temperature with a larger set of features. [8]

Similarly, D. Deva Hema et al. (2019) found that CO_2 plays a major role in temperature change, followed by CH_4 and N_2O. They evaluated several machine learning algorithms (such as Linear Regression, Multi-Regression Tree, Support Vector Regression (SVR), and Lasso) to predict annual global warming and found that Linear Regression and Linear model are the best methods to predict and forecast temperature and greenhouse gases for the next 10 years on average. However, these authors only considered the amount of CO_2 affecting the environment, without exploring the sources of CO_2 emissions. [9]

In this project, we focus on evaluating CO_2 emissions from light-duty vehicles using fossil fuel (gasoline) and deciding whether these cars will have a harmful impact on the environment if their emissions exceed the allowed level.

III. Our Solution

To provide a solution to the problem of evaluating whether gasoline-fueled light-duty vehicles exceed the allowable threshold for CO_2 emissions per kilometer driven, we propose the following approach:

A. Description of Dataset

The dataset used in this project was taken from Kaggle and contains information on fuel consumption ratings and estimated carbon dioxide emissions for vehicles from 2000 to 2022. The data is in CSV format with 22.556 rows and 13 columns, including attributes such as year, model, vehicle class, fuel, fuel consumption, and emissions. [10]

To perform our approach, examples that were not gasoline were dropped, and the emission attribute was converted into a binary label based on a threshold calculated from the amount of CO_2 emissions from driving one kilometer. We specifically calculate the threshold based on the amount of CO_2 emissions from a gallon of gasoline, and the average gasoline vehicle on the road today has a fuel economy of about ≈ 35.4kilometers per gallon. [11] Therefore, the average vehicle, when driving one mile, has tailpipe CO_2 emissions of about 251 grams per kilometer:

$$CO_2\text{per km} = \frac{CO_2 \text{ per gallon}}{MPG} = \frac{8.887}{35.4} = 251 \text{ grams}$$

Various visualization techniques were employed to gain insights into the dataset. One such visualization is a histogram that displayed the distribution of features. It was observed that most features followed a normal distribution, except for the year column.

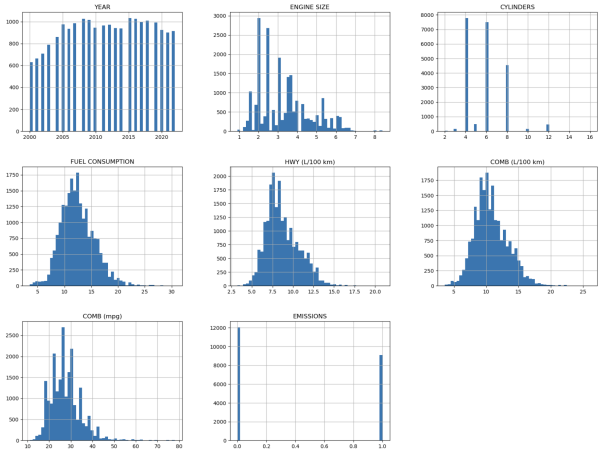

Fig. 1. A histogram for each numerical attribute

A scatterplot was also generated to investigate the relationship between fuel consumption and emissions. The scatterplot illustrated that the data was separated into two distinct groups based on the label of the emissions attribute.

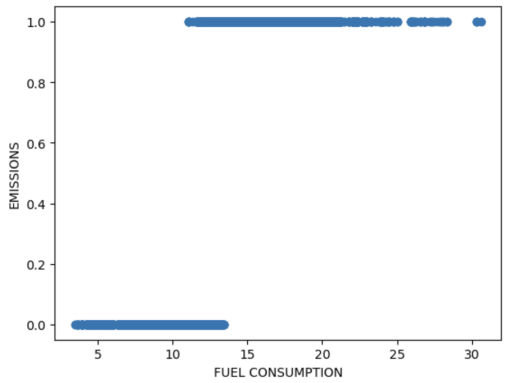

Fig. 2. A scatter plot for fuel consumption and emission

The correlation coefficient of each attribute with the emissions ranged from -0.7 to 0.7, with most attributes exhibiting strong positive correlation, and the comb attribute displaying strong negative correlation. [12] The correlation matrix heatmap depicted the strength of the correlations between pairs of variables, revealing strong correlations between most variables, except for the year attribute. The year attribute in this context does not represent a numerical value, but rather indicates categories based on the year from 2000 to 2022. It is important to consider the relationship between the year category and the emission variable, as there are 23 distinct year categories in total. Hence, we retain this attribute as part of our training data for the model.

