Ministry Of Higher Education and Scientific Research

University Of Kasdi Merbah Ouargla

Faculty of New Technologies of Information and Communication

Department of Computer Science and Information Technology

Dissertation

Presented for the attainment of the degree of

Academic Bachelor in Computer Science

Topic

SVM-based classification of fruit images

Presented By:

- ABANOU Baya
- BENTAYEB Ayat Arrahmane

Supervised By:

- MAZOUZ Mihoub

Academic Year: 2022/2023

Table of content

List of figures

الملخص

نظرًا للتطور الكبير في التكنولوجيا في معظم المجالات والقطاعات، من الصناعة والصحة والتسويق والإعلام إلى الزراعة، فقد خصصنا مجال التهجين، وخاصة تهجين الفاكهة.

تهجين الفاكهة هو عملية إنتاج فاكهة جديدة بواسطة التلقيح الاصطناعي لنوعين مختلفين، ويمكن أن يحدث التهجين تلقائيًا في الطبيعة. بسبب درجة التشابه العالية بين أنواع معينة، غالبًا ما يكون من الصعب التمييز بين سلالات مختلفة من الفاكهة. في هذا العمل، نطور برنامجًا يستخدم التعلم الآلي وطريقة SVM للتعرف على الثمار من صورهم.

الكلمات الرئيسية: بايثون، الذكاء الاصطناعي، آلة التعلم، آلة ناقل الدعم (SVM)

ABSTRACT

Due to the significant development in technology in most areas and sectors, from industry, health, marketing, and media to agriculture, we have dedicated the field of hybridization, especially fruit hybridization.

Fruit hybridization is the process of producing new fruit by IVF for two different species, and hybridization can occur automatically in nature. Because of the high degree of similarity between certain species, it is often difficult to distinguish between different strains of fruit. In this work, we develop software that uses machine learning and the SVM method to recognize fruits from their images

Key words: Python, Artificial Intelligence, Learning Machine, Support vector machine (SVM)

Introduction

Building models and systems capable of learning, adapting, and making decisions based on available data is the goal of machine learning, an important and evolving field of artificial intelligence. Machine learning is the process of detecting patterns and knowledge in data and using them to make decisions or carry out activities in a smart and effective manner.

Data is a key component of machine learning, and models and systems are trained using different data sets relevant to the task at hand. These data are examined and interpreted by machine learning algorithms, which then extract patterns, reports, and classifications from them.

Machine learning technology includes deep learning, Hebrew learning, learning transfer, and more. For her, these innovations stand out.
In our first chapter, we will radicalize the definition and types of machine learning. In the second axis, we identified the classification in machine learning as well as the SVM algorithm that will be the focus of our attention, and in chapter III, entitled "Implementation", we will address the nucleus of our research, "Fruit classification," and the stages of its completion from the language used to the data sets and pre-processed data sets and their implementation.

Chapter 1

Introduction to Machine Learning

1. Introduction

As interest in artificial intelligence grows, the term "machine learning" appears more frequently.

Machine learning is the ability of computer systems to learn and improve by interacting with data. Instead of explicitly teaching and programming computer systems in detail, they have swept across a wide range of fields, including health, economics, and financial engineering, among others.

In this chapter, we will first define machine learning before moving on to its various types and the majority of its algorithms.

2. Machine learning

Machine learning is a branch of artificial intelligence which is concerned with creating and developing algorithms that allow for the building of learned systems capable of self-improvement based on a set of data that is sorted. In other words, it is a way to make the machine learn to extract data patterns and predict them to make decisions about new data.

Or, as Tom Mitchell knew it : "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E" **[1]** .

Machine learning can be divided into three categories: supervised learning, unsupervised learning and Reinforcement learning.

