VISVESVARAYA TECHNOLOGICAL UNIVERSITY

Jnana Sangama, Belagavi – 590 018.

An Internship Report

On

"Machine Learning"

18.07.2020 – 20.08.2020

Submitted in partial fulfilment of the for the award of the degree of

Bachelor of Engineering

In

Computer Science & Engineering

Submitted by

Rajashree **1VI17CS114**

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

VEMANA INSTITUTE OF TECHNOLOGY

BENGALURU – 560034

2020-2021

Karnataka Reddy Jana Sangha®

VEMANA INSTITUTE OF TECHNOLOGY

(Affiliated to Visvesvaraya Technological University, Belagavi)

Koramangala, Bengaluru-560034.

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

CERTIFICATE

This is to certify that the Internship/Professional Practice work entitled "**MACHINE LEARNING**" is a bonafide work carried out by **Ms. RAJASHREE (1VI17CS114)** during the academic year 2020-21 in partial fulfilment of the requirement for the award of **Bachelor of Engineering in Computer Science & Engineering** of the **Visvesvaraya Technological University, Belagavi.**

It is certified that all corrections/suggestions indicated for internal assessment have been incorporated in the report. The internship report has been approved as it satisfies the academic requirements in respect of the Internship/ Professional Practice prescribed for the said degree.

Guide	**Head of the Department**	**Principal**
(Mr. Noor Basha)	(Dr. Ramakrishna M)	(Dr. Vijayasimha Reddy B.G)

External Viva

Name of the Examiner — Signature with date

1._____

2._____

ACKNOWLEDGEMENT

I sincerely thank Visvesvaraya Technology University for providing a platform to do the internship work.

I express my sincere thanks to **Dr. Vijayasimha Reddy B G**, Principal, Vemana Institute of Technology, Bengaluru, for providing necessary facilities and motivation to carry out internship work successfully.

I express heartfelt gratitude and humble thanks to **Dr. Ramakrishna M**, HoD, CSE, Vemana Institute of Technology, for his constant encouragement, inspiration and help to carry out internship work successfully.

I am very thankful to my external guide, **Farheen Farhath,** chief Execuive officer, at Company name who has given in-time valuable instructions and put me in contact with experts in the field, with extensive guidance regarding practical issues.

I would like to express my sincere gratitude towards my internal guide, **Mr.Noor Pasha** ,Assistant Professor for providing encouragement and inspiration throughout the internship.

I thank **Ms. Rachitha M V** and **Ms. Ruma Panda**, Assistant Professor, the internship coordinators for their boundless cooperation and support during the internship work.

I am thankful to all the teaching and non-teaching staff members of Electronics & Communication Engineering Department for their help and much needed support throughout the internship.

Rajashree

1VI17CS114

Internship Certificate given from the company

Prinston Smart Engineering

ABSTRACT

The Machine Learning field, which can be briefly defined as enabling computers make successful predictions using past experiences, has exhibited an impressive development recently with the help of the rapid increase in the storage capacity and processing power of computers.

Heart disease is a major life threatening disease that can cause either death or a serious long term disability. However, there is lack of effective tools to discover hidden relationships and trends in e-health data. Medical diagnosis is a complicated task and plays a vital role in saving human lives so it needs to be executed accurately and efficiently. An appropriate and accurate computer based automated decision support system is required to reduce cost for achieving clinical tests. This paper provides an insight into machine learning techniques used in diagnosing various diseases. Various data mining classifiers have been discussed which has emerged in recent years for efficient and effective disease diagnosis.

However using data mining technique can reduce the number of test that are required. In order to reduce from heart diseases there have to be a quick and efficient detection technique. Decision Tree is one of the effective data mining methods used.

This research compares different algorithms of Decision Tree classification seeking better performance in heart disease diagnosis. The algorithms which are tested are SVM algorithm, K Nearest Neighbour algorithm and Random Forest algorithm.

The goal of this study is to extract hidden patterns by applying data mining techniques, which are noteworthy to heart diseases and to predict the presence of heart disease in patients where this presence is valued from no presence to likely

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

INTRODUCTION

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it to learn for themselves. The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide**.** The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly. But, using the classic algorithms of machine learning, text is considered as a sequence of keywords; instead, an approach based on semantic analysis mimics the human ability to understand the meaning of a text. Machine learning involves computer to get trained using a given data set, and use this training to predict the properties of a given new data. For example, we can train computer by feeding it 1000 images of cats and 1000 more images which are not of a cat, and tell each time to computer whether a picture is cat or not. Then if we show the computer a new image, then from the above training, computer should be able to tell whether this new image is cat or not. Process of training and prediction involves use of specialized algorithms. We feed the training data to an algorithm, and the algorithm uses this training data to give predictions on a new test data. There are various machine learning algorithms like Decision trees, Naive Bayes, Random forest, Support vector machine, K-nearest neighbour, K- means clustering, etc. Machine Learning is the art (and science) of enabling machines to learn things which are not explicitly programmed.

1.1 History Of Machine Learning

The term machine learning was coined in 1959 by Arthur Samuel, an American IBMer and pioneer in the field of Computer gaming and artificial intelligence. A representative book of the machine learning research during the 1960s was the Nilsson's book on Learning Machines, dealing mostly with machine learning for pattern classification. Interest related to pattern recognition continued into the 1970s, as described by Duda and Hart in 1973. In 1981 a report was given on using teaching strategies so that a neural network learns to recognize 40 characters (26 letters, 10 digits, and 4 special symbols) from a computer terminal. Tom M.

Mitchell provided a widely quoted, more formal definition of the algorithms studied in the machine learning field: "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*." This definition of the tasks in which machine learning is concerned offers a fundamentally operational definition rather than defining the field in cognitive terms. This follows Alan Turning's proposal in his paper "Computing machinery and Intelligence", in which the question is replaced with the question . Modern day machine learning has two objectives, one is to classify data based on models which have been developed, the other purpose is to make predictions for future outcomes based on these models. A hypothetical algorithm specific to classifying data may use computer vision of moles coupled with supervised learning in order to train it to classify the cancerous moles. Where as, a machine learning algorithm for stock trading may inform the trader of future potential predictions.

1.1.1 Artificial Intelligence

Machine learning (ML), reorganized as a separate field, started to flourish in the 1990s. The field changed its goal from achieving artificial intelligence to tackling solvable problems of a practical nature. It shifted focus away from the Symbolic approaches it had inherited from AI, and toward methods and models borrowed from statistics and probability theory. As of 2020, many sources continue to assert that machine learning remains a subfield of AI. The main disagreement is whether all of ML is part of AI, as this would mean that anyone using ML could claim they are using AI. Others have the view that not all of ML is part of AI where only an 'intelligent' subset of ML is part of AI. The question to what is the difference between ML and AI is answered by Judea Pearl in The Book of Why. Accordingly ML learns and predicts based on passive observations, whereas AI implies an agent interacting with the environment to learn and take actions that maximize its chance of successfully achieving its goals.

1.1.2 Data Mining

Machine learning and Data Mining often employ the same methods and overlap significantly, but while machine learning focuses on prediction, based on known properties learned from the training data, Data Mining focuses on the discovery of unknown properties in the data . Data mining uses many machine learning methods, but with different goals; on the other hand, machine learning also employs data mining methods as "unsupervised learning" or as a

preprocessing step to improve learner accuracy. Much of the confusion between these two research communities comes from the basic assumptions they work with: in machine learning, performance is usually evaluated with respect to the ability to reproduce known knowledge, while in knowledge discovery and data mining (KDD) the key task is the discovery of previously unknown knowledge. Evaluated with respect to known knowledge, an uninformed (unsupervised) method will easily be outperformed by other supervised methods, while in a typical KDD task, supervised methods cannot be used due to the unavailability of training data.

1.1.3 Optimization

Machine learning also has intimate ties to optimization many learning problems are formulated as minimization of some loss function on a training set of examples. Loss functions express the discrepancy between the predictions of the model being trained and the actual problem instances (for example, in classification, one wants to assign a label to instances, and models are trained to correctly predict the pre-assigned labels of a set of examples)

1.1.4 Generalization

The difference between optimization and machine learning arises from the goal of generalization: while optimization algorithms can minimize the loss on a training set, machine learning is concerned with minimizing the loss on unseen samples. Characterizing the generalization of various learning algorithms is an active topic of current research, especially for deep learning algorithms.

1.1.5 Statistics

Machine learning and statistics are closely related fields in terms of methods, but distinct in their principal goal: statistics draws population inferences from a sample, while machine learning finds generalizable predictive patterns. According to Michal I. Jordan, the ideas of machine learning, from methodological principles to theoretical tools, have had a long pre-history in statistics. He also suggested the term data science as a placeholder to call the overall field.Leo Breiman distinguished two statistical modeling paradigms data model and algorithmic model, wherein "algorithmic model" means more or less the machine learning algorithms like Random forest Some statisticians have adopted methods from machine learning, leading to a combined field that they call statistical learning.

1.2 Scope

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves. The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly. The main feature of the project is to generate an approximate forecasting output and create a general idea of future values based on the previous data by generating a pattern. The scope of this project does not exceed more than a generalized suggestion tool

1.3 Machine learning as Python application.

Python is widely used general purpose,high level programming language. It was initially designed by Guido van Rossum in 1991 and developed by Python Software Foundation.It was mainly developed for emphasis on code readability, and its syntax allows programmers to express concepts in fewer lines of code. The most recent major version of Python is Python 3, which we shall be using in this tutorial. Python can be used on a server to create web applications. It can be used along side software to create workflows. It can connect to database system and also read and modify files.

