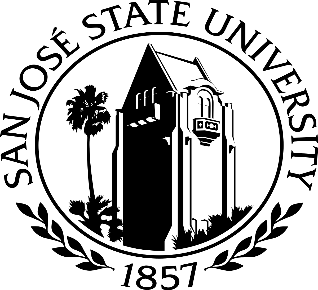
Project Report  
Milestone 1

CMPE 257: Machine Learning  
San Jose State University  
Master of Science in Data Analytics



**Group 8**

Aarathi Nadathur – 013704405  
Amar Ippili- 013728975

Divya Puraswani - 013755391

Sai Chaitanya Tolem - 013008788

Sameer Rajput - 013802841

Shubh Johri - 013750672

Selected Machine Learning Algorithm: Convolutional Neural Networks

Link: Google Colaboratory: <https://colab.research.google.com/drive/1RsvUuIq2oEIHBJ0gHIYIuvorLvGnlHSG>

**Who Did What!**

|  |  |  |
| --- | --- | --- |
| **Milestone 1** | | |
| **What** | **Description** | **Who** |
| Technical | 1. Data Preparation and EDA | Sai Chaitanya |
| 1. Data Validation, Data Preprocessing | Sai Chaitanya, Sameer Rajput |
| 1. Logistic Model | Amar Ippili |
| 1. Random Forest, KNN | Shubh, Aarathi |
| 1. Decision Tress, Naïve Bayes | Shubh, Aarathi |
| 1. SVM, CNN | Divya, Sameer, Chaitanya |
| Presentation | 1. PowerPoint Preparation | Divya, Amar |
| 1. In Class PowerPoint Presentation | Entire Team |
| Documentation | 1. Report Preparation | Sameer, Divya |
| 1. Report Editing | Shubh, Aarathi |

Contents

[1. About the dataset 4](#_Toc7382667)

[2. Problem 4](#_Toc7382668)

[3. Software and libraries 4](#_Toc7382669)

[4. Data Preparation 4](#_Toc7382670)

[5. Data Validation 5](#_Toc7382671)

[6. Data Pre-processing 6](#_Toc7382672)

[7. Machine Learning Algorithms 6](#_Toc7382673)

[7.1 Logistic Regression 6](#_Toc7382674)

[7.2 KNN: 6](#_Toc7382675)

[7.3 Naïve Bayes: 7](#_Toc7382676)

[7.4 Random Forests: 8](#_Toc7382677)

[7.5 Decision Tree Classifier: 8](#_Toc7382678)

[7.6 SVM: 9](#_Toc7382679)

[8. Selected Machine Learning Algorithm: CNN 10](#_Toc7382680)

[8.1 What is CNN? 10](#_Toc7382681)

[8.2 Why is CNN better than other algorithms? 10](#_Toc7382682)

[9. Conclusion 13](#_Toc7382683)

[10. Bibliography 13](#_Toc7382684)

# About the dataset

The CIFAR-100 dataset has 100 classes containing 600 images each. There are 500 training images and 100 testing images per class. The 100 classes in the CIFAR-100 are grouped into 20 super classes. Each image comes with a "fine" label (the class to which it belongs) and a "coarse" label (the superclass to which it belongs) (CIFAR 100 DATASET).

|  |  |
| --- | --- |
| **Super Class** | **Class** |
| Vehicle 1 | Bicycle, Bus, Motorcycle, Pickup truck, Train |
| Vehicle 2 | Rocket, Tank, Tractor, Streetcar, Lawn-mower |

# Problem

Our goal is to perform a binary classification of images in vehicle 1 and vehicle 2 superclass pairs.

# Software and libraries

**Software: Google Colaboratory Environment** (adventuresinmachinelearning, n.d.)

Google colaboratory environment is based on Python Jupyter notebooks which gives the user free access to Tesla K80 GPUs. It provides the ability to experiment with deep learning. To access the environment, you must have a signed in Google Drive account. The .ipynb files that you create will be saved in your Google Drive account. We have signed in using SJSU Email Id to access the environment and work with cifar100 dataset.