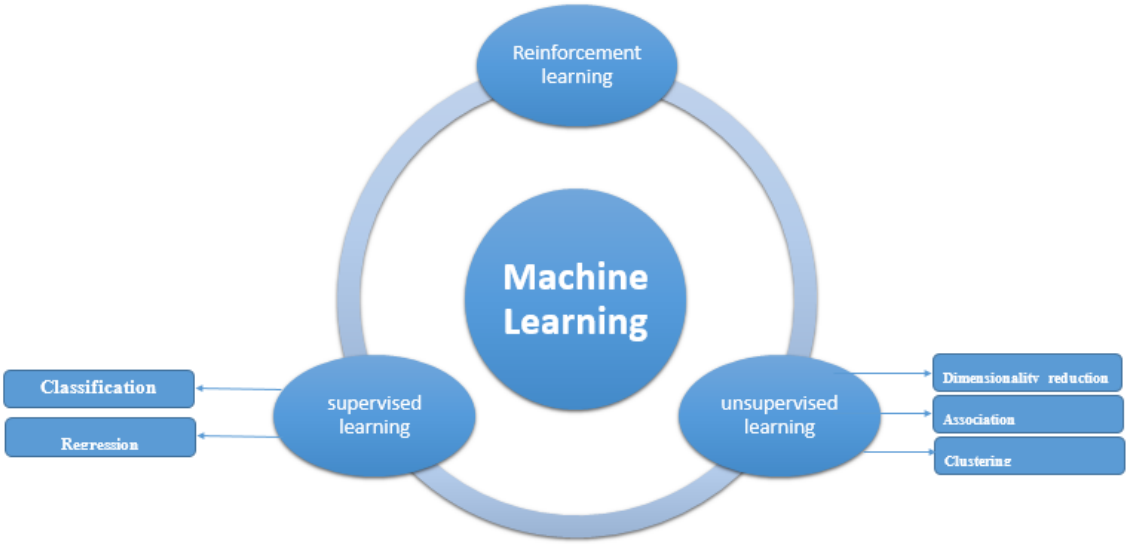

Figure 1.1 : Categories of Machine Learning

2.1 Supervised learning

Supervised learning involves training the machine on a set of data with specific outputs. The algorithm collects the patterns and links them, forming a database under the name of the model. The model is then used to predict the outputs of new input data.

Supervised learning includes several algorithms that are used depending on the research area and the problem to be solved, and the most famous of these are regression and classification.

a. Regression

Regression entails predicting a continuous variable output. When a regression-based system predicts a value, the number of prediction errors is used to assess performance.

For evaluating regression models, statistical methods such as the coefficient of determination, mean absolute error, relative absolute error, and root mean square error are used.

Linear regression, regression trees, support vector machines, k nearest neighbor, and perceptrons are examples of regression algorithms. **[2]**, These algorithms contain outputs that are real or countable. For example, height (4 feet, 5 feet, 6 feet), age (27, 31, 65), or price (100 rupees, 20 pounds, 10 dollars). **[3]**

b. Classification

Classification is a data mining (machine learning) method for predicting group membership for data instances. Although there are numerous machine learning techniques available, classification is the most widely used. Classification is a highly valued task in machine learning, particularly in future planning and knowledge discovery. **[4]**

2.2 Unsupervised learning

Unsupervised learning involves training the machine on a set of Unlabeled data, where it classifies and divides the data after extracting similarities between them, or in other way, the machine learns to determine patterns without direct guidance, we let the computer learn on its own to find the patterns, trends, and facts.

So when we use unsupervised learning, you can say that the computer program is not being explicitly programmed to learn. It is learning on its own by discovering the facts, patterns, and trends. But we do program it by selecting the algorithms it will use to discover them. **[5]**

a. Clustering

One of the most fundamental types of unsupervised learning is clustering. The process of discovering relationships within data to split data into clusters based on similarity is the main goal.

