|  |  |
| --- | --- |
| Internship Project Title | TCS ion RIO-125: Image Classification - using Different Machine Learning Algorithms |
| Name of the Company | TCS iON |
| Name of the Industry Mentor | Imran Khan |
| Name of the Institute | Mepco Schlenk Engineering College , Sivakasi. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Start Date | End Date | Total Effort (hrs.) | Project Environment | Tools used |
| 07/09/2021 | 07/12/2021 | 130 | Google Colab | Python 3 |

**TABLE OF CONTENT**

* Acknowledgments
* Objective
* Introduction / Description of Internship
* Internship Activities
* Approach / Methodology
* Assumptions
* Exceptions / Exclusions
* Charts, Table, Diagrams
* Algorithms
* Challenges & Opportunities
* Risk Vs Reward
* Reflections on the Internship
* Recommendations
* Outcome / Conclusion
* Enhancement Scope
* Link to code and executable file

**ACKNOWLEDGMENTS**

I am conveying my sincere gratitude towards my industry mentor, Mr.Imran Khan, and academic mentor, Dr. Muthukumar for helping me throughout this project till now and providing me this wonderful platform to complete this project. I am also thankful for answering my queries at every phase of the project. I also want to thank all my friends who helped me with valuable suggestions during this project.

**OBJECTIVE**

The objective of this model is to make a image classification using machine learning algorithms .The model should be able to classify the images of the fruits by inputting the images.

**INTRODUCTION**

The dataset that I have selected for the project and training of the model contains the dataset of 90380 images with 131 different fruits. Throughout my entire process of the industry project, I overcame the milestones that have been defined by TCS iON. During the period of completing the first milestone, I cleaned and sanitized the dataset and had done some EDA. The next milestone was on training a model for classifying the images. For this, I went through some articles and videos to understand some of the concepts on model training and also for understanding various classification techniques. Then I developed some models for the second milestone. The models were on Logistic regression and Random forest. By the end of the second milestone, I tuned the logistic regression and predicted salary using a tuple of user-defined input. I have generated classification reports for the models I have used. At the end of the project, I have used Support Vector Machine(SVM) as another classification model and tuned it also.

The model that I have developed classifies the image and helps in creating a dashboard. the main objective of this model is to classify the images by different aspects.

**INTERNSHIP ACTIVITIES**

* Completed the RIO – pre-assessment test.
* Gone through the project reference materials that have been available such as the welcome kit, day-wise plan, project reference material, etc.
* Watched the webinars and recorded lectures.
* Created a dataset that is suitable for this project.
* Cleaned and sanitized the dataset.
* Gone through many articles and videos to learn about classification models and training techniques.
* Trained the dataset to predict the salary for an HR by providing details of that particular person.
* Compared 3 different classification techniques i.e. SVM, logistic regression, random forest by using the scores and classification reports.
* Wrote activity reports and project interim reports.

**APPROACH / METHODOLOGY**

The approach I took for the internship project for completing the milestones is understanding the concepts of the requirements. Reading articles and watching videos helped in achieving knowledge about the requirements. Google Colab has been used for doing the programming. Jupyter Notebook has also been used for much faster execution. A GitHub account has been created for publishing the codes.

There are several methods to make a binary classification, which include SVM, logistic regression, random forest, etc. So I have trained and tested my data using Logistic Regression, Random Forest, and SVM, and compared them.

