A

SUMMER INTERNSHIP REPORT

at

Zujo Tech Pvt. Ltd.

on

Pregnant Women Diabetes Prediction

For partial fulfillment towards

Summer Internship (3170001)

By

Rana Akash Rakeshkumar (180840131055)

Under the Guidance

of

Mr. Dhaval J. Rana

Department of Computer Science & Engineering

R. N. G. Patel Institute of Technology, Isroli-Bardoli

Gujarat Technological University, Ahmedabad

June 2021

R. N. G. Patel Institute of Technology, Isroli-Bardoli

Department of Computer Science & Engineering

Academic Year 2021-22

CERTIFICATE

This is to certify that **Rana Akash Rakeshkumar (180840131055)** of Computer Science & Engineering has submitted SUMMER INTERNSHIP project report entitled **"Pregnant Women Diabetes Prediction"** in partial fulfillment of requirement for the completion of SUMMER INTERNSHIP as prescribed by Gujarat Technological University, Ahmedabad during the academic year 2021-22. It has been found to be satisfactory and hereby approved for the submission. He has work under my guidance.

Date: / / 2021

Mr. Dhaval J. Rana
(Assistant Professor)

Dr. M. B. Desai
(Head of Department)

Signature of Examiner

Certificate of Completion

About the Company

ZUJO
TECH PVT. LTD.

Vision, Mission, and Values of Zujo Tech Pvt. Ltd.

Zujo is an Ed-tech company that aims to strengthen students' core knowledge from a grassroots level. We are a company based in the City of Diamonds - Surat, delivering our products worldwide. We are constructed on the idea of providing skillful education to the learners, considering the demands of the industry.

Our mission

- To provide learners with all the skills they must have before entering the industry.
- To create a community of learners and educators for a better future

Our vision

- To promote an ethical educational environment as a mandate for living
- To construct leadership in the younger minds

Zujo.co is a platform that works on cutting-edge technologies and frames a better work environment in the technology industry. Technically, it is one of the best providers of real-time communication and advancements. We possess a strict research and development team who looks after developing new technologies every time making them useful for all.

On Zujo.co, dedicated teachers for each skill are within easy reach who make each student brighten up & embellish their goal making. Everything at Zujo is homegrown. We have crafted all the resources on our own to make learning more significant and respectful. Our development in video conferencing, APIs and SDKs, live streaming, and much more makes us put our heart and soul towards the skillful blooming of a student.

Zujo.co is a leading learning platform opted by students, making their educational background stronger and optimistic. In the long run, Zujo delivers a world-class platform for students to achieve big. Zujo looks to cater to the best work for their learners in the technological industry, helping them reach the best companies for their chosen careers.

ZUJO
9th Floor, Infinity Tower, Near Ayurvedic College, Station Rd, Surat, Gujarat 395003
+91-261-288-2299
support@zujo.io
www.zujo.co

ACKNOWLEDGEMENTS

The satisfaction that accompanies the successful completion of this project would be incomplete without mentioning the people who made it possible, without whose constant guidance and encouragement would have made efforts go in vain. I consider myself privileged to express gratitude and respect towards all those who has guided through the completion of projects.

I extend my deepest gratitude to my mentor **Mr Dhaval J. Rana** for guiding, supporting and helping me in every possible way.

During my summer internship, My supervisor **Arjun Kava** from **Zujo Tech Pvt. Ltd.** Was very helpful and extended his/her valuable guidance and provide help whenever we required.

I am grateful to **Dr. Madhavi Desai**, Head of the Department, Computer Science & Engineering, RNGPIT for giving us the support and encouragement that was necessary for the completion of this project.

I would also like to express my deepest gratitude to **Dr. Latesh Chaudhari**, Principal of R. N. G. Patel Institute of Technology for providing us congenial environment to work in.

I would also like to thank Placement Cell of the department for giving me an opportunity to be the part of this internship. I extend my gratitude to all the faculty members for their understanding and guidance that gave me strength to work to long hours for developing a project and preparing the report.