As an example:

Clustering is used in a variety of medical applications, including the following:

• Grouping patients with similar profiles together for monitoring

• Detecting anomalies or outliers in claims or transactions

• Defining treatment groups based on medication or condition **[6]**

b. Association

The process of identifying associations between different observations using provided data is known as association. Association rule learning **[7],** for example, has historically been best applied to online shopping checkout basket datasets gathered on users' purchasing habits. For example, when someone buys a pizza, they may also buy wine, just as someone who buys lettuce may also buy tomatoes, cucumbers, and onions. The likelihood of associations can be predicted by analyzing transactional datasets. We know that certain items (whether food, clothing, or disease) frequently occur together, and association rule learning attempts to comprehend these relationships.**[8]**

c. Dimensionality reduction algorithms

Although the message is generally "the more data, the better," datasets can often contain many variables, making capturing the signal of the data a more difficult task. Dimensionality reduction algorithms are very useful to data scientists for a variety of

reasons, including sparse values, missing values, identifying relevant features, resource efficiency, and more straightforward interpretation.**[9]**

2.3 Reinforcement learning

Reinforcement learning: Reinforcement learning, where ML algorithms learn through their interaction in an environment, where the ML algorithm obtains feedback about the accuracy of its response. **[10]**

This method is similar to training a pet, It sends positive signals to the machine when it gives the desired output, to let it know that it is right and to help it learn better. Similarly, it sends negative signals to a machine if it provides an incorrect output.**[11]**

3. Conclusion

Finally, machine learning (ML) is a powerful artificial intelligence field.

It has numerous applications in a variety of fields, including image recognition, natural language processing, fraud detection, personal recommendations, and predictive analysis. Businesses and organizations can gain valuable insights from their data, automate processes, improve decision-making, and improve overall efficiency and performance by leveraging ML techniques, It has the potential to revolutionize many industries and drive innovation in many areas as it advances. ML algorithms, with their ability to learn and adapt, can address complex problems and contribute to real-world challenges.

Chapter 2

Classification

1. Introduction

Classification is one of the most important and widely used techniques in machine learning. It is a type of supervised learning, where the algorithm is trained on a labeled dataset to predict the class label of a new and unseen data point.

The classification problem will be defined in this chapter before its many forms and the majority of its algorithm.

2. Classification Problem

Classification techniques include estimation and prediction. categorization methods including Naive Bayes, Support Vector Machines (SVM), and Bayesian based on Bayes' theorem. The primary goal of classification in data mining is to assign objects in a group to training class categories and accurately forecast the test set for each case in the data. The sympathetic of Classification must have a basic understanding of training data **[12].** A well-known technique for extracting useful information that relates to how a machine learns.

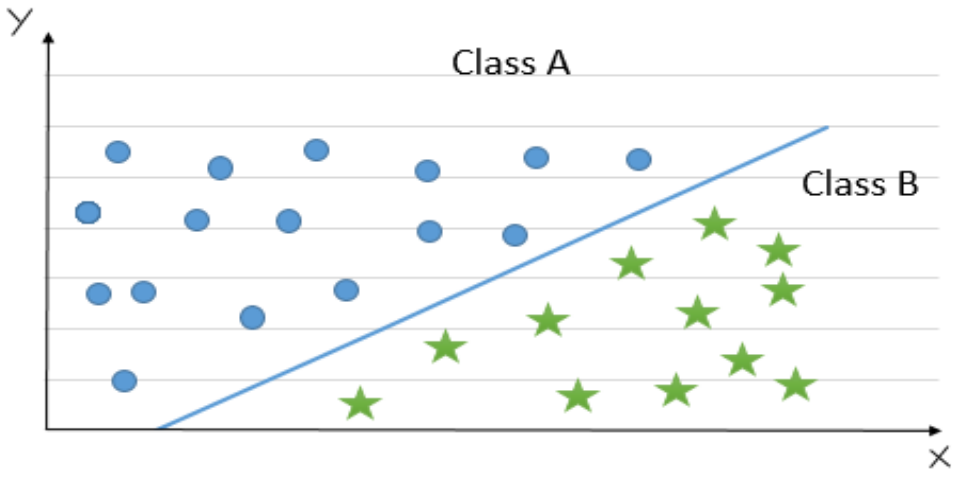

Figure 2.1 : Classification in Machine Learning

3. Types of classification

3.1 Binary Classification

Binary is a type of problem in classification in machine learning that has only two possible outcomes. For example, yes or no, true or false, spam or not spam, etc. Some common binary classification algorithms are logistic regression, decision trees, simple bayes, and support vector machines.