**ASSUMPTIONS**

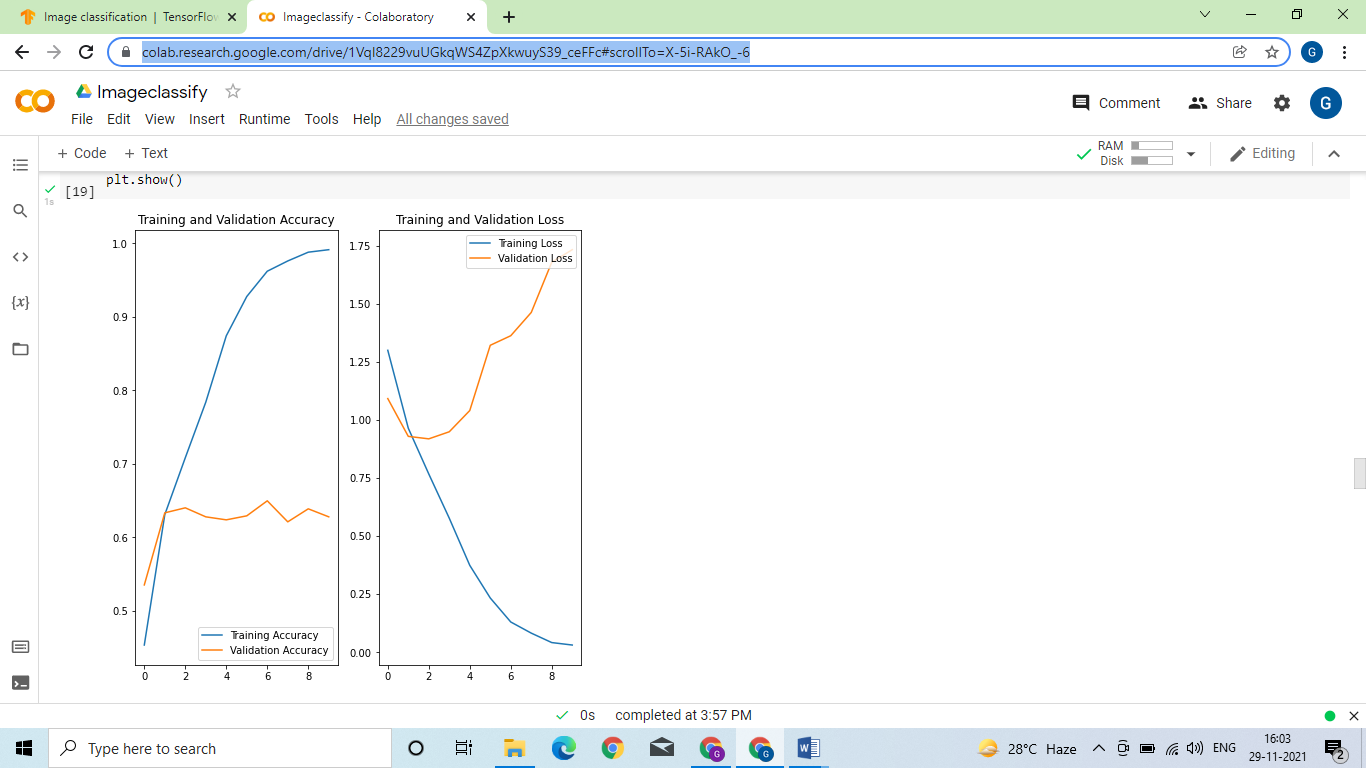
There was some errors or misvalues in the dataset. i.e. some troubles that incurred in the cleaning process. So I have taken a more number of epochs for training the model