Python Libraries: Keras, Sklearn, Numpy, Matplotlib, Pandas, Scipy, Tensorflow

Keras is a high-level neural networks API, written in Python and we are running it on top of TensorFlow, CNTK, or Theano. It supported both convolutional networks and recurrent networks, as well as combinations of the two and ran seamlessly on CPU and GPU. (Keras Documentation, n.d.)

# Data Preparation

Cifar 100 dataset consists of train, test and meta data. We loaded them into colaboratory and combined the images(X), coarse labels(Y) (0-19) data into data frame and filtered this data frame on target variable (vehicle1 =18, vehicle2 = 19) as in Figure 1.

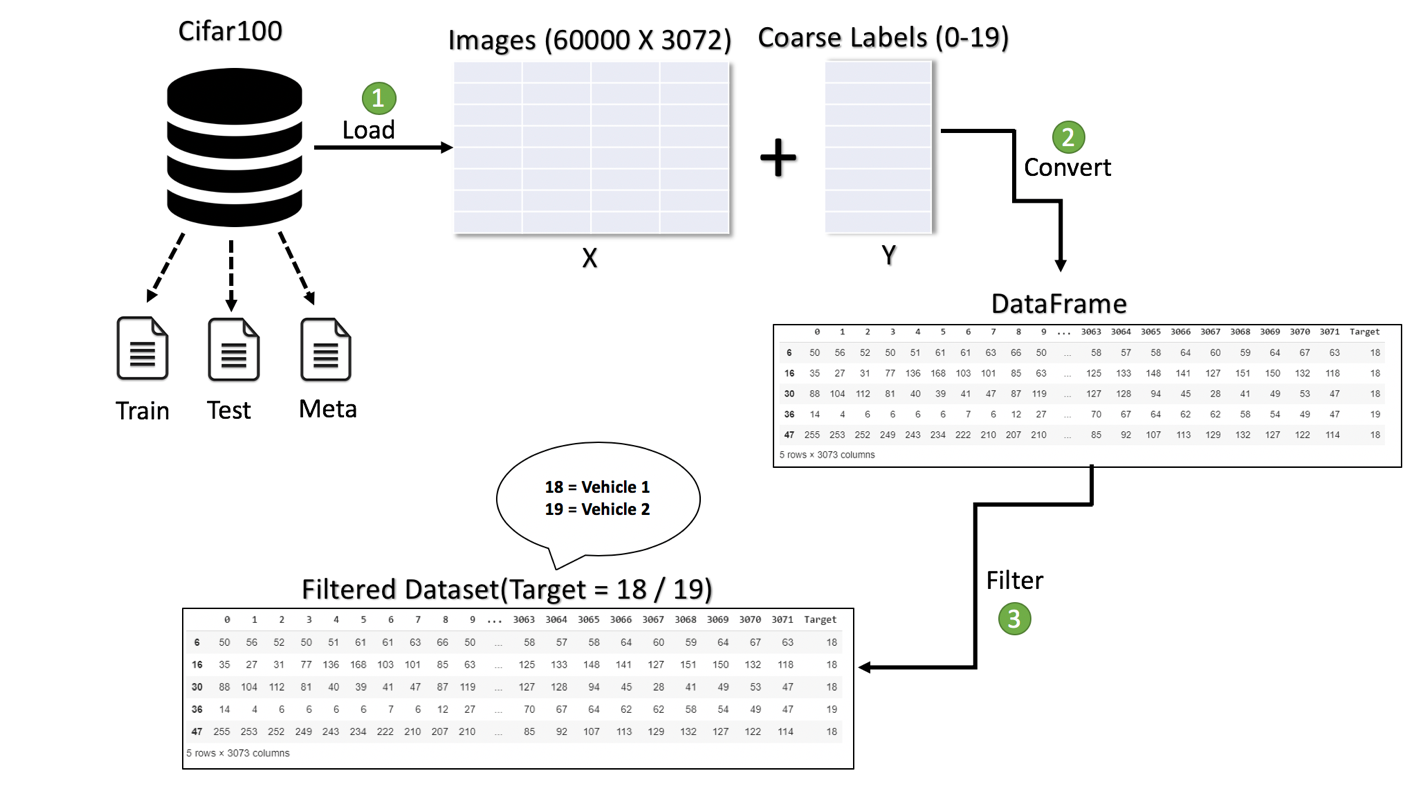


Figure 1

# Data Validation

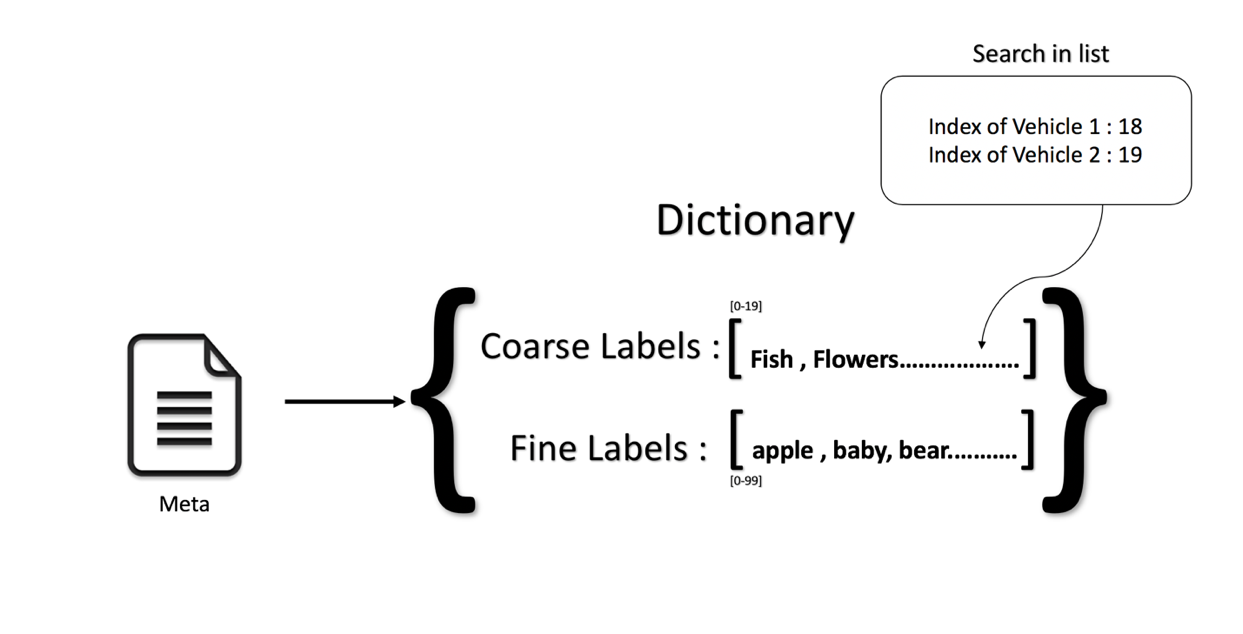


Figure 2

We used the meta data file in cifar 100 which consists of a serialized dictionary object to search for index of vehicle 1, vehicle 2 in the coarse label list as in Figure 2 and validated that 18 corresponds to vehicle 1 and 19 corresponds to vehicle 2 through code and we have crosschecked it by plotting few images from the filtered dataset.

# Data Pre-processing

To split the dataset into two sets, one for training and the other for testing, we have split the dataset in an 80–20 ratio such that 80% data is trained.

To reshape our dataset inputs (X\_train and X\_test) to the shape that our model expects when we train the model, we used astype to cast a pandas object into float. And then we used ‘one-hot-encode’ on our target variable using get\_dummies to convert categorical variable into dummy variables and to\_categorical to convert a class vector (integers) to binary class matrix. This means that a column will be created for each output category and a binary variable is inputted for each category.