3.2 Multi-Class Classification

Multi-class is a type of classification problem with more than two outcomes and does not have the concept of normal and abnormal outcomes. Here each outcome is assigned to only one label. For example, classifying images, classifying species, and categorizing faces, among others. Some common multi-class algorithms are choice trees, progressive boosting, nearest k neighbors, and rough forest.

3.3 Multi-Label-Classification

Multi-label is a type of classification problem that may have more than one class label assigned to the data. Here the model will have multiple outcomes. For example, a book or a movie can be categorized into multiple genres, or an image can have multiple objects. Some common multi-label algorithms are multi-label decision trees, multi-label gradient boosting, and multi-label random forests.

3.4 Imbalanced Classification

Most machine learning algorithms assume equal data distribution. When the data distribution is not equal, it leads to imbalance. An imbalanced classification problem is a classification problem where the distribution of the dataset is skewed or biased. This method employs specialized techniques to change the composition of data samples. Some examples of imbalanced classification are spam filtering, disease screening, and fraud detection **[13].**

4. Classification Algorithms

Classification algorithms have many different types in terms of accuracy and usage. In this title, we will touch on some types, and we will devote the next title to the SVM algorithm that will be the nucleus of our search.

4.1 Logistic Regression

It is a supervised learning classification technique that forecasts the likelihood of a target variable. There will only be a choice between two classes. Data can be coded as either one or yes, representing success, or as 0 or no, representing failure. The dependent variable can be predicted most effectively using logistic regression. When the forecast is categorical, such as true or false, yes or no, or a 0 or 1, you can use it. **[14]**

4.2 Decision Tree

Decision trees are non-parametric techniques that have a structure akin to a tree or a flowchart and can be used to categorize issues **[15].**

A tree structure is used to model the many links between the features and the potential output data in decision trees, which are effective algorithms for classifying data. The reason the ML method has this name is because it mimics the way a real tree grows, starting with a broad trunk and branching out into narrower branches as it goes up. Similar to this, DT employs a branching choice architecture, starting with the primary question for a given problem that must be resolved in order to move on to the secondary question that must be resolved in order to continue decomposing the data and classifying results.

4.3 K- Nearest Neighbor

In the K-nearest neighbor (KNN) method, the distance between two neighbors is assessed in relation to the value of k, which determines how many neighbors must be examined in order to describe the class of a sample data point **[14].**

The two types of nearest neighbor algorithms are structure-based KNN and structure-less KNN. The structure-based approach deals with the data's fundamental structure, which has fewer mechanisms connected to training data samples. **[16]**

4.4 Naive Bayes

A Bayesian Network (BN) is a graphical representation of the probabilistic relationships between a group of variables **[17].**

The directed acyclic graph (DAG) that makes up BN structure S's nodes is in one-to-one communication with the X features. The arcs represent unanticipated effects between nodes, and S's lack of arcs encodes conditional liberties **[18].** Normal Bayesian network learning tasks can be broken down into two subtasks: (a) learning the network's DAG structure, and (b) determining the parameters.