Rana Akash Rakeshkumar
(180840131056)

TABLE OF CONTENTS

1:INTRODUCTION AND OVERVIEW

1.1 Abstract

Now a days the food and diet cycle of any person is so inappropriate. Due to this many long-life disease happened with that person. Diabetes is one of the disease that may end the human life. I generate a machine learning code for the diabetes specially for pregnant women. [1] This report presents a method for Pregnant Women Diabetes Prediction using a XgBoost and Random Forest Classifier. We Predict the Pregnant Women is Diabatic or not according to some parameters. We corelate those parameters with each other and according to its value we conclude that women is diabatic or not.

1.2 Introduction

Full Stack Machine Learning Internship

R. N. G. Patel Institute of Technology - RNGPIT, Bardoli Department of Computer Science & Engineering Summer Internship Training 2021 on “Full Stack Machine Learning” in association with Zujo Tech Pvt. Ltd., Surat from 31/05/2021 to 12/06/2021

What we learn
Introduction to Machine Learning
Supervised Machine Learning Algorithm
- Linear Regression
- Gradient Descent

Supervised Machine Learning Algorithm
- Classification Algorithm
- Support Vector Machine

Artificial Neural Network (ANN):
- Introduction
- Mathematics
- Backpropagation
- SGD

Convolution Neural Network (CNN):
- Introduction
- Mathematics
- Adam and RMS prop

Natural Language Processing
- RNN
- LSTM
- Encoder / Decoder

Introduction to transformers
- Introduction
- BERT,

Introduction to flask
- Setup
- Create API
- Inference a model

1.3 Problem statement

In the generation of Fast-food and inappropriate routine , a Human life suffer from many long term disease. Those disease are much dangerous that if we do not start proper treatment on time then it take a life of human. Aside of these problem many people are dying in village due to this disease because of lack of knowledge , less medical equipment , less hospital , time consuming reports and many more reason. Those disease are different types of cancer, asthma , Heart disease , diabetes and many more [1]. Diabetes is major disease that it occurs when pancreas doesnot produce enough of the hormone insulin.

It is also major problem for Pregnant women. Gestational diabetes occurs during pregnancy but may resolve after the baby is delivered [2]. Having a family history of diabetes makes it more likely that a woman will develop gestational diabetes, which suggests that genes play a role [6]. But in many cases due to not proper treatment ontime both baby and women may die. The main reason is many reports are too much time consuming and due to this we will not start and treatment on time.

2:TRAINING ACTIVITIES

2.1 Machine Learning : Supervised ML Algorithm – Linear Regression

Linear Regression in Machine Learning:

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

The linear regression model provides a sloped straight line representing the relationship between the variables.

Mathematically, we can represent a linear regression as:

$y= a_0+a_1x+ \varepsilon$

Here,

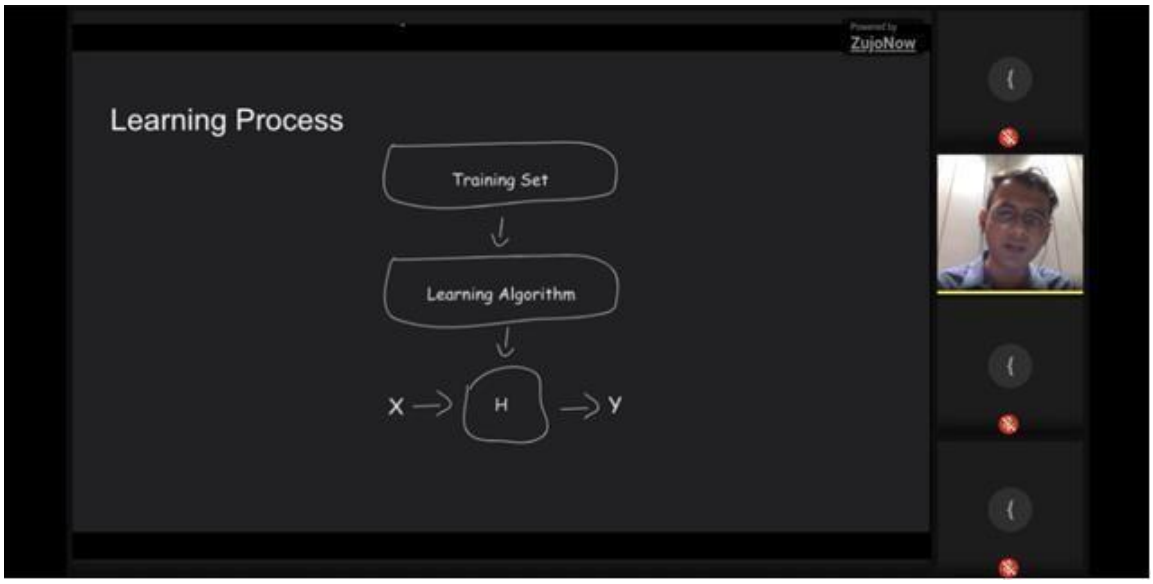
Y= Dependent Variable (Target Variable)