Chapter 2

ORGANIZATION PROFILE

Company Name : Prinston Smart Engineers

Founder :Mr. Asif Akhter

PRINSTON SMART ENGINEERS traces its roots back to 2004 in New Delhi and since then have never looked back. With Dozens of successful projects under our belt, we can proudly say that we are one of the most trusted Engineering, Maintenance &Training Services in Delhi,and also extended the services across India in various states such as Karnataka mainly in Bangalore & Mysore,Jaipur,Gujarat In 2016 we expanded our services in Skill development and Training program for Engineering students in different domain.India today produce 1.5 million Engineering graduates a year, it is however agreed by all that 75% of these graduates are unemployed,agreeing to the fact that there is need to provide Skill training to the Engineering students.In 2018 when AICTE announced that Internship is mandatory for Engineering students to ensure that technical students get exposure to the Industrial environment, current technology relevant to their subject and opportunities to learn understand and sharpen real time technical and managerial skills. On request from colleges we started providing quality internship from our experts team to the students. In 2020 Prinston Smart Engineers collaborated with Wedir-Tech Trading Contracting & Services W.L.L, Doha, Qatar for mutual benefit and better service. Prinston Smart Engineers is an certified training division specializing in effective training of B.Tech and Diploma students.They trained over 30000+ Students for science & Technology program in various schools and various Technical Certifications For Engineering Students.Courses are designed to cover all aspects of most of our real projects over the years to provide for smooth transfer of Real knowledge and expertise that we have gained over the years.Training is Hands-on and covers all required fundamental and specialist know-how in system operation, troubleshooting & maintenance, designing and installation

2.1 Training Domain produced

- HVAC Design and Drafting for Mechanical Engineers
- UI Design and Development for CS,ECE,IS Students
- Electrical Design and Lighting Design for Electrical Engineers

- Machine Learning
- Cyber security
- Software testing
- python

Chapter 3

IMPLEMENTATION

3.1 PHASES OF MACHINE LEARNING

3.1.1 Data Acquisition

Data acquisition Machine learning needs two things to work, data (lots of it) and models. When acquiring the data, be sure to have enough features (aspect of data that can help for a prediction, like the surface of the house to predict its price) populated to train correctly your learning model.

3.1.2 Data Preparation

Data preparation involves five sub-processes to be followed. They are selection, cleansing, construction, integration, and formatting of data. In other words, all these steps comprise all the activities that must be performed for construction of the final data set.

3.1.3 Choosing and Training a Model

When designing a Machine Learning solution for a real-world problem, it's important to remember that the goal is not just to train a model to make accurate predictions on a representative dataset, but to train a model to make accurate predictions on data points seen in the field.

3.1.4 Evaluation and Parameter Tuning

In machine learning, the specific model you are using is the function and requires parameters in order to make a prediction on new data. Whether a model has a fixed or variable number of parameters determines whether it may be referred to as "parametric" or "nonparametric". The weights in an artificial neural network.

3.1.5 Prediction

Prediction refers to the output of an algorithm after it has been trained on a historical dataset and applied to new data when forecasting the likelihood of a particular outcome Prediction is

at the heart of almost every scientific discipline, and the study of generalization (that is, prediction) from data is the central topic of machine learning and statistics, and more generally, data mining. Machine learning and statistical methods are used throughout the scientific world for their use in handling the "information overload" that characterizes our current digital age.

3.2 Classification Of Machine Learning

- Logistic Regression
- K Nearest Neighbours
- Decision Trees
- Random Forest
- Support Vector Machine

3.2.1 Logistic Regression

It is a classification algorithm in machine learning that uses one or more independent variables to determine an outcome. The outcome is measured with a dichotomous variable meaning it will have only two possible outcomes. The goal of logistic regression is to find a best-fitting relationship between the dependent variable and a set of independent variables. It is better than other binary classification algorithms like nearest neighbour since it quantitatively explains the factors leading to classification.

Advantages and Disadvantages

- Logistic regression is specifically meant for classification, it is useful in understanding how a set of independent variables affect the outcome of the dependent variable.
- The main disadvantage of the logistic regression algorithm is that it only works when the predicted variable is binary, it assumes that the data is free of missing values and assumes that the predictors are independent of each other.

3.2.2 K Nearest Neighbours

It is a lazy learning algorithm that stores all instances corresponding to training data in n-dimensional space. It is a lazy learning algorithm as it does not focus on constructing a general internal model, instead, it works on storing instances of training data. Classification is

computed from a simple majority vote of the k nearest neighbours of each point. It is supervised and takes a bunch of labeled points and uses them to label other points. To label a new point, it looks at the labeled points closest to that new point also known as its nearest neighbours. It has those neighbours vote, so whichever label the most of the neighbours have is the label for the new point. The "k" is the number of neighbours it checks.

Advantages And Disadvantages

- This algorithm is quite simple in its implementation and is robust to noisy training data. Even if the training data is large, it is quite efficient.
- The only disadvantage with the KNN algorithm is that there is no need to determine the value of K and computation cost is pretty high compared to other algorithms.

3.2.3 Decision Trees

The decision tree algorithm builds the classification model in the form of a tree structure. It utilizes the if-then rules which are equally exhaustive and mutually exclusive in classification. The process goes on with breaking down the data into smaller structures and eventually associating it with an incremental decision tree. The final structure looks like a tree with nodes and leaves. The rules are learned sequentially using the training data one at a time. Each time a rule is learned, the tuples covering the rules are removed. The process continues on the training set until the termination point is met. The tree is constructed in a top-down recursive divide and conquer approach. A decision node will have two or more branches and a leaf represents a classification or decision. The topmost node in the decision tree that corresponds to the best predictor is called the root node, and the best thing about a decision tree is that it can handle both categorical and numerical data.

Advantages and Disadvantages

- A decision tree gives an advantage of simplicity to understand and visualize, it requires very little data preparation as well.
- The disadvantage that follows with the decision tree is that it can create complex trees that may bot categorize efficiently. They can be quite unstable because even a simplistic change in the data can hinder the whole structure of the decision tree.

3.2.4 Random Forest

Random decision trees or random forest are an ensemble learning method for classification, regression, etc. It operates by constructing a multitude of decision trees at training time and outputs the class that is the mode of the classes or classification or mean prediction(regression) of the individual trees. A random forest is a meta-estimator that fits a number of trees on various subsamples of data sets and then uses an average to improve the accuracy in the model's predictive nature. The sub-sample size is always the same as that of the original input size but the samples are often drawn with replacements.

Advantages and Disadvantages

- The advantage of the random forest is that it is more accurate than the decision trees due to the reduction in the over-fitting.
- The only disadvantage with the random forest classifiers is that it is quite complex in implementation and gets pretty slow in real-time prediction.

3.2.5 Support Vector Machine

The support vector machine is a classifier that represents the training data as points in space separated into categories by a gap as wide as possible. New points are then added to space by predicting which category they fall into and which space they will belong to.

Advantages and Disadvantages

- It uses a subset of training points in the decision function which makes it memory efficient and is highly effective in high dimensional spaces.
- The only disadvantage with the support vector machine is that the algorithm does not directly provide probability estimates.

3.3 Regression in Machine Learning

- Simple Linear Regression
- Multiple Linear Regression
- Polynomial Regression
- Support Vector Regression
- Decision Tree Regression
- Random Forest Regression

3.3.1 Simple Linear Regression

It is also called linear regression. It establishes the relationship between two variables using a straight line. Linear regression attempts to draw a line that comes closest to the data by finding the slope and intercept that define the line and minimize regression errors.If two or more explanatory variables have a linear relationship with the dependent variable, the regression is called a multiple linear regression Many data relationships do not follow a straight line, so statisticians use n nonlinear regression instead. The two are similar in that both track a particular response from a set of variables graphically. But nonlinear models are more complicated than linear models because the function is created through a series of assumptions that may stem from trial and error.

3.3.2 Multiple Linear Regression

It is rare that a dependent variable is explained by only one variable. In this case, an analyst uses multiple regression, which attempts to explain a dependent variable using more than one independent variable. Multiple regressions can be linear and nonlinear. Multiple regressions are based on the assumption that there is a linear relationship between both the dependent and independent variables. It also assumes no major correlation between the independent variables. there are several different advantages to using regression analysis. These models can be used by businesses and economists to help make practical decisions.

3.3.3 Polynomial Regression

In statistics, polynomial regression is a form of regression analysis in which the relationship between the independent variable x and the dependent variable y is modelled as an nth degree polynomial in x. Polynomial regression fits a nonlinear relationship between the value of x and the corresponding conditional mean of y, denoted $E(y|x)$. Although *polynomial regression* fits a nonlinear model to the data, as a statistical estimation problem it is linear, in the sense that the regression function $E(y|x)$ is linear in the unknown parameter that are estimated from the data. For this reason, polynomial regression is considered to be a special case of multiple linear regression.

3.3.4 Support Vector Regression

In SVR, we identify a hyperplane with maximum margin such that the maximum number of data points are within that margin. SVRs are almost similar to the SVM classification algorithm. We will discuss the SVM algorithm in detail in my next article.Instead of minimizing the error rate as in simple linear regression, we try to fit the error within a certain threshold. Our objective in SVR is to basically consider the points that are within the margin**.** Our best fit line is the hyperplane that has the maximum number of points.

3.3.5 Decision Tree Regression

Decision trees can be used for classification as well as regression. In decision trees, at each level, we need to identify the splitting attribute. In the case of regression, the ID3 algorithm can be used to identify the splitting node by *reducing the* standard deviation. A decision tree is built by partitioning the data into subsets containing instances with similar values (homogenous). Standard deviation is used to calculate the homogeneity of a numerical sample. If the numerical sample is completely homogeneous, its standard deviation is zero.