5. Support Vector Machine

The support vector machine (SVM) carries out the following: It transforms the input vectors x into a high-dimensional feature space z through some nonlinear mapping that is predetermined. In this region, an optimal separation hyperplane is constructed **[19].**

5.1 hyperplane

In geometry, a hyperplane is a subspace of one dimension less than its ambient space. Given the linearly separable data, we can use a hyperplane to perform binary classification. The hypothesis is defined as follows:

$$h(x_i) = \begin{cases} +1 & if\ w.x_i + b \geq 0 \\ -1 & if\ w.x_i + b < 0 \end{cases}$$

5.2 kernel method

A kernel is a function that outputs the outcome of a dot operation carried out in another location. Formally speaking, we could write :

$$K(x, x') = \langle \emptyset(x), \emptyset(x')v \rangle$$

It has several types:

- **Linear kernel** it defined by $k(x, x') = x.x'$
- **Polynomial kernel** it defined by $k(x, x') = (x.x' + c)^d$
- **RBF kernel** it defined by $k(x, x') = \exp(-\gamma\|x - x'\|^2)$. **[20]**

6. Conclusion

In the field of machine learning, classification is used to classify data into different categories based on available attributes and properties. The main objective of classification is to build a model that is used to predict the class of the new sample based on the available data. We have shortly described some algorithms of classification as logistic regression which is suitable for binary classification as well as SVM algorithm.

Chapter 3

implementation of the model

1. Introduction

Fruit classification and differentiation is a difficult procedure that involves human expertise, patience, and effort. But as AI develops quickly, we can use it to dramatically enhance fruit classification.

We shall investigate the use of artificial intelligence technologies in fruit classification in this study.

2. Fruits Classification

In the context of machine learning, the classification of fruits refers to the process of identifying the type of fruit based on its specific characteristics. Fruit data is processed using machine learning techniques to build a model that can recognize and categorize various fruits according to their characteristics.

A group of distinguishing characteristics that can be utilized to categorize fruits are discovered. Color, size, form, texture, and other attributes are examples of characteristics. In the training dataset, these characteristics of each fruit are measured.

3. Model implémentation

3.1 Language & tools

a. Python

It is very common for the programming language Python to be one of the most widely used programming languages in the world because to its advantages over other programming languages, especially in the field of education.

Python was used in this work since it can be automated.

The Python programming language was created in the Netherlands by Guido van Rossum near the end of the 1980s. Python was developed after ABC, a programming language that can handle exceptions and communicate with the "Amoeba" operating system. People have been using it since 1989.

Unlike other programming languages like Java, the Python programming language allows programmers to express concepts with a very little number of lines of code.[21]

b. IDE

Google Research created "Colab" as a product. Colab is particularly well suited to machine learning, data analysis, and education. It enables anyone to create and execute arbitrary Python code through the browser. Technically speaking, Colab is a hosted Jupyter notebook service that offers free access to computer resources, including GPUs, and requires no setup to use. [22]

c. Different Libraires

In this work, we used a number of libraries to create a model that allows us to recognize fruit, which is as follows :

- **NumPy**

The most essential Python package for numerical computing is called NumPy.

Python and C are both used to create NumPy (for speed). The following are a handful of NumPy's key features as described on its website:

• A fast and efficient multidimensional array object ND array

• Functions for performing element-wise computations with arrays or mathematical operations between arrays

• Tools for reading and writing array-based data sets to disk

• Linear algebra operations, Fourier transform, and random number generation

• Tools for integrating connecting C, C++, and Fortran code to Python. [23]

- **Pillow**

All of the fundamental image processing capabilities are available in the Pillow library. You can modify, rotate, and resize images.

You can extract some statistics from an image using the Pillow module's histogram function, and then use those statistics for automatic contrast amplification and statistical analysis. [24]

3.2 Data set

a. Description

A data set is a structured collection of data that is organized and stored for analysis, research, or machine learning. It is a key resource for a variety of data-driven tasks as well as a foundation for generating valuable insights and training models for decision-making.

Spreadsheets, databases, text files, and specialized file formats designed for specific types of data, such as images or audio, are all examples of datasets. It can contain various types of data, such as numerical or textual information.

We will use an image data set in this study.

Horea Muresan compiled the list under the heading "Image Classification Network for Fruit Datasets".