**EXCEPTIONS**

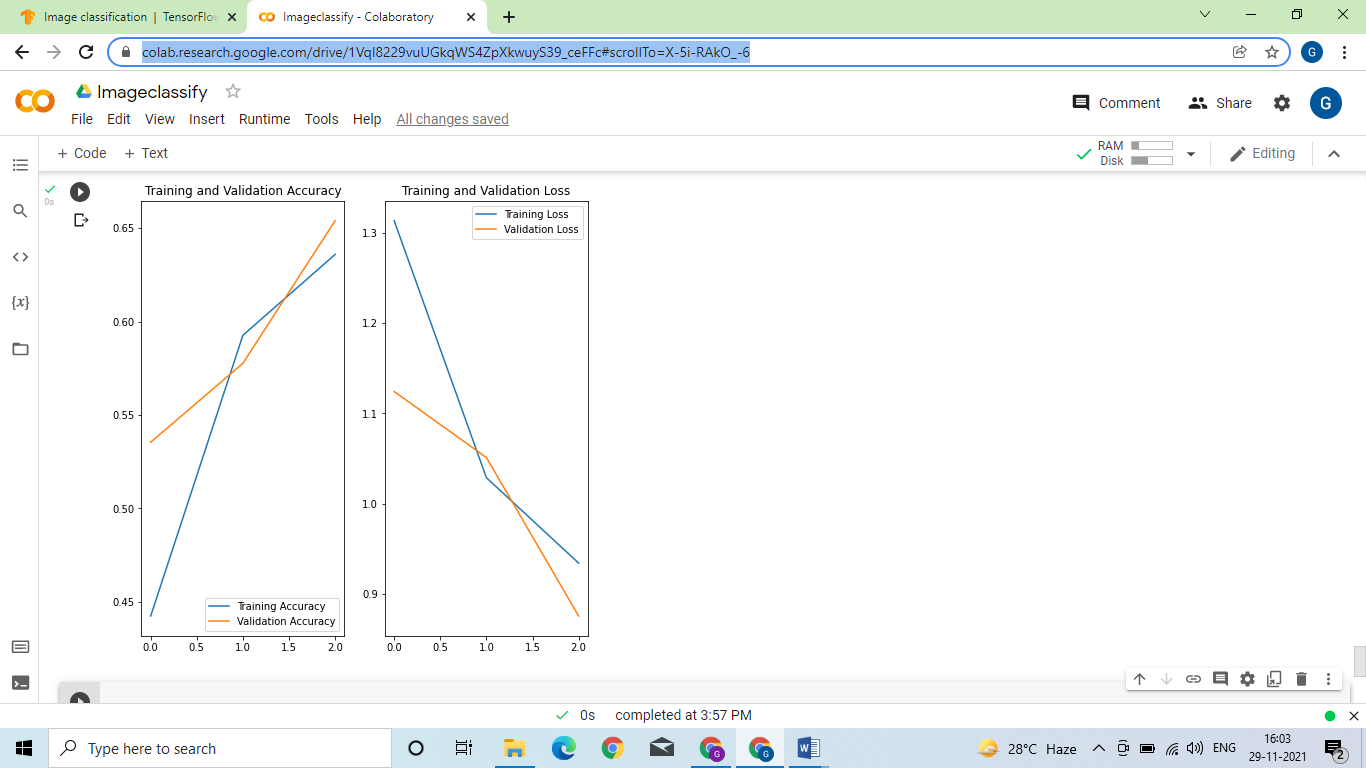
The exception incurred during the classification is that this project cannot classify with 100 percent accuracy , but it surely classifies the image with atmost 90 percent accuracies.