X= Independent Variable (predictor Variable)

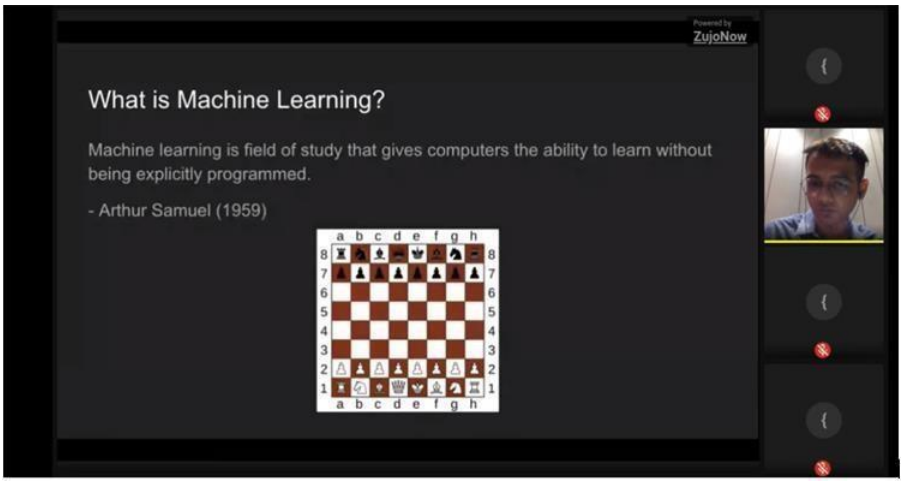
a0= intercept of the line (Gives an additional degree of freedom)

a1 = Linear regression coefficient (scale factor to each input value).