From Kaggle (URL: https://www.kaggle.com/code/mitch9090/fruit-dataset-image-classification-network/input)

We selected a range of fruits to undergo our model training and testing:

[Apple Red 1 , Kiwi , Grape White , Cherry 1 , Mango , Pear , Avocado , Apple Red Yellow 1 , Banana , Banana Red , Corn , fig , Orange]

b. Preprocessing

As a machine learning engineer, data preparation or cleansing is a critical step, and most ML engineers spend a significant amount of time doing so before building the

model. External detection, lost value treatments, and the removal of unwanted or noisy data are some examples of advanced data processing.

In our case, the data we have is essentially net, so we don't need to purify the impurities or adjust their size, leaving us with one stap:

Upload photos and turn them into vector

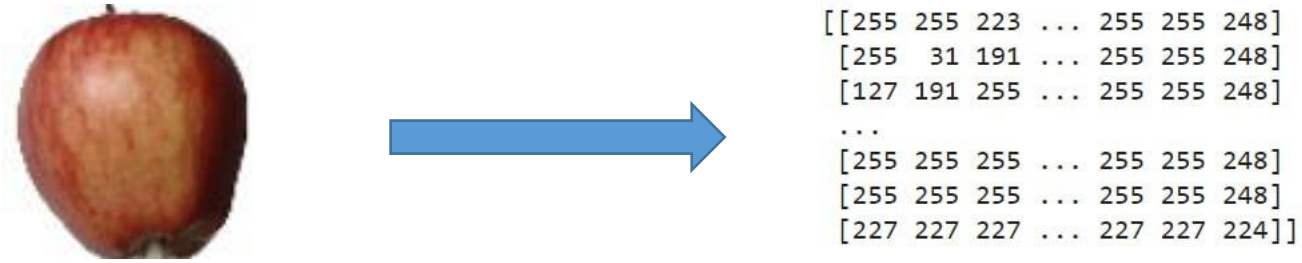

Figure 3.1 : Preprocessing image

3.3 SVM-based Model

a. Approach

- **Input**

In machine learning, the term "input" refers to the data provided to the training or prediction machine learning model. It denotes information or data that can be used to make predictions or infer patterns. Machine learning input forms vary depending on the problem and nature of the data, as a machine learning model can process numerical values, mandatory variables, texts, images, voice signals, or any other type of data.

Here we use a photo-themed dataset to address our problem

for example

Figure 3.2 : Inputs image

- **Output**

In machine learning, the term "output" refers to the results or predictions provided by the machine learning model based on the inputs provided to it. Output can be a projected digital value, such as predicting the next day's share price, or classifications, such as classifying images as "dog" or "cat."

The efficiency and accuracy of the machine learning model, as well as its ability to predict results correctly and accurately based on inputs provided to it, determine output quality.

In this project, our output is classifications where value is string

for example:

Appel Red Mango Cherry Kiwi

These are some examples of inputs and their outputs:

Figure 3.1: inputs and their outputs

b. Training

Simply said, training a model entails learning (deciding) appropriate values for each weight and bias from labeled samples. In supervised learning, an algorithm uses a

large number of instances to develop a model in an effort to discover the model with the lowest possible loss.

In our problem, features were extracted from the vector previously extracted in the preprocessing of the model training.

c. Test

After a machine learning software has been trained on an initial training data set, it is tested using a secondary (or tertiary) data set called a test set. The concept is that, as opposed to being examined from a programming standpoint, predictive models always contain some form of untested capability that needs to be tested.

d. Evaluation

After studying this model we reached satisfactory results, as it became easy to recognize the fruit from its own image.

4. Difficulties

The following issues were encountered as the program was being developed: how to extract fruit features from images like size, color, and other features in order to train the model to recognize fruits

5. Conclusion

We applied the SVM algorithm to fruit image classification in this axis, where we converted the DataSat consisting of images into matrices and then used it as an input to train the growth with it, and the results were good, so we classified most species correctly.

