|  |  |
| --- | --- |
| 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 |
| 08.09.2021 | 07.12.2021 | | 20 | | Google colab | Python3 |
| Milestone # | 2 | Milestone: | | Image Classifier using CNN | | |

**TABLE OF CONTENT**

* Acknowledgements
* Objective
* Introduction
* Internship Activities
* Approach / Methodology
* Outcome/Conclusion
* Link to code and executable file

**ACKNOWLEDGEMENTS**

I am conveying my sincere gratitude towards my industry mentor Mr.Imran Khan, and academic mentor, Dr. S. Muthu Kumar 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 classifier using Machine learning Algorithms.

**INTRODUCTION**

the next 10 days of my project, I visualized all the features or attributes of image classifictions. After all the visualization part, I trained the dataset using logistic regression and random forest. Classification reports for each classifier were generated. At the end using a tuple of user defined data, a small prediction was made using the logistic regression classifier.

**INTERNSHIP ACTIVITIES**

* + Watched the welcome kit videos.
  + Done preparations for RIO – pre-assessment.
  + Attended the RIO – pre-assessment test.
  + Went through the day-wise plan.
  + Read the project reference material.
  + Read the industry project material.
  + Watched webinar 1.
  + Watched webinar 2.
  + Gone through all posts in the digital discussion room.
  + I went through the linear regression YouTube videos.
  + Read the linear regression article.
  + Watched the lectures provided and other videos for further understanding.
  + Created a GitHub account.
  + Searched and found out a proper data set for this project.
  + Wrote activity reports.
  + Checked and clarified the data set whether it has enough data for the project.
  + Read articles and find out how to clean and sanitize the data.
  + Cleaned the data set.
  + Sanitized the data set.
  + Done Exploratory Data Analysis(EDA)
  + Watched videos on model training
  + Used Logistic Regression and trained it
  + Used Random Forest Classifier and trained it.
  + Generated Classification reports for both classifiers.
  + Did Hyperparameter tuning for the logistic regression and generated the classification report after the tuning process.
  + Did a prediction using the logistic regression classifier.
  + Watched the following videos for understanding about model training and hyperparameter tuning.
    - <https://www.youtube.com/watch?v=D7p5g8PMDvQ&list=PLLy_2iUCG87CNafffzNZPVa9rW-QmOmEv&index=41>
    - <https://www.youtube.com/watch?v=HdlDYng8g9s&list=PLeo1K3hjS3uvCeTYTeyfe0-rN5r8zn9rw&index=17>
  + Gone through the following articles
    - <https://www.pluralsight.com/guides/cleaning-data-using-python>
    - <https://www.analyticsvidhya.com/blog/2020/08/types-of-categorical-data-encoding/>
    - <https://www.geeksforgeeks.org/hyperparameter-tuning/>
    - <https://scikit-learn.org/stable/modules/grid_search.html>

**APPROACH / METHODOLOGY**

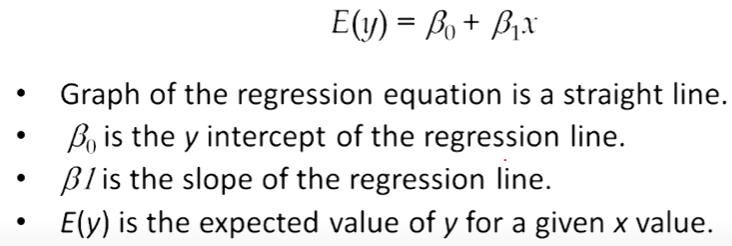
The approach I took for the internship project for completing the 2nd milestone is understanding the concepts of the requirements. Reading articles and watching videos helped in achieving knowledge about the requirements. Google colab has been used for doing the programming.Jupyter Notebook has also been used for much faster execution. A GitHub account has been created for publishing the codes.

**OUTCOME**

From these 15 days of learning and implementing, I have grasped much knowledge about the following:

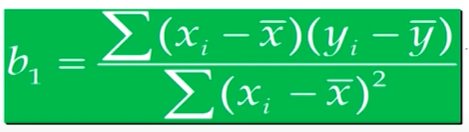
**Linear Regression**

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.