The steps for finding the splitting node is briefly described below:

- Calculate the standard deviation of the target variable

- Split the dataset on different attributes and calculate the standard deviation for each branch (standard deviation for target and predictor). This value is subtracted from the standard deviation before the split. The result is the standard deviation reduction.
- The attribute with the largest standard deviation reduction is chosen as the splitting node.
- The dataset is divided based on the values of the selected attribute. This process is run recursively on the non-leaf branches until all data is processed.
- To avoid overfitting, the Coefficient of Deviation **(**CV**)** is used which decides when to stop branching. Finally, the average of each branch is assigned to the related leaf node (in regression mean is taken whereas in classification mode of leaf nodes is taken)

3.3.6 Random Forest Regression

Random forest is an ensemble approach where we take into account the predictions of several decision regression trees. Select K random points Identify n where n is the number of decision tree regressors to be created. Repeat steps 1 and 2 to create several regression trees. The average of each branch is assigned to the leaf node in each decision tree. To predict output for a variable, the average of all the predictions of all decision trees are taken into consideration. Random Forest prevents overfitting (which is common in decision trees) by creating random subsets of the features and building smaller trees using these subsets.

3.4 Convolution Neural Network

Technically, the convolution as described in the use of convolutional neural networks is actually a " cross-correlation". Nevertheless, in deep earning, it is referred to as a " convolution operation.Convolution is a specialized kind of linear operation. Convolutional networks are simply neural networks that use convolution in place of general matrix multiplication in at least one of their layers A convolutional neural network consists of an input and an output layer In deep learning, a convolutional neural network (CNN, or ConvNet) is a class of deep neural networks, most commonly applied to analyzing visual imagery. They are also known as shift invariant or space invariant artificial neural networks (SIANN), based on their shared-weights architecture and translation invariance characteristics.

3.5 Components in machine learning

Machine learning solutions are used to solve a wide variety of problems, but in nearly all cases the core components are the same. Whether you simply want to understand the skeleton of machine learning solutions better or are embarking on building your own, understanding these components - and how they interact - can help.

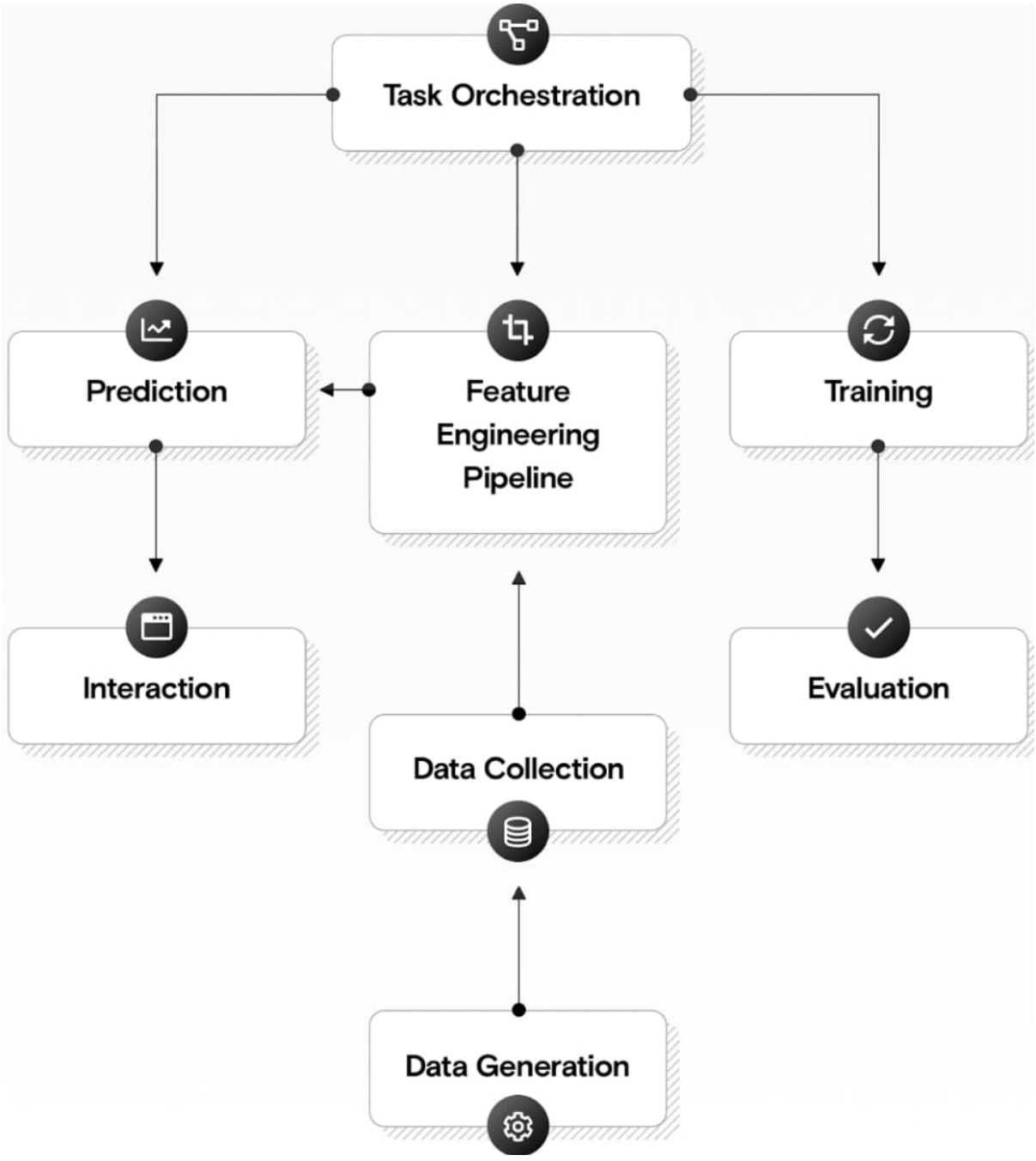


Fig 3.5.1 : Architecture Of Machine Learning

The components of a machine learning solution

1. **Data Generation:** Every machine learning application lives off data. That data has to come from somewhere. Usually it's generated by one of your core business functions.
2. **Data Collection:** Data is only useful if it's accessible, so it needs to be stored – ideally in a consistent structure and conveniently in one place.
3. **Feature Engineering Pipeline:** Algorithms can't make sense of raw data. We have to select, transform, combine, and otherwise prepare our data so the algorithm can find useful patterns.
4. **Training:** This is where the magic happens. We apply algorithms, and they learn patterns from the data. Then they use these patterns to perform particular tasks.
5. **Evaluation:** We need to carefully test how well our algorithm performs on data it hasn't seen before (during training). This ensures we don't use prediction models that work well on "seen" data, but not in real-world settings.
6. **Task Orchestration:** Feature engineering, training, and prediction all need to be scheduled on our compute infrastructure (such as AWS or Azure) – usually with non-trivial interdependence. So we need to reliably orchestrate our tasks.
7. **Prediction:** This is the money maker. We use the model we've trained to perform new tasks and solve new problems – which usually means making a prediction.
8. **Infrastructure:** Even in the age of the cloud, the solution has to live and be served somewhere. This will require setup and maintenance.
9. **Authentication:** This keeps our models secure and makes sure only those who have permission can use them.
10. **Interaction:** We need some way to interact with our model and give it problems to solve. Usually this takes the form of an API, a user interface, or a command-line interface.
11. **Monitoring:** We need to regularly check our model's performance. This usually involves periodically generating a report or showing performance history in a dashboard.

Chapter 4

ENVIRONMENT SETUP

4.1 GOOGLE COLABS

Google Colab is a great platform for deep learning enthusiasts, and it can also be used to test basic machine learning models, gain experience, and develop an intuition about deep learning aspects such as hyperparameter tuning, preprocessing data, model complexity, overfitting and more. Colaboratory by Google (Google Colab in short) is a Jupyter notebook based runtime environment which allows you to run code entirely on the cloud.

This is necessary because it means that you can train large scale ML and DL models even if you don't have access to a powerful machine or a high speed internet access. Google Colab supports both GPU and TPU instances, which makes it a perfect tool for deep learning and data analytics enthusiasts because of computational limitations on local machines. Since a Colab notebook can be accessed remotely from any machine through a browser, it's well suited for commercial purposes as well.