General conclusion

After careful consideration of this issue, solutions were found to the problems highlighted. The goal was to identify the fruit through its image and definition. The results were satisfactory and adjustable in terms of using the idea in purchasing applications (fruits) or in shops.

References

[1]: Mitchell, T. (1997)." Machine Learning". p. 2

[2]: Panesar, A. (2019). Machine Learning Algorithms. In: Machine Learning and AI for Healthcare p128

[3]: Nikita Silaparasetty, "Machine Learning Concepts with Python and the Jupyter Notebook Environment", 2020, p32

[4]: A,.A. Soofi , "Classification Techniques in Machine Learning: Applications and Issues",August 2017, INTRODUCTION ,p127

[5]: Puneet Mathur, 2019" Machine Learning Applications Using Python - Cases Studies from Healthcare, Retail, and Finance by Puneet Mathur (z-lib.org)", Process of Technology Adoption, p2

[6]: Panesar, A. (2019). Machine Learning Algorithms. In: Machine Learning and AI for Healthcare p158

[7]: Nikita Silaparasetty, "Machine Learning Concepts with Python and the Jupyter Notebook Environment", 2020, p37

[8]: Panesar, A. (2019). Machine Learning Algorithms. In: Machine Learning and AI for Healthcare p160

[9]: Panesar, A. (2019). Machine Learning Algorithms. In: Machine Learning and AI for Healthcare p162

[10]: Pineda-Jaramillo, Juan D,2019 "A review of Machine Learning (ML) algorithms used for modeling travel mode choice"p4

[11]: Nikita Silaparasetty, "Machine Learning Concepts with Python and the Jupyter Notebook Environment", 2020, p37

[12]: C. Jalota and R. Agrawal, "Analysis of Educational Data Mining using Classification," 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Faridabad, India, 2019, pp. 243-247, doi:10.1109/COMITCon.2019.8862214

[13]: R. S. Gaur and A. Kumar, "Binary Classification in Machine Learning: A Review," International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 3, no. 3, pp. 451-

458, 2018. [Online]. Available: https://www.ijsrset.com/paper/1334.pdf. [Accessed: 02-May-2023].

[14]: Samuel, A., Some studies in Machine Learning using the game of checkers. IBM Journal of Research and Development, 3(3), pp. 210-229, 1959. DOI: 10.1147/rd.33.0210

[15]: Bhavsar, P., Safro, I., Bouaynaya, N., Polikar, R. and Dera, D., Machine learning in transportation data analytics. In: Chowdhury, M ., Apon, A . and Dey K ,., Eds. Data analytics for intelligent transportation system, Elsevier, 2017, pp. 283-307. DOI: 10.1016/ B978-0-12-809715-1.00012-2 [2]: Kotsiantis SB, Zaharakis I, Pintelas P. Supervised machine learning: A review of classification techniques. ed, 2007

[16]: Abduljabbar, R., Dia, H., Liyanage, S. and Bagloee, S., Applications of artificial intelligence in transport: an overview. Sustainability, 11(1), pp. 189-190, 2019. DOI: 10.3390/su11010189

[17]: Phyu TN. Survey of classification techniques in data mining, in Proceedings of the International MultiConference of Engineers and Computer Scientists 2009; pp. 18-20

[18]: Kotsiantis SB, Zaharakis I, Pintelas P. Supervised machine learning: A review of classification techniques. ed, 2007

[19]: Vapnik VN. The Nature of Statistical Learning Theory, 1995.

[20]: Alexandre Kowalczyk . Support Vector Machine Succinctly p27, p75 .

[21]: Gowrishankar S. Veena A. Introduction to Python Programming.

[22]: https://colab.research.google.com/#scrollTo=5fCEDCU_qrC0

[23]: Wes McKinney.(2013). Python for Data Analysis

[24]: https://www.tutorialspoint.com/python_pillow/python_pillow_tutorial.pdf