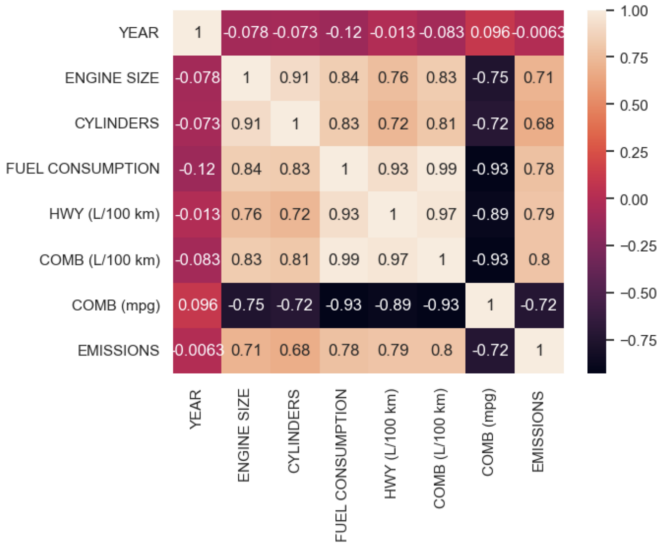

Fig. 3. A correlation matrix for the numerical variables

As the amount of CO_2 produced from burning one gallon of fuel is influenced by the carbon content in the fuel, a threshold was established for estimating CO_2 emissions from fuel consumption. As a result, this feature was removed from the dataset to prevent overfitting when training the model. In order to ensure the accuracy of our analysis, we employed the Tukey method to identify and remove any outliers. The fuel consumption variable's interquartile range (IQR) was calculated and values outside of 1.5 times the IQR were considered outliers, leading to the detection and removal of 340 outliers from the dataset. [13]

To prepare the dataset for training, we used the OrdinalEncoder() method to convert the string values into the numerical value. The final dataset for implementing the machine learning model was obtained by concatenating the matrix of label code with the other attributes. Lastly, we split the data into training and testing sets using a 90:10 ratio.

B. Machine Learning Algorithms

To implement the model for this project, choosing multiple models is a common practice in machine learning to increase the chances of finding the best performing model for a specific problem. Each model has its own strengths and weaknesses, and the performance of a model can vary depending on the specific dataset and problem at hand.This is a binary classification problem, and the features are highly relevant to the target variable. The sample size is large enough to accurately estimate the coefficients, and the dataset is linearly separable. [14]. Hence, we have chosen five models to train a binary classification problem:

- Logistic regression: For binary classification issues, logistic regression is a straightforward and understandable approach. It can handle a high number of characteristics and can be applied to both linear and nonlinear decision boundaries. The likelihood of each class is also provided via logistic regression, which is helpful in some applications. [15]

- LinearSVC: A linear model called a Linear Support Vector Classifier (SVC) seeks out the ideal hyperplane that divides the two classes. It is effective when the data can be separated linearly and is computationally efficient, especially for large datasets.

- Random Forest: To enhance classification performance, Random Forest is an ensemble model that blends various decision trees. In comparison to a single decision tree, it can handle nonlinear decision boundaries and is less prone to overfitting. Additionally, Random Forest can deal with irrelevant characteristics and missing values. [16]

- Neural network: A neural network is an effective model that can discover intricate patterns and connections in data. It consists of many layers of nodes that alter the input characteristics nonlinearly. Large datasets may be handled by neural networks, which can also be applied to both linear and nonlinear decision limits.

- Naive Bayes: Naive Bayes is a straightforward and quick model that excels in classifying text as well as other

high-dimensional datasets. It determines the likelihood of each class based on the occurrence of each feature and makes the assumption that the features are independent. Naive Bayes can be applied to problems involving binary and multiple classes in classification. [15]

By default, we initialized the parameters for the models excepting the random forest to see how they perform on our dataset. However, we also plan to optimize the models by searching for the best parameters using various techniques, such as SelectKBest(), StandardScaler(), GridSearchCV(), or RandomizedSearchCV(). These techniques will help us fine-tune the models to achieve better accuracy and performance.