- Getting around in Google Colab
- Installing python libraries in Colab
- Downloading large datasets in Colab
- Training a Deep learning model in Colab
- Using TensorBoard in Colab

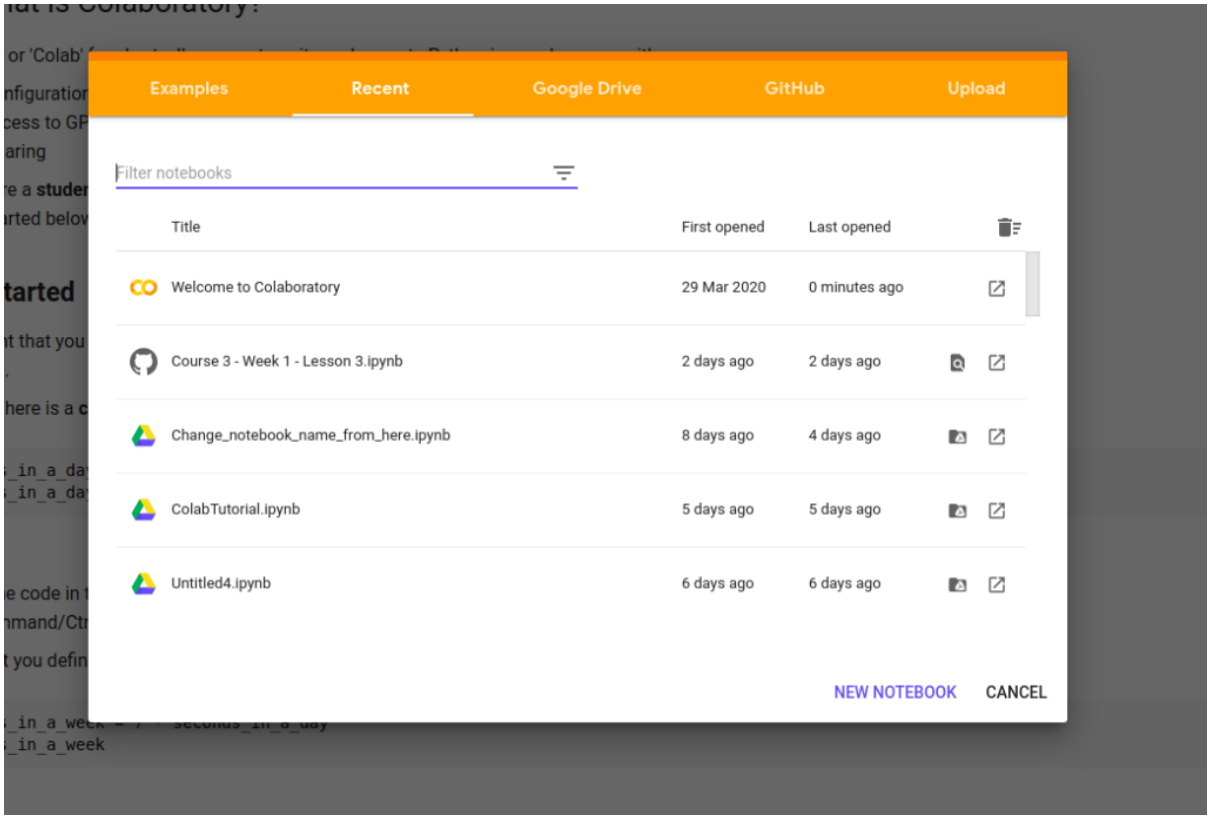
4.1.1 Creating first .ipynb notebook in colab

- Open a browser of your choice and go to colab.research.google.com and sign in using your Google account. Click on a new notebook to create a new runtime instance.

- In the top left corner, you can change the name of the notebook from "Untitled.ipynb" to the name of your choice by clicking on it.

- The cell execution block is where you type your code. To execute the cell, press shift + enter.

- The variable declared in one cell can be used in other cells as a global variable. The environment automatically prints the value of the variable in the last line of the code block if stated explicitly



4.1.2 Training a sample tensorflow model

Training a machine learning model in Colab is very easy. The best part about it is not having to set up a custom runtime environment, it's all handled for you. For example, let's look at training a basic deep learning model to recognize handwritten digits trained on the MNIST dataset. The data is loaded from the standard Keras dataset archive. The model is very basic, it categorizes images as numbers and recognizes them.

Setup:

```
#import necessary libraries
import tensorflow as tf
#load training data and split into train and test sets
mnist = tf.keras.datasets.mnist

(x_train,y_train), (x_test,y_test) = mnist.load_data()
```

```
x_train, x_test = x_train / 255.0, x_test / 255.0
```

The output for this code snippet will look like this:

```
Downloading data from https://storage.googleapis.com/tensorflow/tf-keras-datasets/mnist.npz
11493376/11490434 [==============================] - 0s 0us/step
```

4.1.3 Installing packages in Google Colab

You can use the code cell in Colab not only to run Python code but also to run shell commands. Just add a **!** before a command. The exclamation point tells the notebook cell to run the following command as a shell command. Most general packages needed for deep learning come pre-installed. In some cases, you might need less popular libraries, or you might need to run code on a different version of a library. To do this, you'll need to install packages manually.

The package manager used for installing packages is pip.

```
[1]  import tensorflow as tf
```

```
[2]  tf.__version__

     '2.3.0'
```

To install a particular version of TensorFlow use this command:

```
!pip3 install tensorflow==1.5.0
```

The following output is expected after running the above command:

```
!pip3 install tensorflow==1.5.0

Collecting tensorflow==1.5.0
  Downloading https://files.pythonhosted.org/packages/04/79/a37d0b373757b4d283c674a64127bd8864d69f881c639b1e
     |████████████████████████████████| 44.4MB 90kB/s
Requirement already satisfied: wheel>=0.26 in /usr/local/lib/python3.6/dist-packages (from tensorflow==1.5.0
Requirement already satisfied: absl-py>=0.1.6 in /usr/local/lib/python3.6/dist-packages (from tensorflow==1.
Requirement already satisfied: protobuf>=3.4.0 in /usr/local/lib/python3.6/dist-packages (from tensorflow==1
Collecting tensorflow-tensorboard<1.6.0,>=1.5.0
  Downloading https://files.pythonhosted.org/packages/cc/fa/91c06952517b4f1bc075545b062a4112e30cebe558a6b962
     |████████████████████████████████| 3.0MB 40.0MB/s
Requirement already satisfied: numpy>=1.12.1 in /usr/local/lib/python3.6/dist-packages (from tensorflow==1.5
Requirement already satisfied: six>=1.10.0 in /usr/local/lib/python3.6/dist-packages (from tensorflow==1.5.0
Requirement already satisfied: setuptools in /usr/local/lib/python3.6/dist-packages (from protobuf>=3.4.0->t
Requirement already satisfied: werkzeug>=0.11.10 in /usr/local/lib/python3.6/dist-packages (from tensorflow-
Collecting bleach==1.5.0
  Downloading https://files.pythonhosted.org/packages/33/70/86c5fec937ea4964184d4d6c4f0b9551564f821e1c357590
Requirement already satisfied: markdown>=2.6.8 in /usr/local/lib/python3.6/dist-packages (from tensorflow-te
Collecting html5lib==0.9999999
  Downloading https://files.pythonhosted.org/packages/ae/ae/bcb60402c60932b32dfaf19bb53870b29eda2cd17551ba56
     |████████████████████████████████| 890kB 40.6MB/s
Requirement already satisfied: importlib-metadata; python_version < "3.8" in /usr/local/lib/python3.6/dist-p
Requirement already satisfied: zipp>=0.5 in /usr/local/lib/python3.6/dist-packages (from importlib-metadata;
Building wheels for collected packages: html5lib
  Building wheel for html5lib (setup.py) ... done
  Created wheel for html5lib: filename=html5lib-0.9999999-cp36-none-any.whl size=107220 sha256=d9383b6974fb8
  Stored in directory: /root/.cache/pip/wheels/50/ae/f9/d2b189788efcf61d1ee0e36045476735c838898eef1cad6e29
Successfully built html5lib
Installing collected packages: html5lib, bleach, tensorflow-tensorboard, tensorflow
  Found existing installation: html5lib 1.0.1
    Uninstalling html5lib-1.0.1:
      Successfully uninstalled html5lib-1.0.1
  Found existing installation: bleach 3.2.1
    Uninstalling bleach-3.2.1:
      Successfully uninstalled bleach-3.2.1
  Found existing installation: tensorflow 2.3.0
    Uninstalling tensorflow-2.3.0:
      Successfully uninstalled tensorflow-2.3.0
Successfully installed bleach-1.5.0 html5lib-0.9999999 tensorflow-1.5.0 tensorflow-tensorboard-1.5.1
WARNING: The following packages were previously imported in this runtime:
  [tensorboard,tensorflow]
You must restart the runtime in order to use newly installed versions.

RESTART RUNTIME
```

Click on RESTART RUNTIME for the newly installed version to be used.

```
[3]  tf.__version__

     '1.5.0'
```

4.1.4 Downloading a dataset

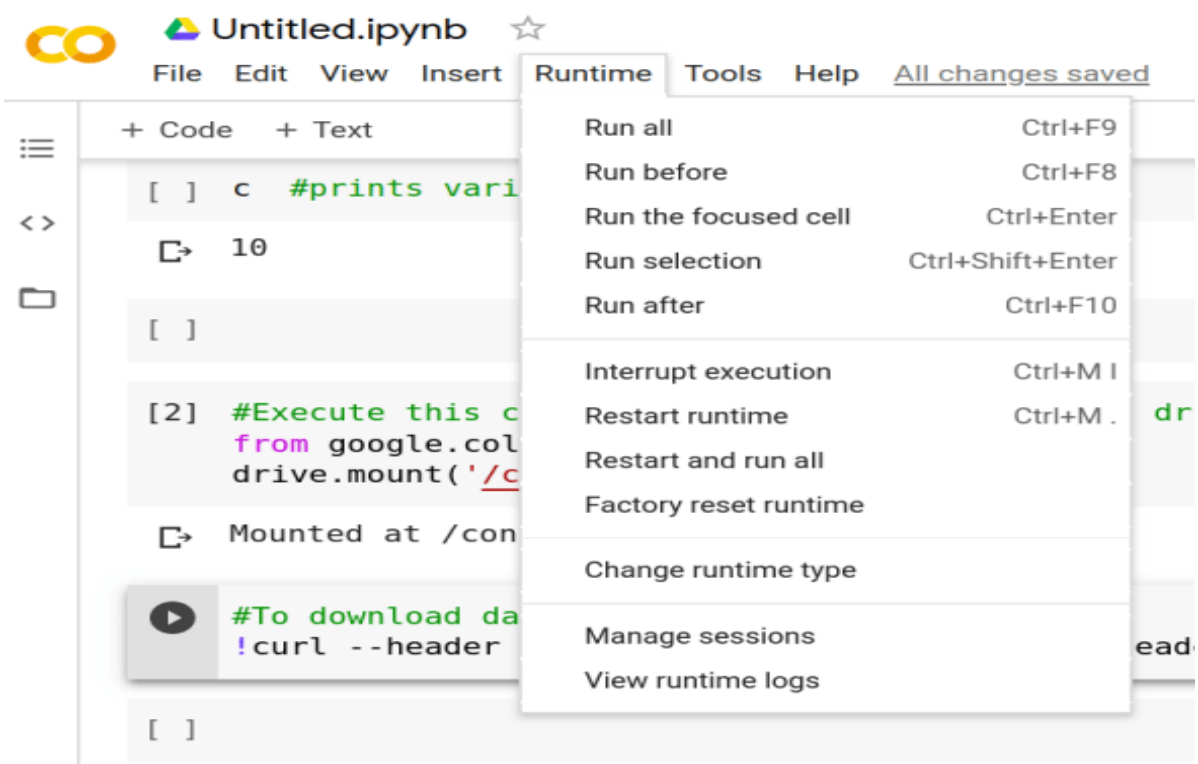
When you're training a machine learning model on your local machine, you're likely to have trouble with the storage and bandwidth costs that come with downloading and storing the dataset required for training a model.