# Machine Learning Algorithms

## 7.1 Logistic Regression

**Accuracy Achieved**: 60.08%

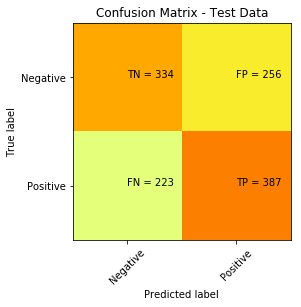
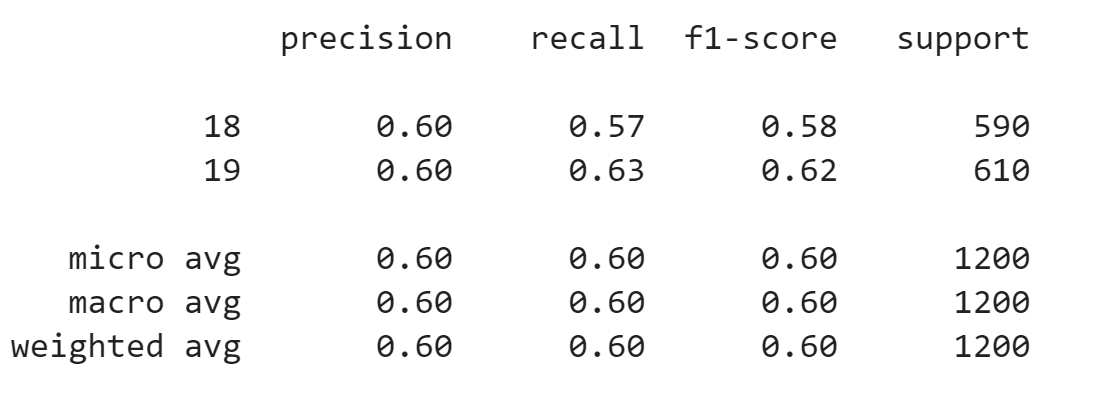