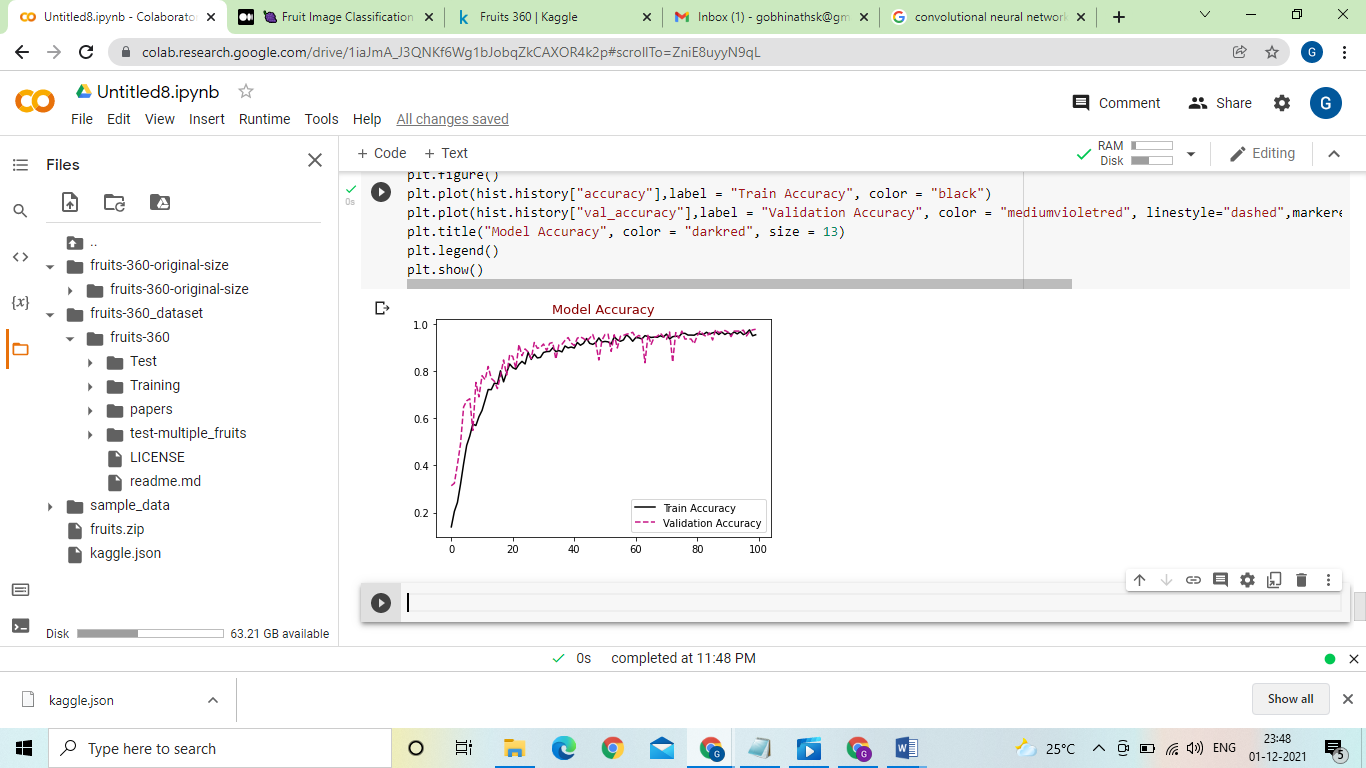
**CHARTS, TABLES, DIAGRAMS**

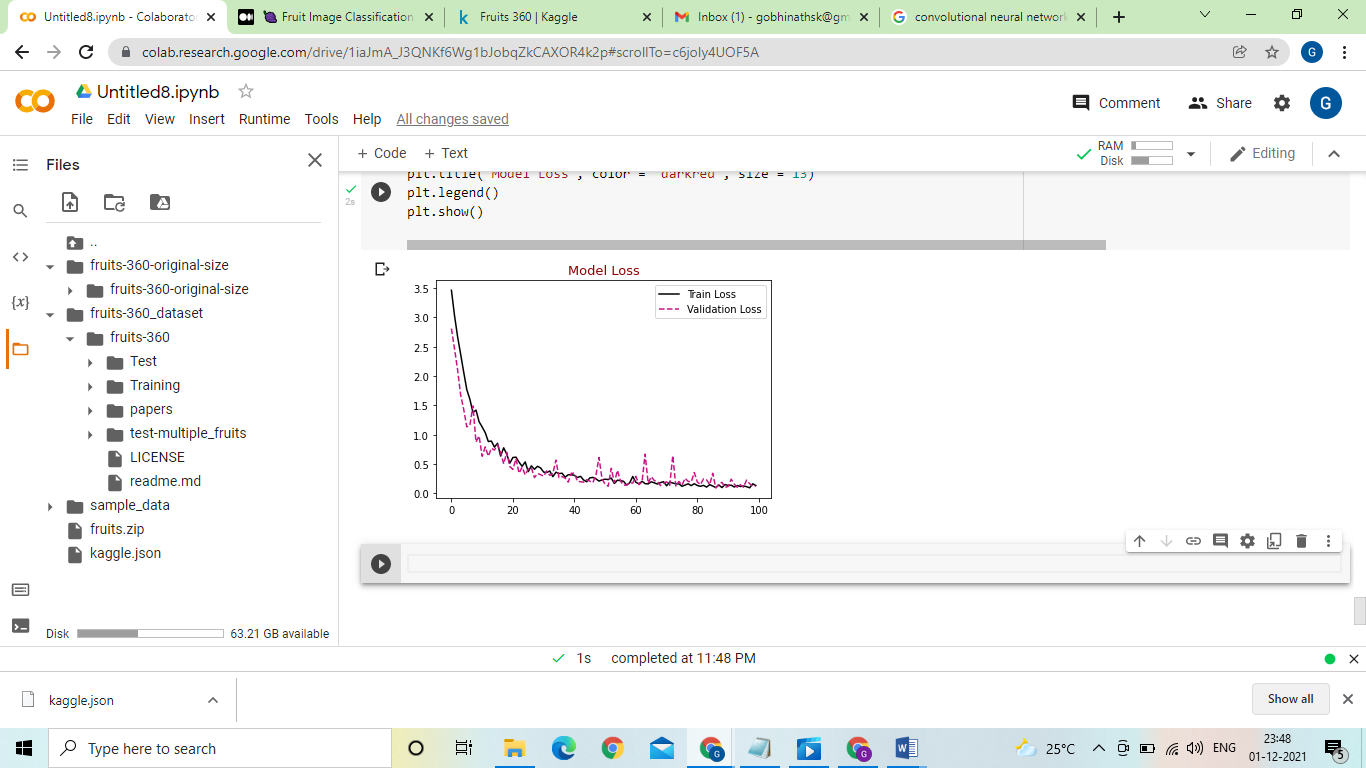
The following are the charts and diagrams that I have created as part of the EDA and Visualization.