Deep learning datasets can be massive in size, ranging between 20 to 50 Gb. Downloading them is most challenging if you're living in a developing country, where getting high-speed internet isn't possible.

The most efficient way to use datasets is to use a cloud interface to download them, rather than manually uploading the dataset from a local machine. Colab gives us a variety of ways to download the dataset from common data hosting platforms.

4.1.5 Initiating a runtime with GPU/TPU enabled

Deep learning is a computationally expensive process, a lot of calculations need to be executed at the same time to train a model. To mitigate this issue, Google Colab offers us not only the classic CPU runtime but also an option for a GPU and TPU runtime as well.The CPU runtime is best for training large models because of the high memory it provides. The GPU runtime shows better flexibility and programmability for irregular computations, such as small batches and non MatMul computations.The TPU runtime is highly-optimized for large batches and CNNs and has the highest training throughput.If you have a smaller model to train, I suggest training the model on GPU/TPU runtime to use Colab to its full potential. To create a GPU/TPU enabled runtime, you can click on runtime in the toolbar menu below the file name. From there, click on "Change runtime type", and then select GPU or TPU under the Hardware Accelerator dropdown menu.



Training more complex and larger models

To train complex models, you often need to load large datasets. It's advisable to load data directly from Google Drive by using the mount drive method.This will import all the data from your Drive to the runtime instance. To get started, you first need to mount your Google Drive where the dataset is stored. You can also use the default storage available in Colab, and download the dataset directly to Colab from GCS or Kaggle.

4.1.6 Mounting a drive

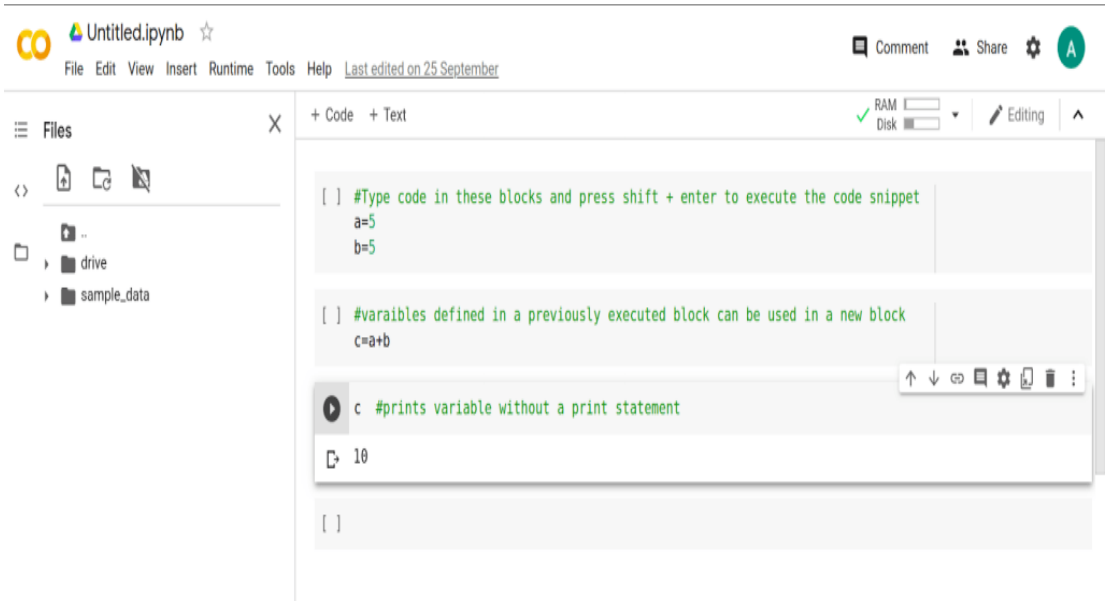
Google Colab allows you to import data from your Google Drive account so that you can access training data from Google Drive, and use large datasets for training.

There are 2 ways to mount a Drive in Colab:

- Using GUI
- Using code snippet

1.Using GUI

Click on the Files icon in the left side of the screen, and then click on the "Mount Drive" icon to mount your Google Drive.



2.Using code snippet

Execute this code block to mount your Google Drive on Colab:

```
from google.colab import drive
drive.mount('/content/drive')
```

Click on the link, copy the code, and paste it into the provided box. Press enter to mount the Drive.

Chapter 5

ASSESSMENTS

5.1 Red Wine Quality

1. Dataset Source:

- Wine Quality Data Set. This dataset is also available from the UCI machine learning
- Dataset Background: Two datasets are included, related to red and white wine samples from the north of Portugal. The goal is to use the red wine samples to model red wine quality based on physicochemical tests.
- Input variables (based on physicochemical tests): 1 - fixed acidity 2 - volatile acidity 3 - citric acid 4 - residual sugar 5 - chlorides 6 - free sulfur dioxide 7 - total sulfur dioxide 8 - density 9 - pH 10 - sulphates 11 - alcohol Output variable (based on sensory data): 12 - quality (score between 0 and 10)

2. Problem Definition & Target Variable:

This project aims to determine which chemical features are the best quality red wine indicators. To be more specific, we define below problems for this analysis:

- Show the contribution of each factor to the wine quality in our model
- Show which features are more important in determining the wine quality
- Show which features are less important in determining the wine quality As mentioned earlier, our target variable will be wine quality, which is scored between 0 and 10.

3. Use Scenario:

The wine industry shows a recent growth spurt as social drinking is on the rise. The price of wine depends on a rather abstract concept of wine appreciation by wine tasters, opinion among whom may have a high degree of variability. Pricing of wine depends on such a volatile factor to some extent. Another critical factor in wine certification and quality assessment is physicochemical tests, which are laboratory-based and consider factors like acidity, pH level, sugar, and other chemical properties. The wine market would be of interest if human quality of

tasting can be related to wine's chemical properties so that certification and quality assessment and assurance processes are more controlled.

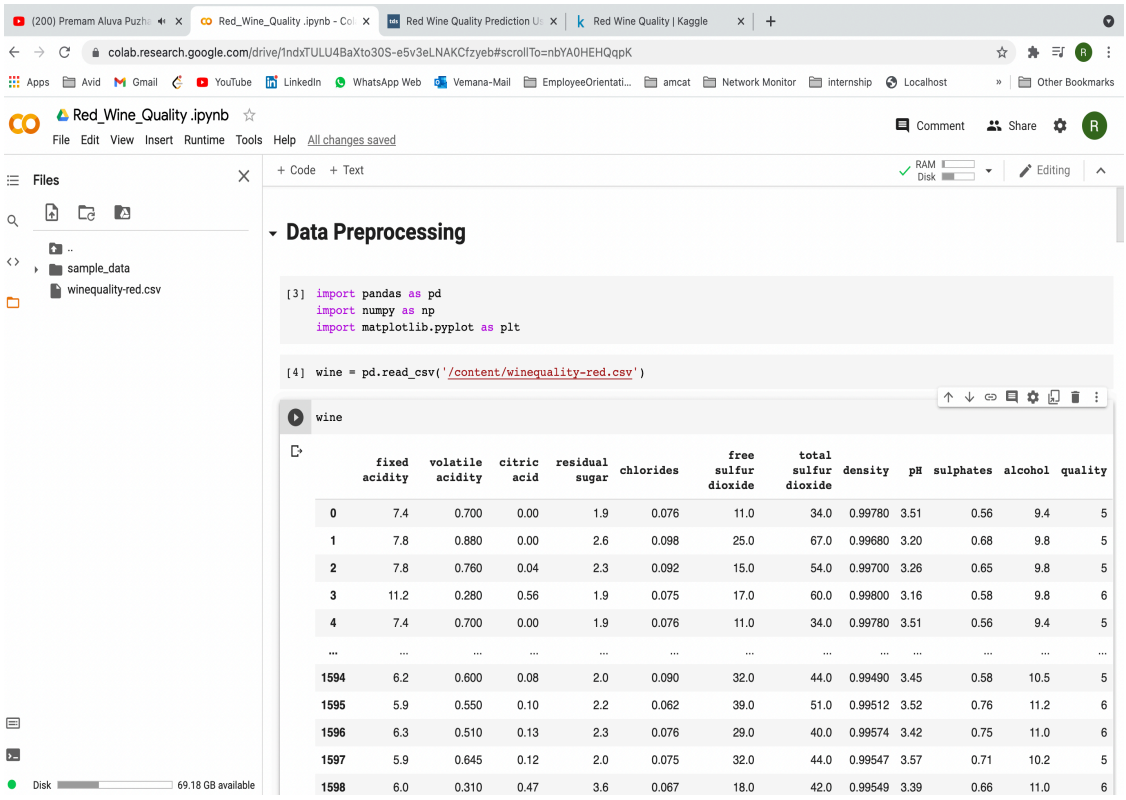
4. Proposed Data Techniques:

- Regression Modeling
- Classification modeling

5. Files included

- Data set: winequality-red.csv

6. Result



	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.700	0.00	1.9	0.076	11.0	34.0	0.99780	3.51	0.56	9.4	5
1	7.8	0.880	0.00	2.6	0.098	25.0	67.0	0.99680	3.20	0.68	9.8	5
2	7.8	0.760	0.04	2.3	0.092	15.0	54.0	0.99700	3.26	0.65	9.8	5
3	11.2	0.280	0.56	1.9	0.075	17.0	60.0	0.99800	3.16	0.58	9.8	6
4	7.4	0.700	0.00	1.9	0.076	11.0	34.0	0.99780	3.51	0.56	9.4	5
...	...	...	...	...	...	...	...	...	...	...	...	...
1594	6.2	0.600	0.08	2.0	0.090	32.0	44.0	0.99490	3.45	0.58	10.5	5
1595	5.9	0.550	0.10	2.2	0.062	39.0	51.0	0.99512	3.52	0.76	11.2	6
1596	6.3	0.510	0.13	2.3	0.076	29.0	40.0	0.99574	3.42	0.75	11.0	6
1597	5.9	0.645	0.12	2.0	0.075	32.0	44.0	0.99547	3.57	0.71	10.2	5
1598	6.0	0.310	0.47	3.6	0.067	18.0	42.0	0.99549	3.39	0.66	11.0	6

Fig 5.1.1 :Implementing the Data Preprocessing for red wine quality data set

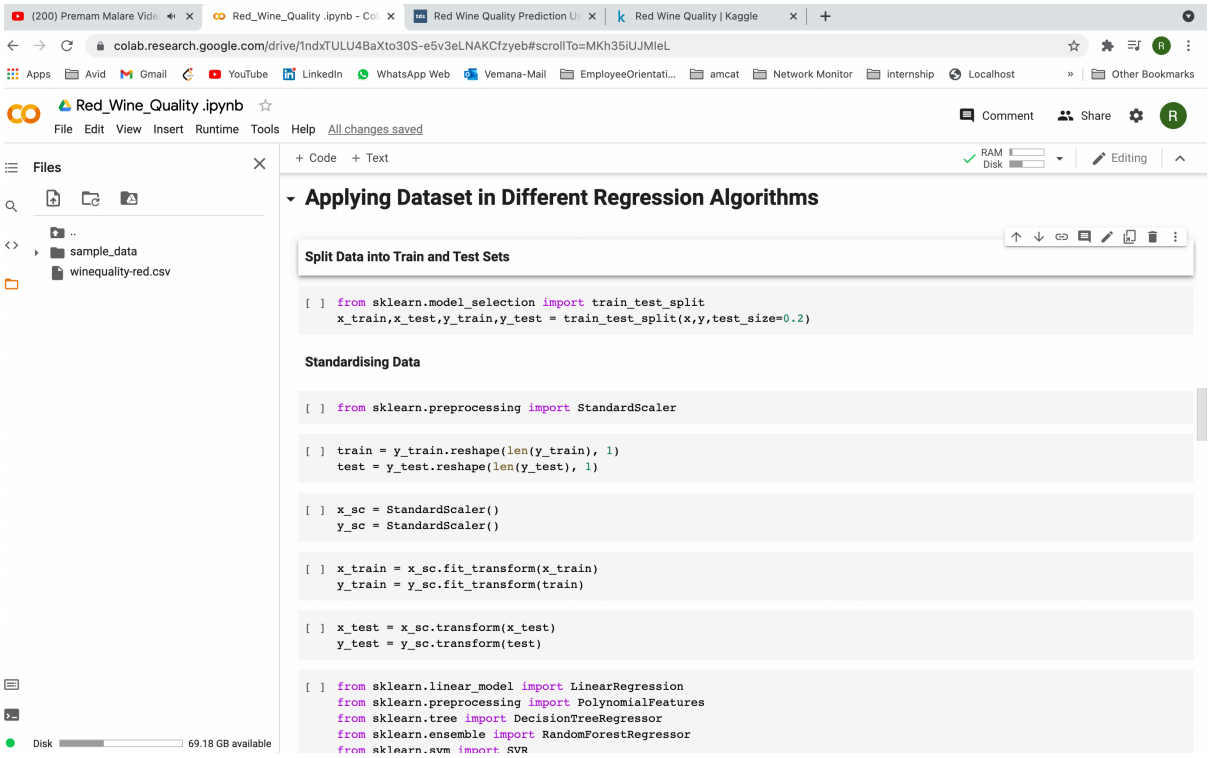


Fig 5.1.2 : Applying different regression algorithms for red wine quality dataset