Figure 3



Figure

7.2 KNN:

**Accuracy Achieved**: 61.66%

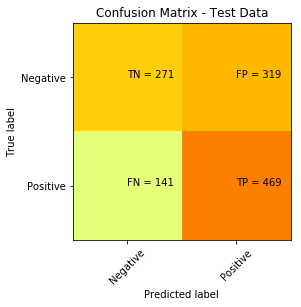


Figure 5

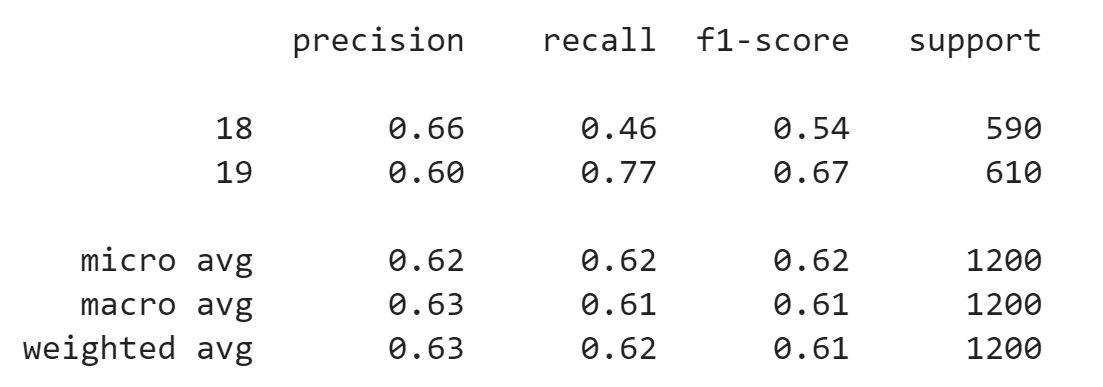


Figure 6

7.3 Naïve Bayes:

**Accuracy achieved:** 55.83%



|  |  |
| --- | --- |
| Figure 7    Figure 8 |  |
|  |  |

7.4 Random Forests:

**Accuracy achieved:** 69.83%

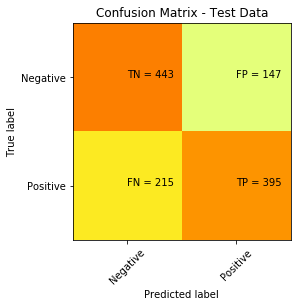


Figure 9

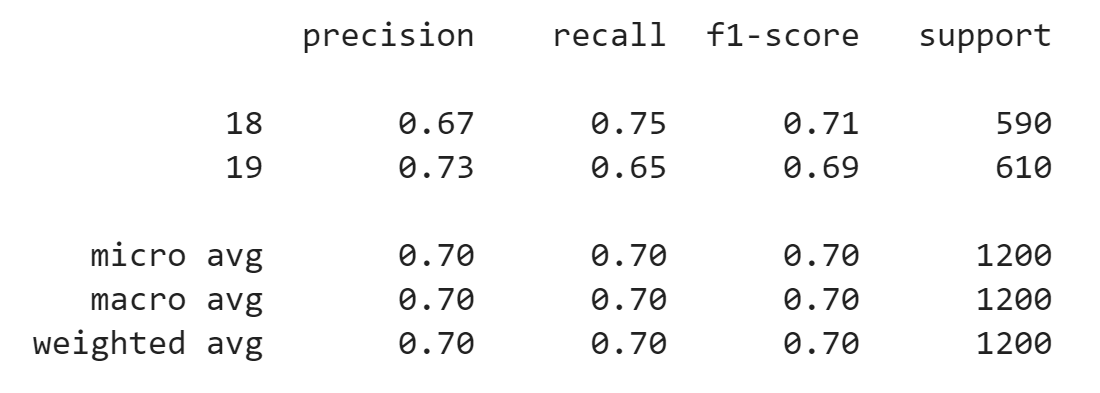


Figure 10

7.5 Decision Tree Classifier:

**Accuracy achieved:** 58.16%

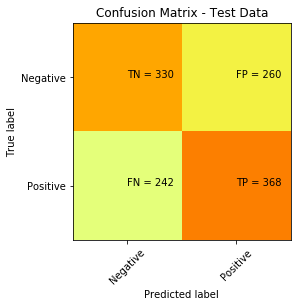


Figure 11

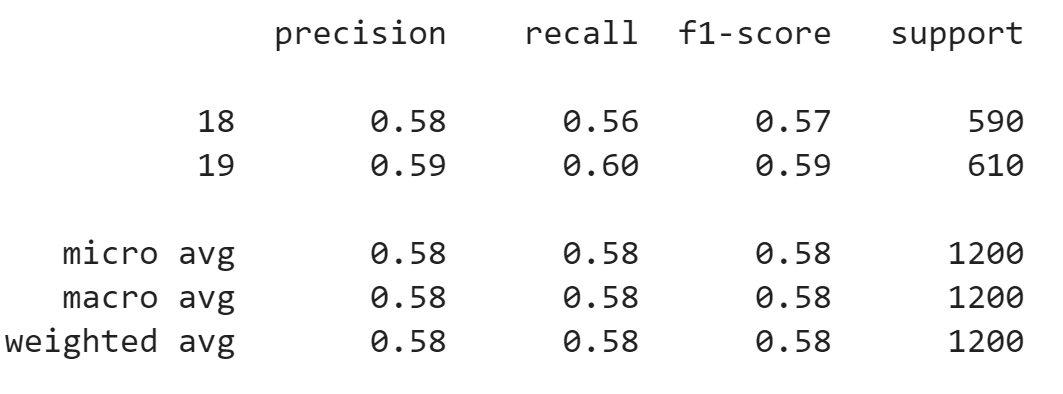


Figure 12

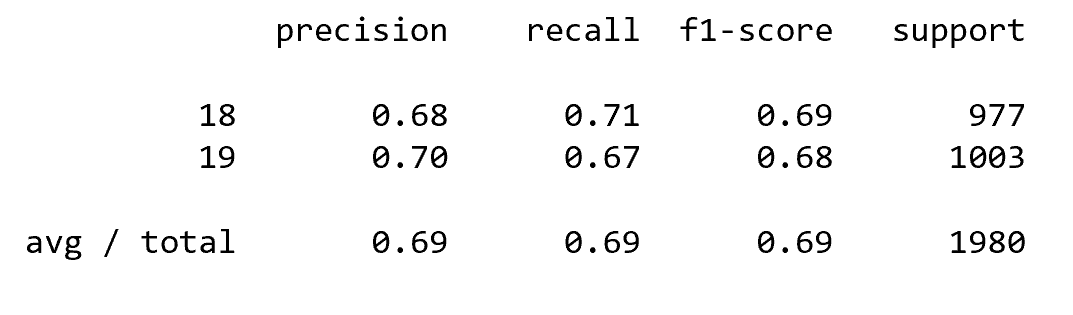
## 7.6 SVM:

**Accuracy achieved:** 68.88%

Confusion Matrix – Test Data

|  |  |  |
| --- | --- | --- |
| **Predicated Label** | **Negative** | **Positive** |
| **True Label** |
| **Negative** | TN = 696 | FP = 281 |
| **Positive** | FN = 335 | TP = 668 |

*Figure 13*



*Figure 14*