C. Implementation Details

Initially, the logistic regression model achieved a high accuracy score of 97.44%. However, to further improve its performance, we utilized GridSearchCV to optimize the model's parameters, resulting in a 1.2% increase in accuracy. By incorporating Pipeline and GridSearchCV techniques to select the top features and optimal parameters, the accuracy increased by 1.6%.

The accuracy of the linear SVC model was 97.35% initially, but cross-validation showed a lower accuracy of 92.97%, revealing the model's bias. To improve the model's accuracy, we utilized GridSearchCV to find the best parameters and obtained an accuracy of 98.67%.

For the Random Forest model, we initially set the parameters to n_estimators = 5, max_leaf_nodes = 5, random_state = 42, which achieved an accuracy of 99.13%. To optimize hyperparameters, we used RandomizedSearchCV instead of GridSearchCV, which randomly samples hyperparameters from a distribution and evaluates them, reducing computation time, including the number of trees, the depth of each decision tree, and the number of features, and obtained the best hyperparameters, resulting in an accuracy of 99.83%. [17]

The neural network model had a lower initial accuracy of 92.23%, but the average accuracy improved to 94.92% after using cross-validation. We used RandomizedSearchCV to optimize the model's parameters and obtained an accuracy of 98.67%.

The Naïve Bayes model initially achieved a high accuracy score of 95.8%. However, after using GridSearchCV to optimize the var_smoothing parameter, the model's performance worsened. To improve its accuracy, we incorporated feature selection techniques, such as SelectKBest with f_classif score function, and resampling techniques. By defining a pipeline to combine the feature selection step with the classifier and using GridSearchCV to search for the best value of k, we significantly improved the accuracy score.

IV. Comparison

A. The classification report

Based on the accuracy score and classification report, we can make the following observations about the five models:

Random Forest has the highest accuracy score (99.86%) and f1-scores for both classes are 1.00 which indicates the model is doing exceptionally well in predicting both classes.

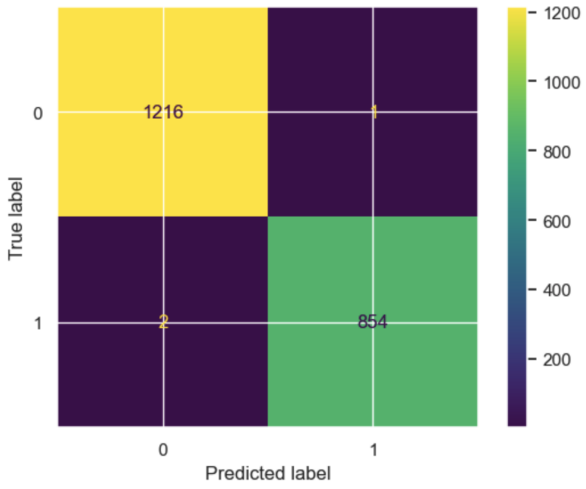

Fig. 4. A confusion matrix for the predicted output and the actual values for the Random Forest model

Logistic Regression has an accuracy score of 98.65% which is close to Random Forest, and its precision, recall and f1-scores are also high, indicating a good performance.

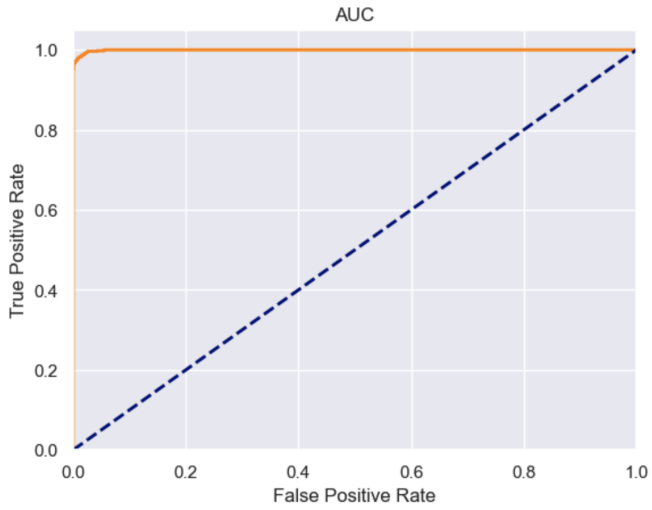

Fig. 5. Plot the AUC for the Logistic Regression model

Linear SVC has an accuracy score of 97.88% which is slightly lower than the other two models, but still performs well in terms of precision, recall and f1-scores.