This is the graph plotted by taking training accuracy values and validation accuracy values as the x axis and y axis .







**ALGORITHMS**

The Algorithm used in the development of all the classifiers is as follows:

1. Start
2. Importing necessary libraries.
3. Vectorize character values.
4. Normalizing all the values.
5. Splitting the dataset to training and testing data.
6. Hyper-parameter tuning with the corresponding classification method.
7. Training the model.
8. Testing the model using test data.
9. End

**CHALLENGES & OPPORTUNITIES**

During this industry project, the challenges that I had faced was on visualizing each attribute’s relationships and developing a model for the classification. I only had a rough idea of these classification techniques from my academic background. But these daily activities made me much more knowledgeable on these techniques. So I had the opportunities to build a good and conceptual knowledge of these models. Also, I learned about hyper-parameter tuning which I wasn’t aware of it earlier.

**RISK Vs REWARD**

In order, to predict a class according to the attributes that I had was not enough. This is because no previous work experience or any project experience was not available. Therefore I had to use these available attributes to do the classification process.

**REFLECTIONS ON THE INTERNSHIP**

It was my first internship that I had done in my academic career. So everything was new to me. the digital discussion room helped in connecting various people who are from different backgrounds and cultures. This helped me to develop a systematic approach to doing the project. The activity reports and interim reports helped me to analyze my process and doings. This helped in refurbishing some of the concepts throughout the project.

**RECOMMENDATIONS**

I felt that much more resources regarding the project can be made available in the project reference part. the reference materials were a bit small. I had to do some research out of the given project references.

**OUTCOME / CONCLUSION**

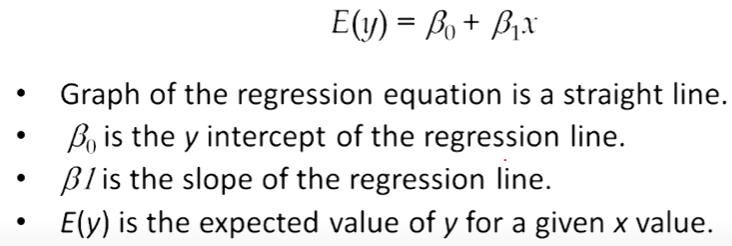
By looking back from the beginning stage of the project, I have cleaned and sanitized the data, i.e data has been preprocessed. A logistic regression model has been trained and tested at the end of the milestone. A random forest classifier has also been implemented to understand the difference between certain models. Classification reports have been generated for both models. The parameters of the logistic regression model have been tuned for showing better performance. Also for the tuned model, a classification report has been generated. After the second milestone, I started working on the support vector machine part and trained it for the classification part.

I was able to enhance my programming skills and even developed some methods to clean and identify errors from a dataset. I came to know more about various visualization libraries, classification models, and how to implement these classification models using python.

From these 30 days of learning and implementing, I have grasped much knowledge about the following:

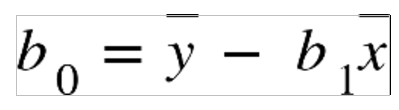
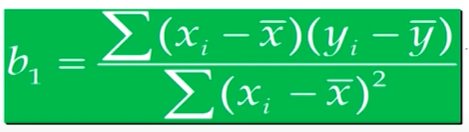
**Linear Regression**

Linear regression is perhaps one of the most well-understood and well-known algorithms in statistics and machine learning. Linear regression was borrowed from statistics to machine learning, which makes a statistical algorithm and machine learning algorithm. It is a linear model that assumes a linear relationship between the input variables (x) and the single output variable (y). The equation that describes how y is related to x and an error term is called the regression model. The simple linear regression model is



Linear regression helps in finding the best line of fit through the data by searching for the regression coefficient (B1) that helps in minimizing the total error of the model.