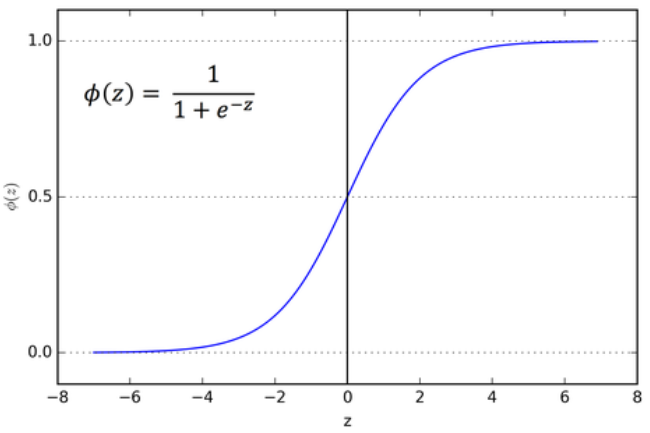
Now we can find or estimate the values for the parameters beta1 and beta2 using the equation for calculating the error. Then we define a model by minimizing the residual error.

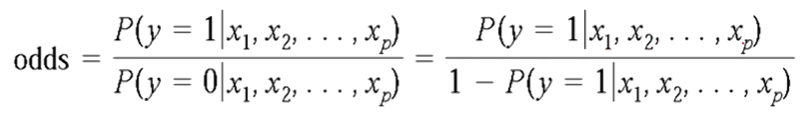
Gradient descent technique is an optimization algorithm that helps in finding the values of parameters of a function to minimize the cost function.

**Logistic regression**

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



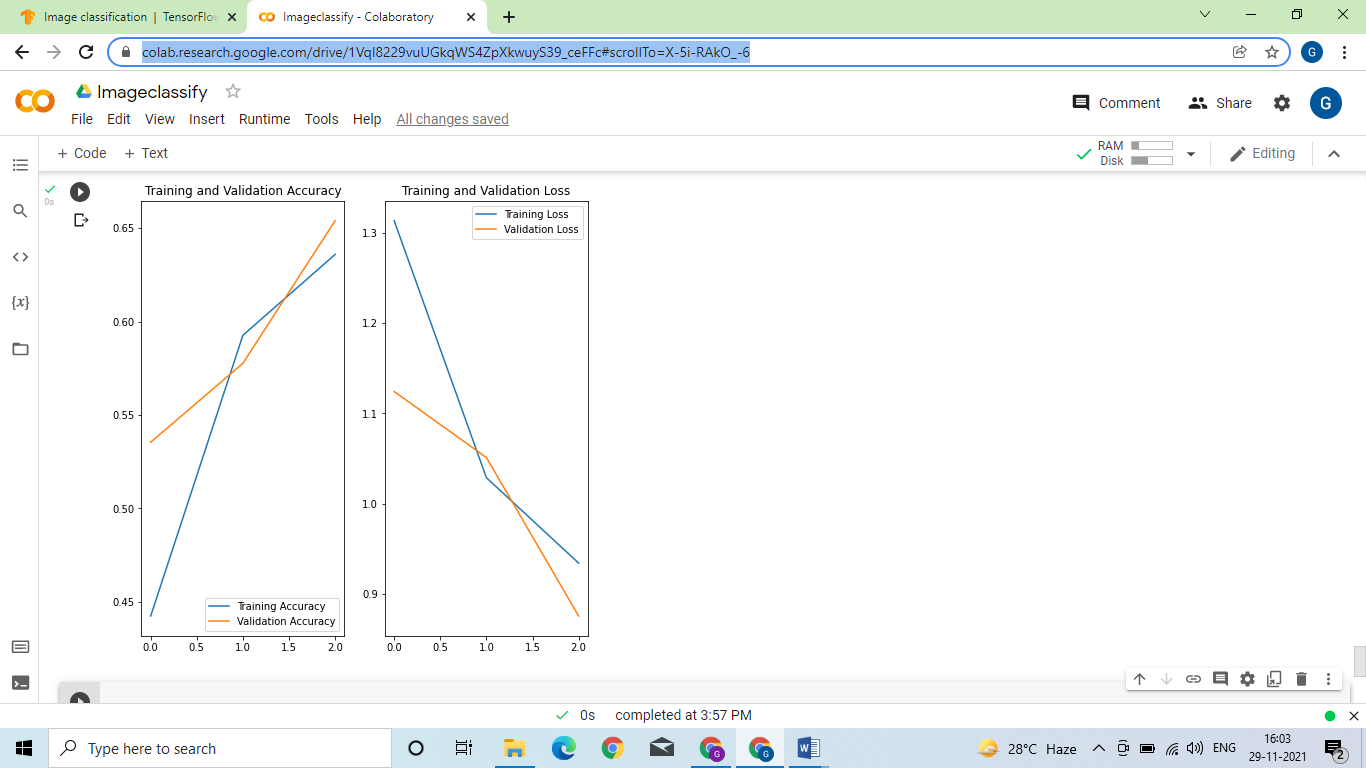
If ‘z’ tends to infinity, y tends to 1 and if ‘z’ tends to negative infinity, y tends to 0. So the outputs for a logistic regression will be 0 and 1. By using the function mentioned in the figure, we can define an odds ratio as