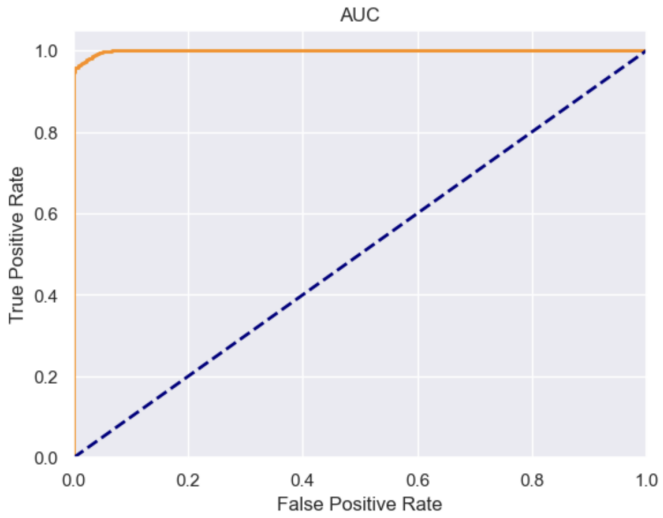

Fig. 6. Plot the AUC for the Linear SCV model

Neural Network has an accuracy score of 97.06% and its f1-score is not as high as the other models for predicting the class 1.

```
Classification report:
              precision    recall  f1-score   support

           0       0.95      1.00      0.98      1217
           1       1.00      0.93      0.96       856

    accuracy                           0.97      2073
   macro avg       0.98      0.96      0.97      2073
weighted avg       0.97      0.97      0.97      2073
```

Fig. 7. The classification report for the Neural Network model

Naive Bayes has the lowest accuracy score among the five models (96.33%) and its precision, recall and f1-scores are slightly lower than the other models.

Overall, Random Forest seems to be the best performing model in this case based on the accuracy score and the f1-scores for both classes. However, Logistic Regression also performs very well and might be a better choice if the interpretability of the model is important.

B. The amount of time train the optimized model

We set a function to determine the time it takes to train and predict using five different models: logistic regression, SVM, random forest, neural network, and naive Bayes.

```
model_name =["logistic","svm",\
"random forest","neural network",\
"naive bayes"]
model_time = [0]*len(model_name)
for i in range(len(model_name)):
    start = time.time()
    model_train = model(train_x,train_y,\
    test_x,test_y,model=model_name[i])
    end = time.time()
    time_count = end-start
    model_time[i]=round(time_count,2)
dict_time=dict(zip(model_name,model_time))
print(dict_time)
```

The time taken for training and prediction varies for each model, and the results are as follows:

- Logistic regression: 0.05 seconds
- SVM: 0.83 seconds
- Random forest: 0.73 seconds
- Neural network: 1.43 seconds
- Naive Bayes: 0.01 seconds

From the above results, we can see that the Naive Bayes is the fastest model to train and predict, whereas the neural network is the slowest. The SVM model also takes a considerable amount of time to train and predict, followed by the random forest model. The logistic regression model is the fastest among all the other models to train and predict.

In summary, the Random Forest model demonstrated the highest accuracy score and f1-scores for both classes, making it the top performer. However, if interpretability is a priority, Logistic Regression also showed strong performance and might be a better option. Additionally, the logistic regression model was the fastest to train and predict, making it a good choice for time efficiency. Considering the trade-offs between accuracy, interpretability, and time efficiency, the logistic regression model could be a suitable option.

V. Future Directions

There are several potential avenues for future research:

Incorporating more data: The analysis could benefit from including more data on the types and amounts of emissions from cars that impact the environment.

Further exploring feature engineering: In addition to the feature selection method used in this study, future research could investigate other feature engineering techniques such as feature extraction or scaling to identify the most crucial aspects in the analysis.