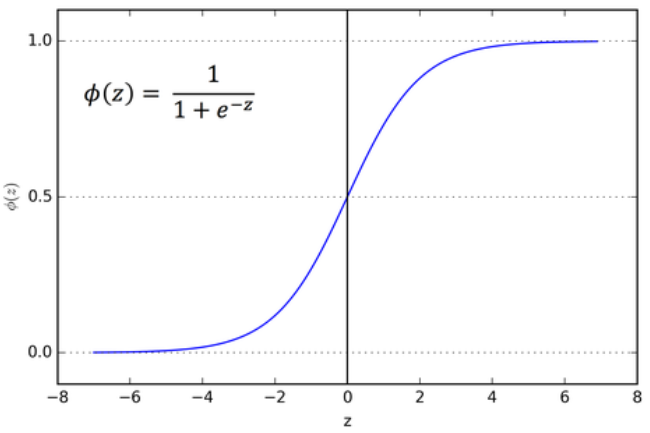
Now we can find or estimate the values for the parameters beta1 and beta2 using the equation for calculating the error. Then we define a model by minimizing the residual error.

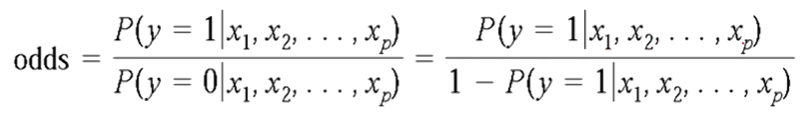
The gradient descent technique is an optimization algorithm that helps in finding the values of parameters of a function to minimize the cost function.

**Logistic regression**

This type of regression is used when the dependent variable is categorical. The function used here is a sigmoid function.



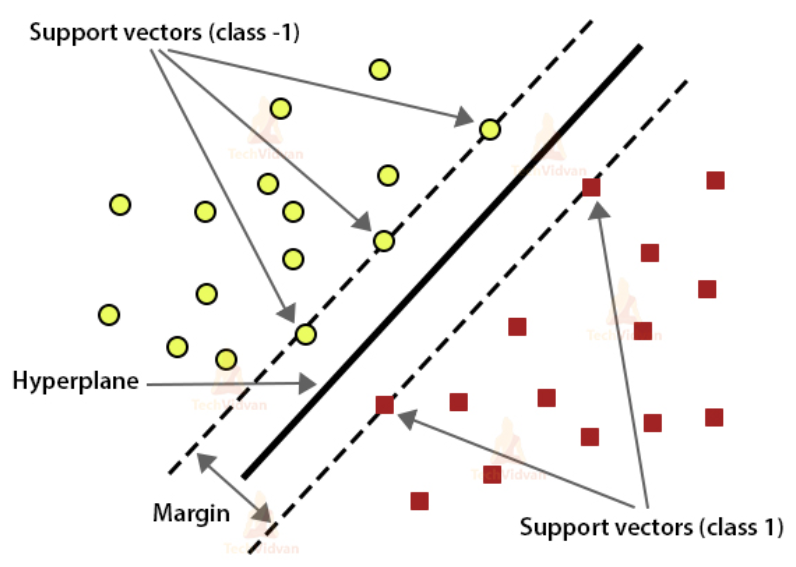
If ‘z’ tends to infinity, y tends to 1 and if ‘z’ tends to negative infinity, y tends to 0. So the outputs for a logistic regression will be 0 and 1. By using the function mentioned in the figure, we can define an odds ratio as

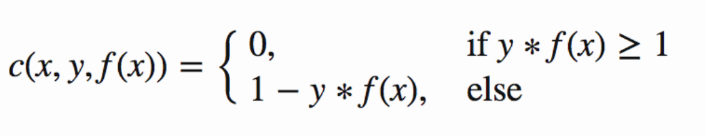


**Support Vector Machine**

In machine learning, support vector machines (SVMs) are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis.

To separate the two classes of data points, there are many possible hyperplanes that could be chosen. Our objective is to find a plane that has the maximum margin, i.e the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence.