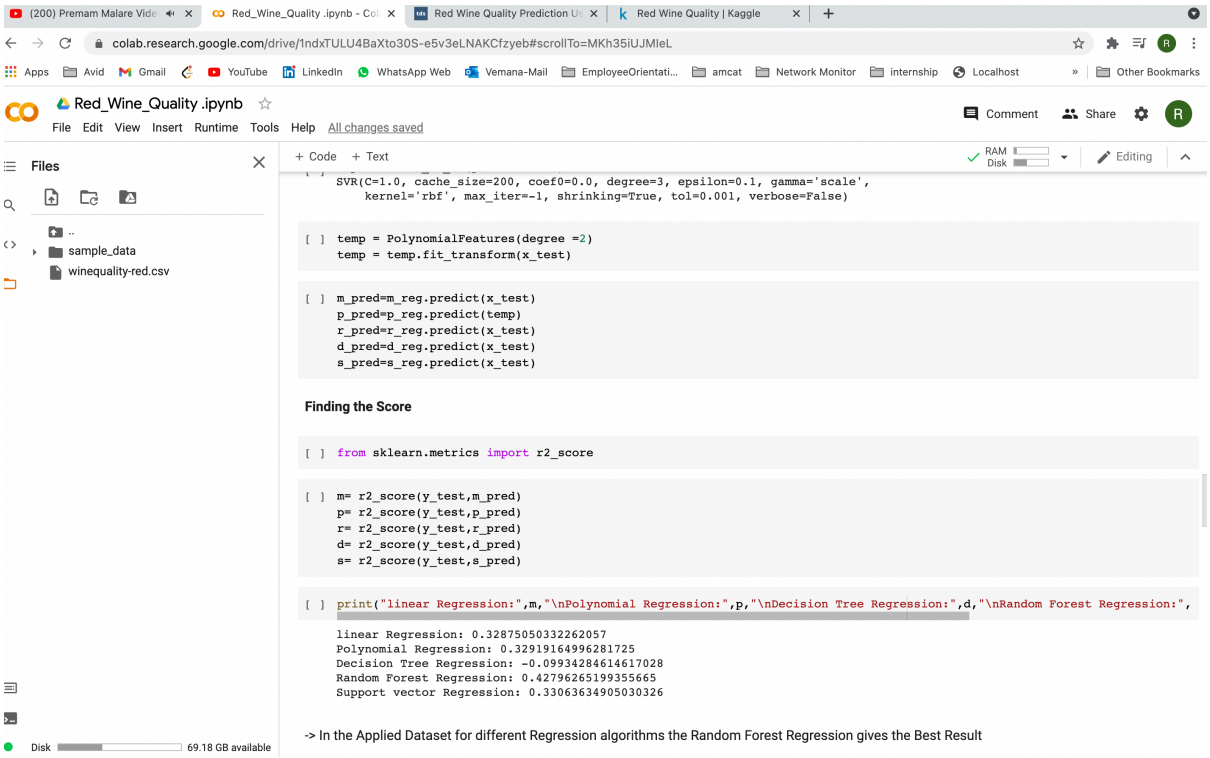


Fig 5.1.3 : Finding the score for red wine quality dataset

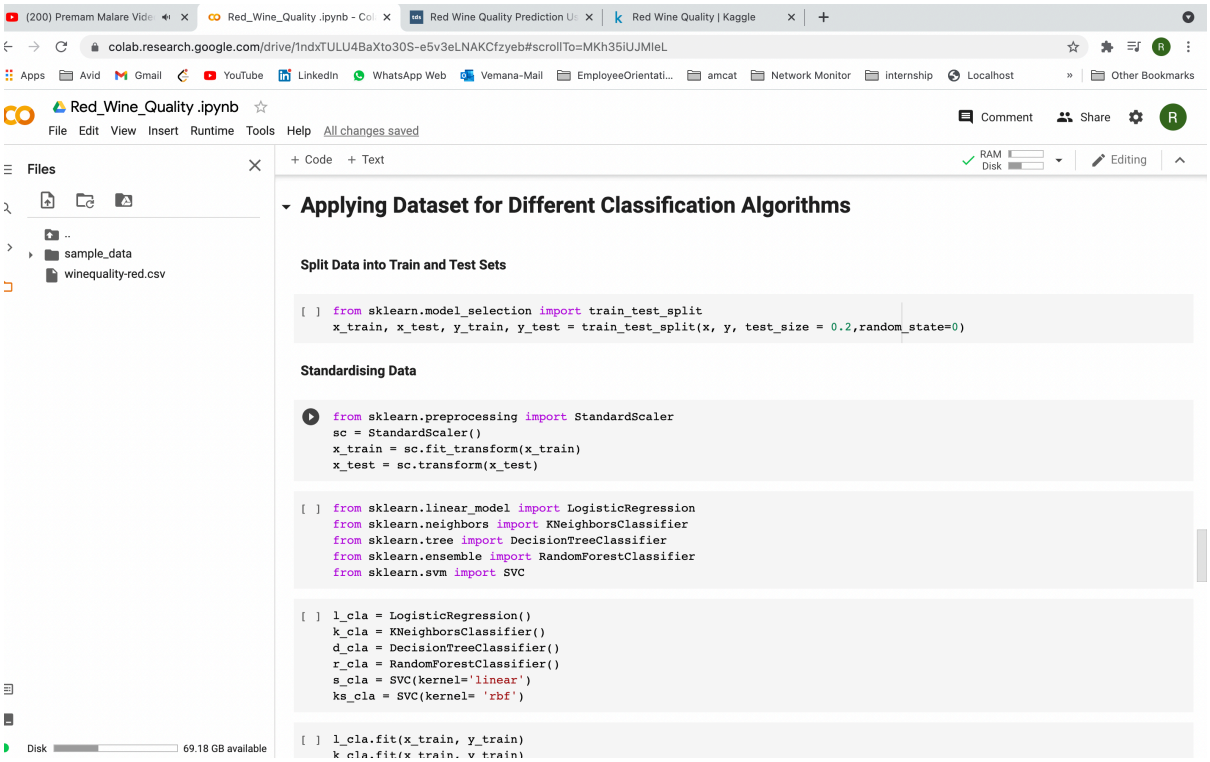


Fig 5.1.4 : Applying different classification algorithms for red wine quality dataset

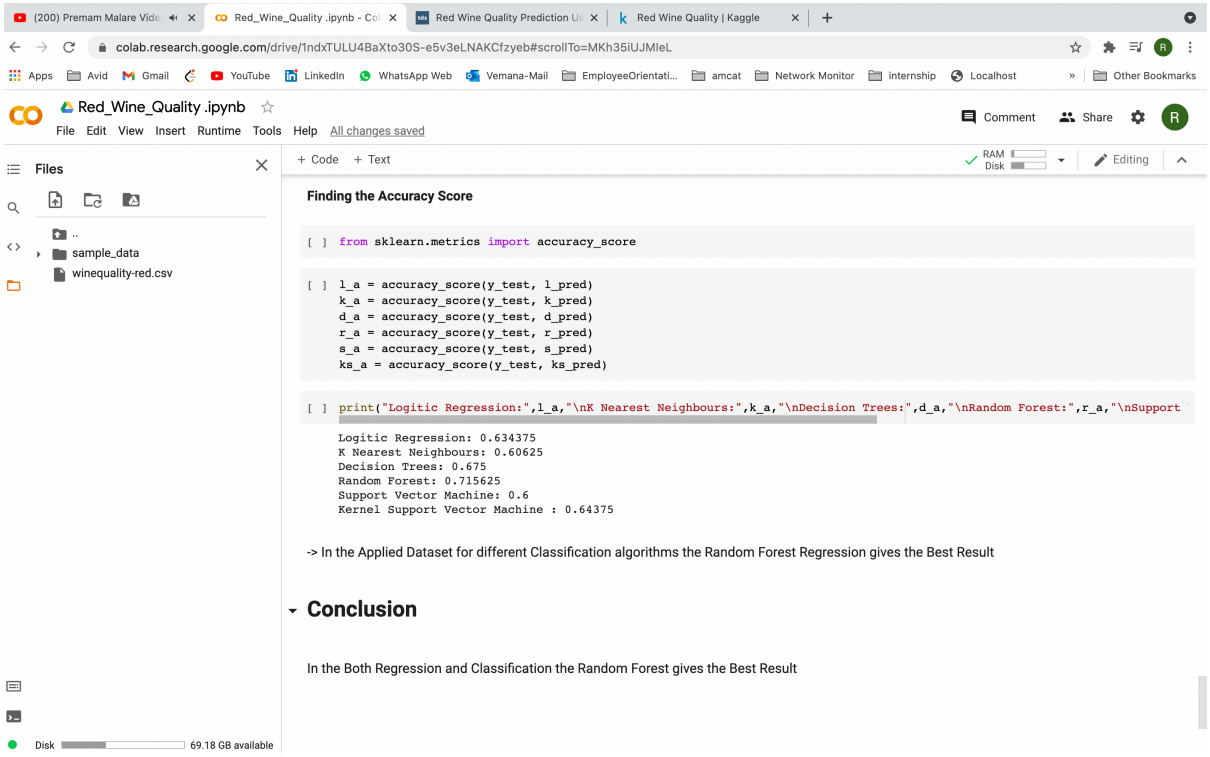
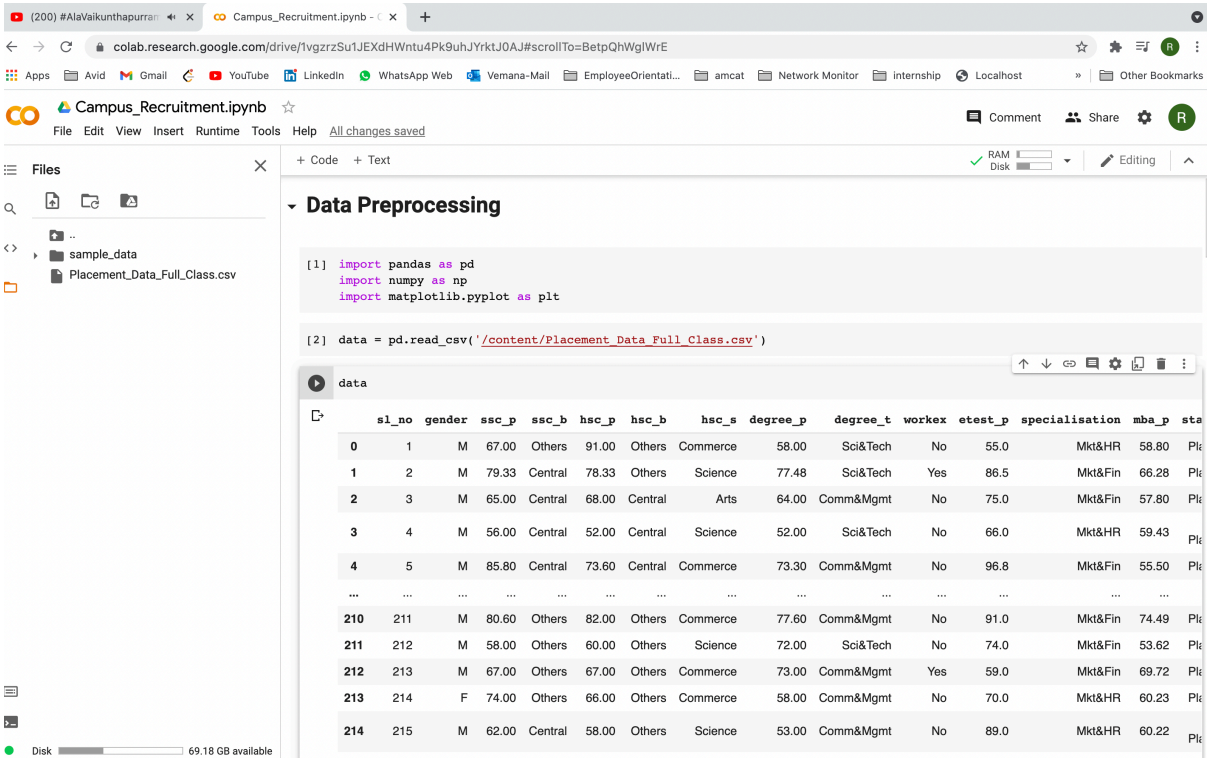


Fig.No 5.1.5 : Finding the Accuracy score for given red wine quality dataset

5.2 Campus Recruitment

This data set consists of Placement data of students in our campus. It includes secondary and higher secondary school percentage and specialization. It also includes degree specialization, type and Work experience and salary offers to the placed students



	sl_no	gender	ssc_p	ssc_b	hsc_p	hsc_b	hsc_s	degree_p	degree_t	workex	etest_p	specialisation	mba_p	sta
0	1	M	67.00	Others	91.00	Others	Commerce	58.00	Sci&Tech	No	55.0	Mkt&HR	58.80	Pla
1	2	M	79.33	Central	78.33	Others	Science	77.48	Sci&Tech	Yes	86.5	Mkt&Fin	66.28	Pla
2	3	M	65.00	Central	68.00	Central	Arts	64.00	Comm&Mgmt	No	75.0	Mkt&Fin	57.80	Pla
3	4	M	56.00	Central	52.00	Central	Science	52.00	Sci&Tech	No	66.0	Mkt&HR	59.43	Pla
4	5	M	85.80	Central	73.60	Central	Commerce	73.30	Comm&Mgmt	No	96.8	Mkt&Fin	55.50	Pla
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
210	211	M	80.60	Others	82.00	Others	Commerce	77.60	Comm&Mgmt	No	91.0	Mkt&Fin	74.49	Pla
211	212	M	58.00	Others	60.00	Others	Science	72.00	Sci&Tech	No	74.0	Mkt&Fin	53.62	Pla
212	213	M	67.00	Others	67.00	Others	Commerce	73.00	Comm&Mgmt	Yes	59.0	Mkt&Fin	69.72	Pla
213	214	F	74.00	Others	66.00	Others	Commerce	58.00	Comm&Mgmt	No	70.0	Mkt&HR	60.23	Pla
214	215	M	62.00	Central	58.00	Others	Science	53.00	Comm&Mgmt	No	89.0	Mkt&HR	60.22	Pla