Leveraging ensemble learning: Ensemble learning, which involves combining the predictions of multiple models, could be explored to improve the performance and robustness of the models in this analysis. Hyperparameter tuning: The performance of the models is greatly influenced by their hyperparameters. Further tuning of these hyperparameters could enhance the models' performance.

Transfer learning: Transfer learning, which involves fine-tuning a pre-trained model on a related task, could be used to enhance the performance of the models, particularly for the neural network model.

Model explanation: As interpretability is important, future research could explore methods to explain the models' decisions and predictions, such as feature importance analysis or partial dependence plots.

VI. Conclusion

Based on the analysis, it can be concluded that the Logistic Regression, LinearSVC, Random Forest, Neural Network, and Naive Bayes models can be trained for binary classification problems. However, to optimize the models, some techniques can be used such as GridSearchCV, RandomizedSearchCV, SelectKBest(), and StandardScaler(). Cross-validation techniques such as cross_val_score() can also be used to evaluate the models' performance on different subsets of data.

In the Logistic Regression model, the combination of GridSearchCV and Pipeline led to the accuracy score of the test set of 98.5%. The LinearSVC model's accuracy was improved by GridSearchCV, and the accuracy score obtained was 97.88%. RandomizedSearchCV improved the Random Forest model, resulting in the highest accuracy score of 99.86%. For the neural network model, the accuracy score obtained was 97.06% after using RandomizedSearchCV. Finally, the Naive Bayes model's accuracy improved significantly after combining the feature selection techniques or the resampling techniques.

Our findings suggest that the Logistic Regression model is a suitable choice for predicting and evaluating whether a car's CO_2 emission exceeds the allowed threshold based on our dataset. However, consdering the other dataset and further optimization of the model's hyperparameters is necessary to improve its accuracy and generalization performance.

References

[1] M. Kayakuş, *Forecasting carbon dioxide emissions in Turkey using machine learning methods*, 2022, vol. 28.

[2] U. Nations, "Causes and effects of climate change," https://www.un.org/en/climatechange/science/causes-effects-climate-.

[3] "Global emissions," https://www.c2es.org/content/international-emissions/.

[4] E. Burns, "machine learning," https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML, 2022.

[5] C. S. et al, "Carbon dioxide emission prediction using support vector machine," https://iopscience.iop.org/article/10.1088/1757-899X/114/1/012148/meta, 2016.

[6] T.-H. Hong, "Predicting the co2 emission of concrete using statistical analysis," https://koreascience.kr/article/JAKO201218559656140.page, 2012.

[7] P. Kadam and S. Vijayumar, "Prediction model: Co2 emission using machine learning," https://ieeexplore.ieee.org/abstract/document/8529498, 2018.

[8] H. Zheng, *Analysis of Global Warming Using Machine Learning*, 2018, vol. 07.

[9] V. L. R. G. D. Deva Hema, Anirban Pal, "Global warming prediction in india using machine learning," https://d1wqtxts1xzle7.cloudfront.net/84056081/A1301109119-libre.pdf?1649844210=&response-content-disposition=inline.

[10] J. YILMAZ, "Fuel consumption 2000-2022," https://www.kaggle.com/datasets/ahmettyilmazz/fuel-consumption, 2022.

[11] U. S. E. P. Agency, "Greenhouse gas emissions from a typical passenger vehicle," https://nepis.epa.gov/Exe/ZyPDF.cgi?Dockey=P100U8YT.pdf.

[12] A. Géron, *Hands-on Machine Learning with Scikit-Learn, Keras TensorFlow*, 2019, vol. 02.

[13] F. Karabiber, "What is binary classification?" https://www.learndatasci.com/glossary/binary-classification/.

[14] S. Chatterjee, "Good data and machine learning," https://towardsdatascience.com/data-correlation-can-make-or-break-your-machine-learning-project.

[15] J. Brownlee, "4 types of classification tasks in machine learning," https://machinelearningmastery.com/types-of-classification-in-machine-learning/.

[16] IBM, "What is random forest?" https://www.ibm.com/topics/random-forest.

[17] P. M. Kouate, "Machine learning: Gridsearchcv randomizedsearchcv," https://towardsdatascience.com/machine-learning-gridsearchcv-randomizedsearchcv-d36b89231b10.