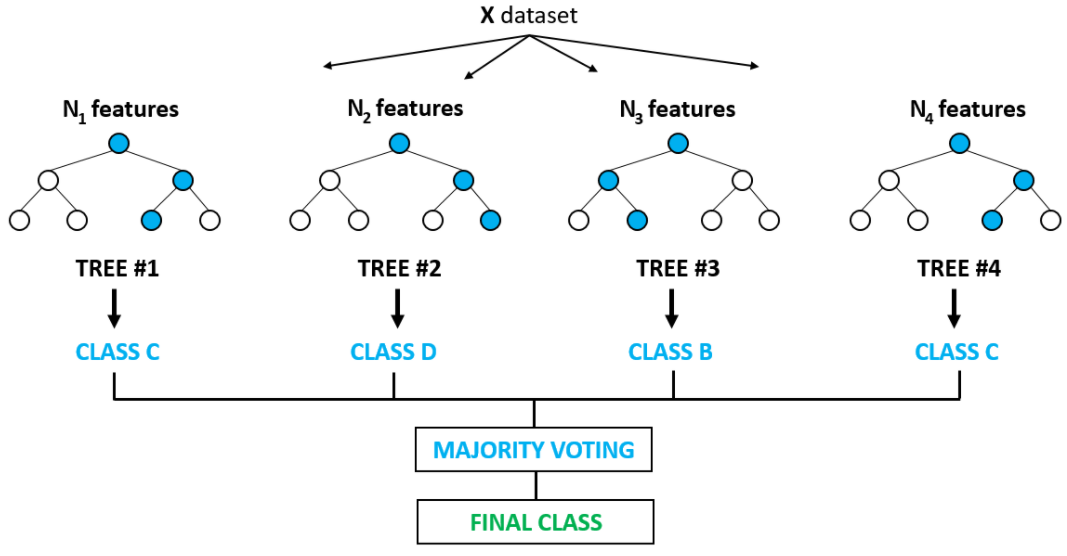
In SVM, the best hyperplane is the one that maximizes the margins from both tags. The loss function that helps maximize the margin is hinge loss.

Now that we have the loss function, we take partial derivatives with respect to the weights to find the gradients. Using the gradients, we can update our weights.

**Random Forest**

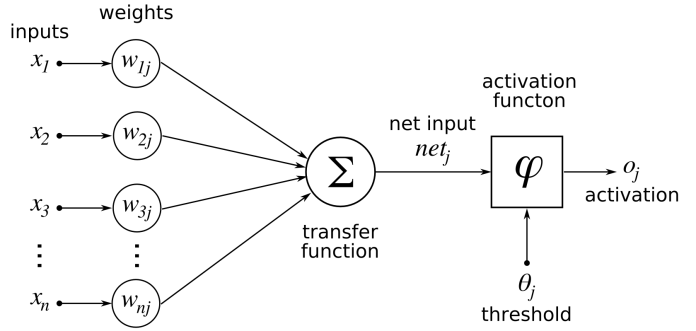
A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. It is a supervised learning algorithm. The "forest" it builds, is an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result.

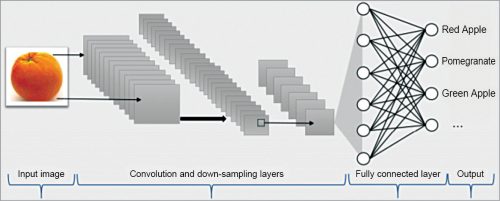
Random forest adds additional randomness to the model while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model.



**Convolutional Neural Network**

convolutional neural networks are composed of multiple layers of artificial neurons. Artificial neurons, a rough imitation of their biological counterparts, are mathematical functions that calculate the weighted sum of multiple inputs and outputs an activation value.





**Project Development**

Now at the end phase of the project, I have cleaned, sanitized, visualized, done some EDA, and preprocessed the dataset enabling it for training. I have also trained and tested the model using logistic regression, SVM, Random Forest. Then I generated a classification report for each of the classifiers. Then I have tuned the parameters of the models, SVM and Logistic Regression, and chose the best ones. And again, I have tested and printed the classification report. Hence it was evident that the SVM model had more accuracy than the other 2. So I chose the SVM classifier over others. In the end, I have tested the model against a user-defined data tuple also.

**Project Inference**

After hyper-parameter tuning and cross-validation of Logistic regression and SVM, the classification reports .

**ENHANCEMENT SCOPE**

This industry project has a wide scope. Using the different training and testing model , we can easily classify the images.

**LINK TO CODE AND EXECUTABLE FILE**

Link to the colab file:

<https://colab.research.google.com/drive/1iaJmA_J3QNKf6Wg1bJobqZkCAXOR4k2p#scrollTo=Trtx5rIQOix_>