Fig.No 5.2.1 : Implementing the Data Preprocessing campus Recruitment data set

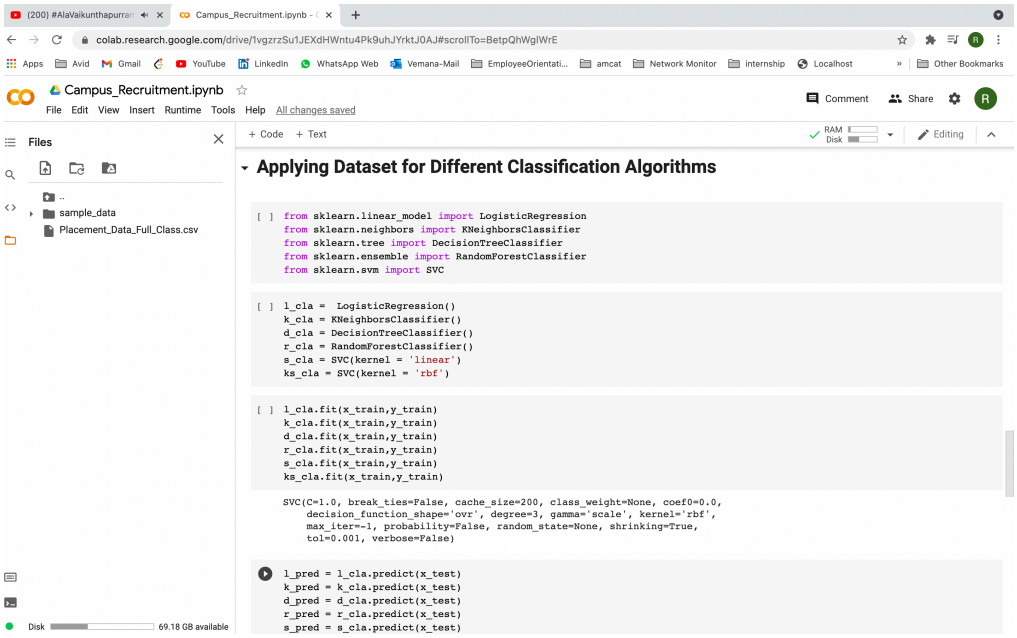


Fig.No 5.2.2 : Applying different classification algorithms for campus Recruitment dataset

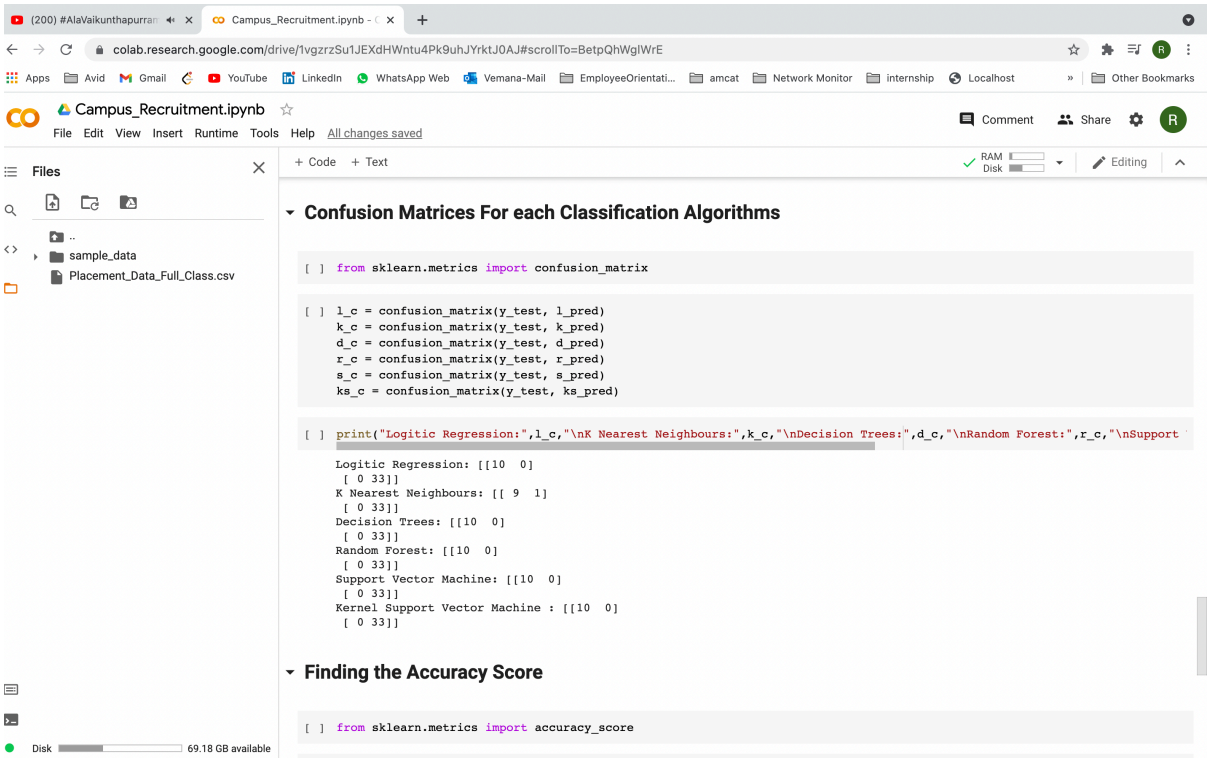


Fig.No 5.2.3 : Applying the confusion matrices for campus recruitment dataset

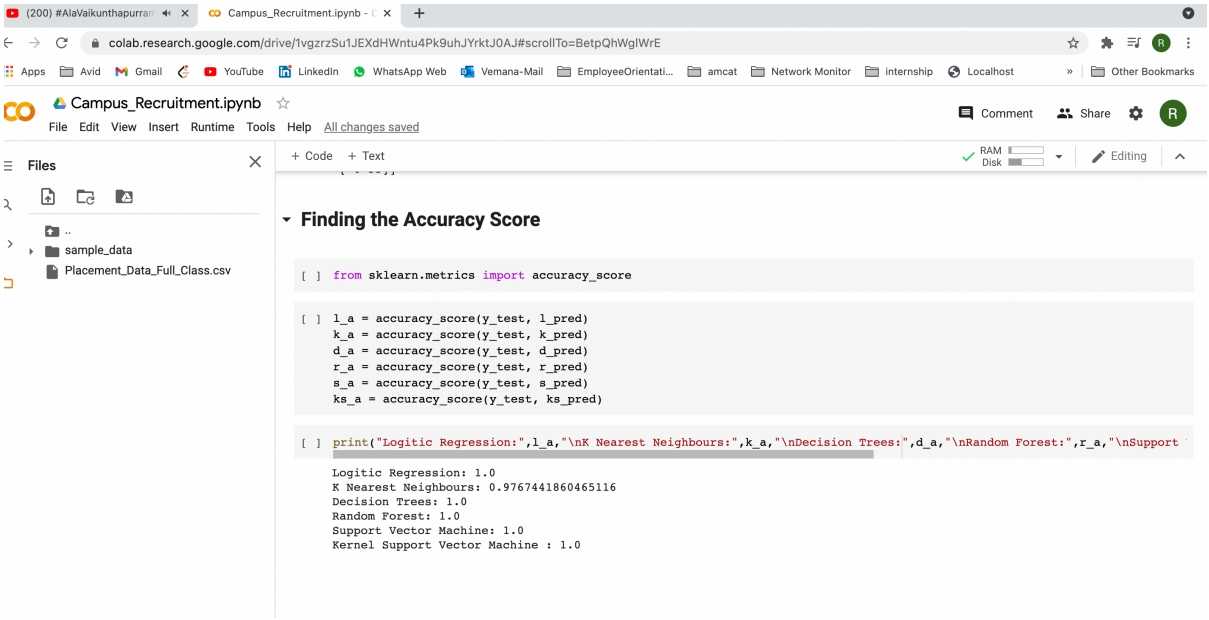


Fig.No 5.2.4 : Finding the Accuracy score for given Campus Recruitment dataset

CONCLUSION

Artificial Intelligence and Machine Learning are products of both science and myth. The idea that machines could think and perform tasks just as humans do is thousands of years old. The cognitive truths expressed in AI and Machine Learning systems are not new either.

Machine learning is quickly growing field in computer science. It has applications in nearly every other field of study and is already being implemented commercially because machine earning can solve problems too difficult or time consuming for humans to solve As a result, we have studied the future of Machine Learning. Also, study algorithms of machine learning. Along with we have studied its application which will help you to deal with real life.

Common methods and popular approaches used in the field, suitable machine learning programming languages, and also covered some things to keep in mind in terms of unconscious biases being replicated in algorithms.

REFERENCES

1. Jure Leskovec, TA: Keith Sillats HW 4
2. R.Wilson and R.Sharda , “Bankruptcy prediction using neural networks” , Decision Support Systems.
3. Neelama Budhani, Dr.C.K.Jha, Sandeep K. Budhani “Application Of Neural Network In Analysis Of Stock Market Prediction”, International Journal Of Computer science And Engineering
4. B. Manjula, S.S.V.N. Sharma, R. Lakshman Naik, G. Shruthi, Stock Prediction using Neural Network, International journal of advantage engineering sciences and technologies.