# Selected Machine Learning Algorithm: CNN

## 8.1 What is CNN?

A Convolutional Neural Network (ConvNet/CNN) is a Deep Learning algorithm which can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other. While in primitive methods filters are hand-engineered, with enough training, ConvNets have the ability to learn these filters/characteristics. (Towards DataScience, n.d.)

Our CNN model runs with 200 epochs and batch size of 128 with steps per epoch as 18 having MaxPool and using dropout and image augmentation we tried to prevent overfitting of our model with a learning rate which is neither too high nor too low. (Towards DataScience, n.d.)

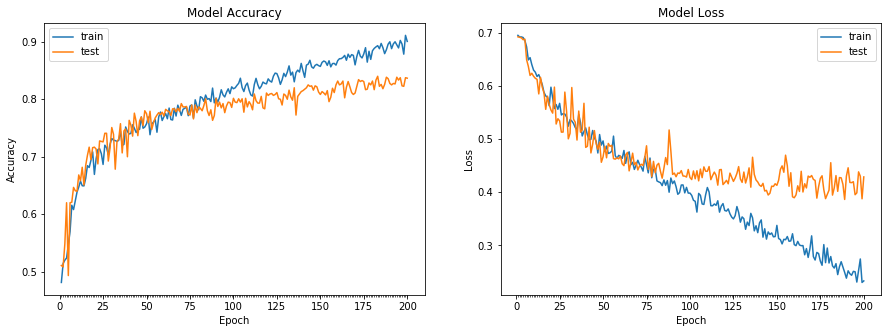


*Figure 15*

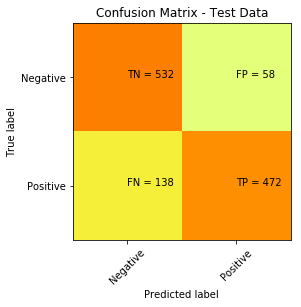
## 8.2 Why is CNN better than other algorithms?

The goal of CNN is to reduce the images into a form which is easier to process without losing features which are critical for getting a good prediction. This means that useful attributes from an already trained CNN can be extracted with its trained weights by feeding data on each level and tune the CNN a bit for the specific task.