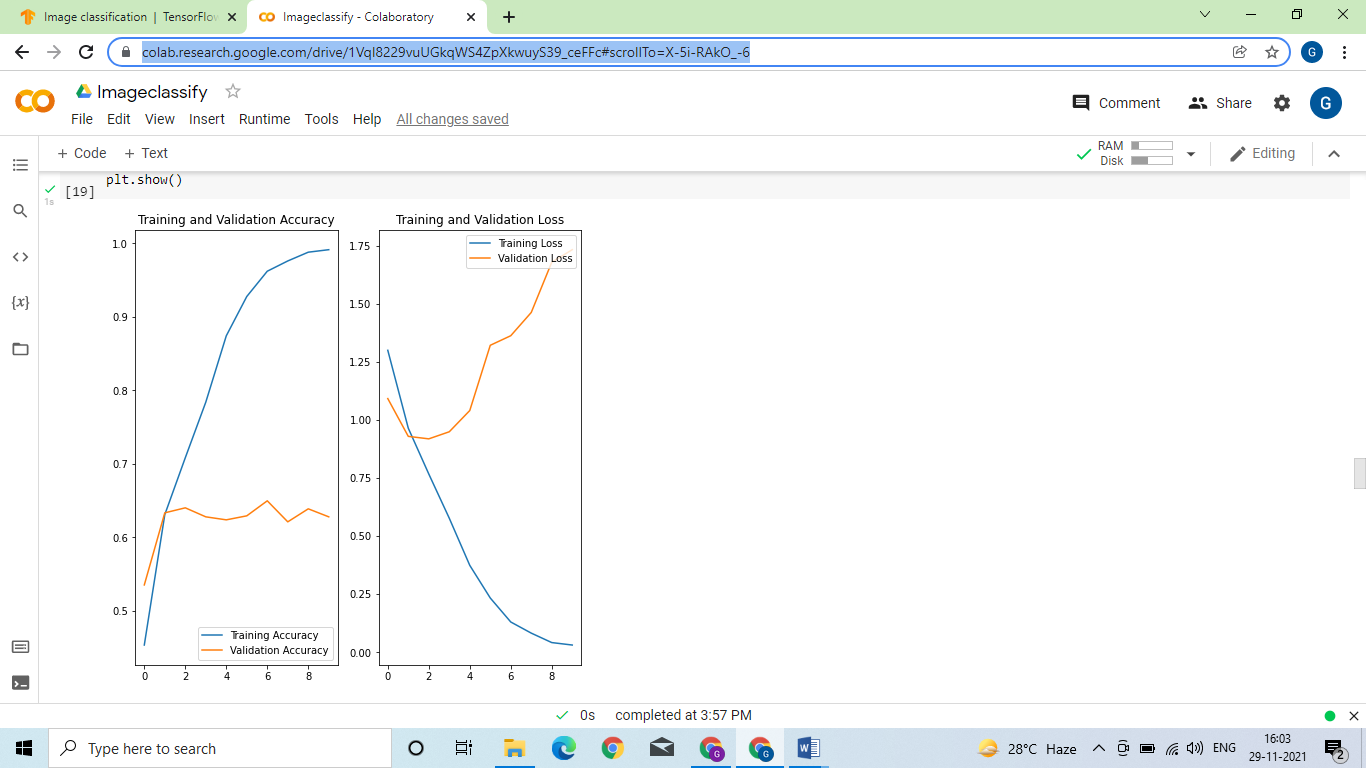


**PROJECT DEVELOPMENT**

The current stage of the project is that 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 the 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.

The plots show that training accuracy and validation accuracy are off by large margins, and the model has achieved only around 60% accuracy on the validation set.





**LINK TO CODE AND EXECUTABLE FILE**

* Link to the code:

<https://colab.research.google.com/drive/1VqI8229vuUGkqWS4ZpXkwuyS39_ceFFc#scrollTo=X-5i-RAkO_-6>