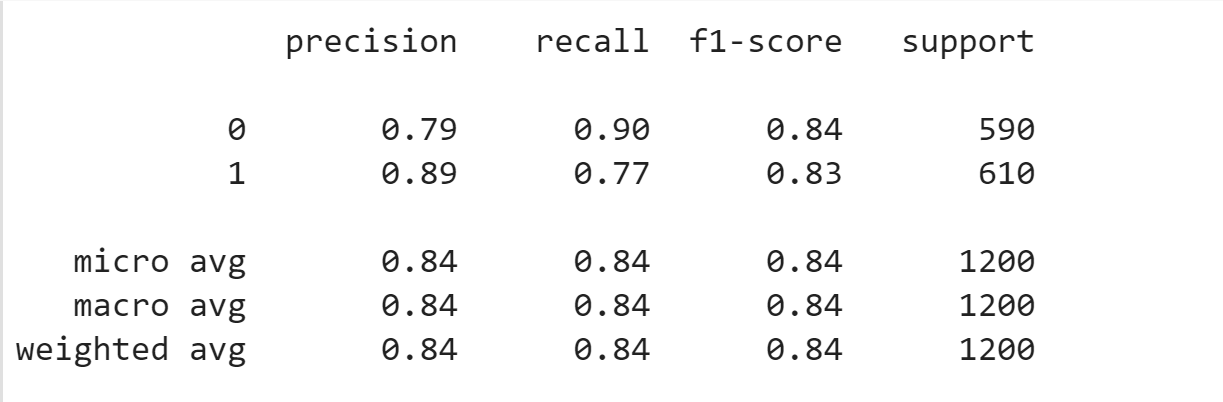
**Accuracy achieved**: 83.67%



*Figure 16*

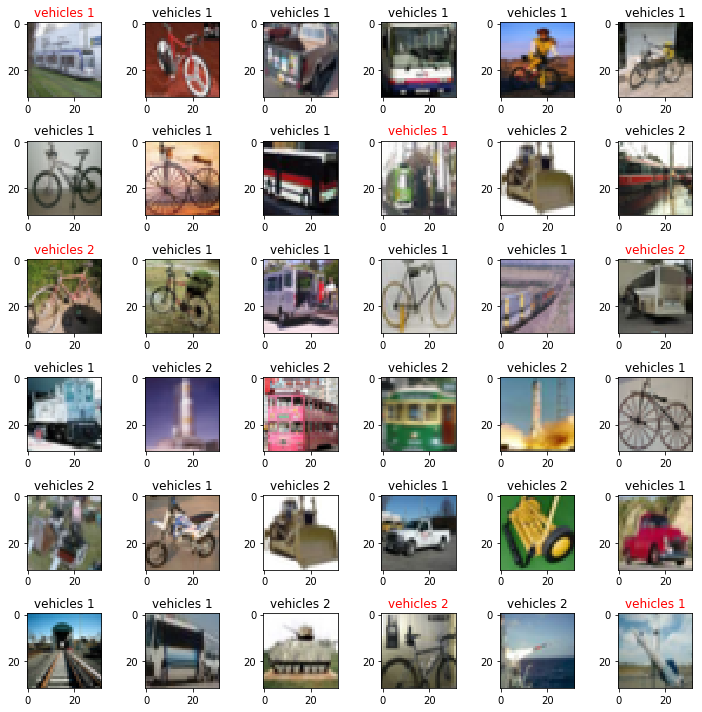


*Figure 17*



*Figure 18*

Output



*Figure 19*

# Conclusion

Accuracy, Computation and loss graphs matter. SVM Algorithm took enough time for generating results with accuracy of nearly 70%. Logistic regression and other algorithms did not generate optimal accuracy and the loss graph which made us decide to use CNN for achieving goal of our project.

Initial CNN model did not result in good accuracy and also was overfitted as it was run without using data augmentation. Later we tried to prevent overfitting and also simultaneously increase accuracy by tuning CNN parameters of learning rate, number of epochs, batch size, steps per epoch, using data augmentation and dropout which should be in 20%-50% range. Trying different layer types and number also helped in achieving accuracy.

# Bibliography

(n.d.). Retrieved from https://adventuresinmachinelearning.com/introduction-to-google-colaboratory/

(n.d.). Retrieved from https://adventuresinmachinelearning.com/introduction-to-google-colaboratory/

(n.d.). Retrieved from adventuresinmachinelearning: https://adventuresinmachinelearning.com/introduction-to-google-colaboratory/

(n.d.). Retrieved from Keras Documentation: https://keras.io/

(n.d.). Retrieved from Towards DataScience: https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

(n.d.). Retrieved from Towards DataScience: https://towardsdatascience.com/epoch-vs-iterations-vs-batch-size-4dfb9c7ce9c9

*CIFAR 100 DATASET*. (n.d.). Retrieved from CIFAR 100 DATASET: https://www.cs.toronto.edu/~kriz/cifar.html