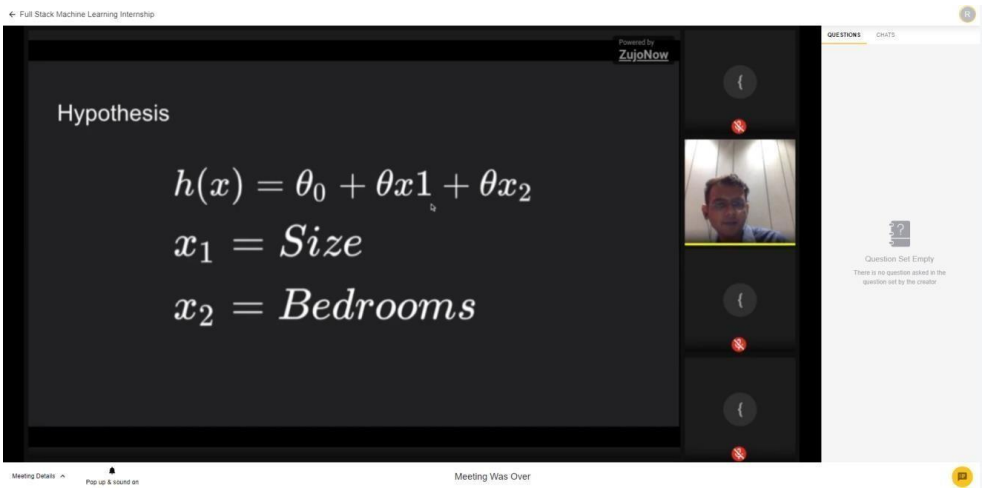
ε = random error



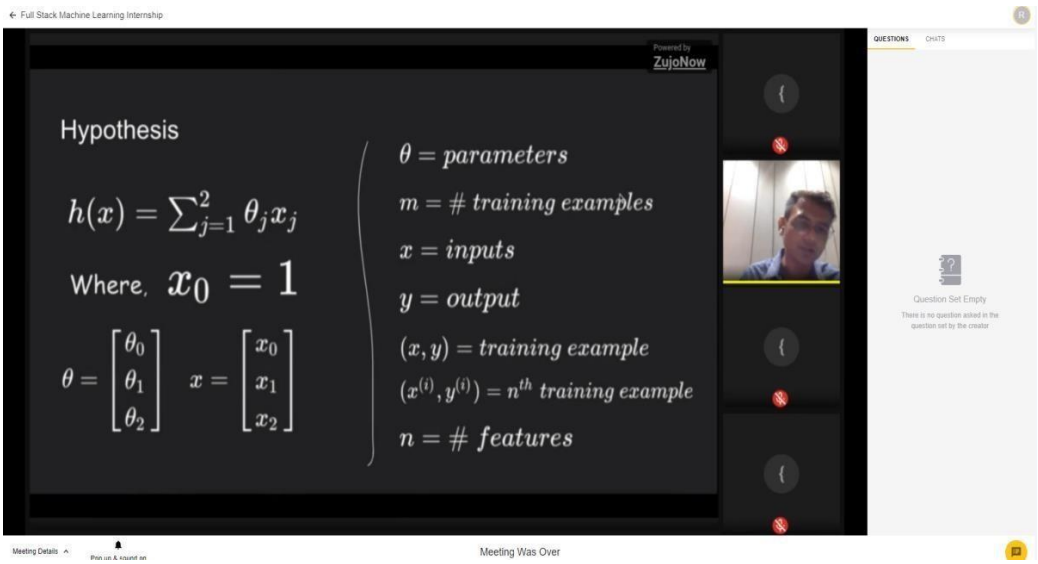
(Fig:-2.1.1):-Machine Learning process



(Fig:-2.1.2):-Definition of Machine learning



(Fig:-2.1.3(a)):-Hypothesis Equation



(Fig:-2.1.3(b)):-Derivation of Hypothesis

2.2 Supervised ML Algorithm – Support Vector Machine

Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and outliers detection.

-The advantages of support vector machines are:

Effective in high dimensional spaces.

Still effective in cases where number of dimensions is greater than the number of samples.

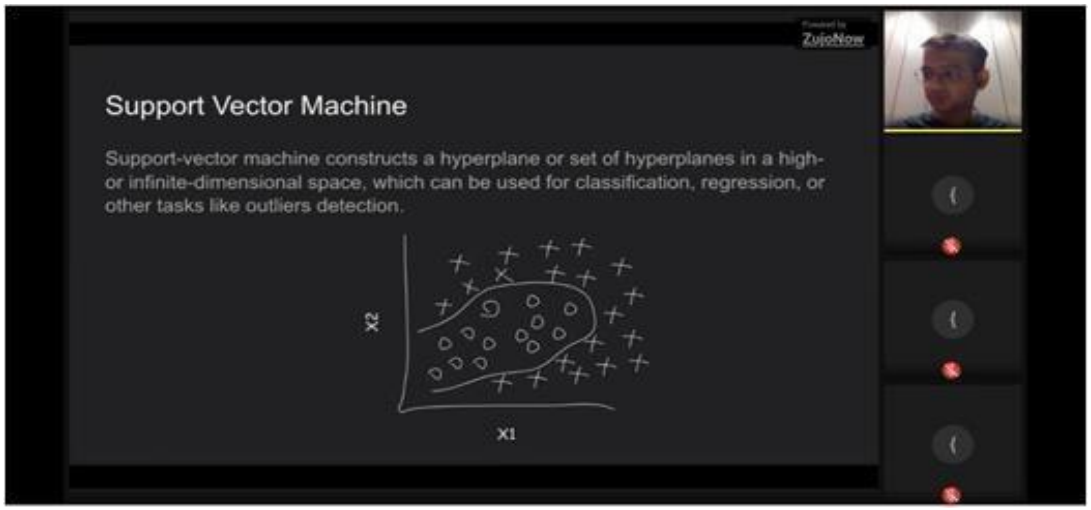
Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.

Versatile: different Kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

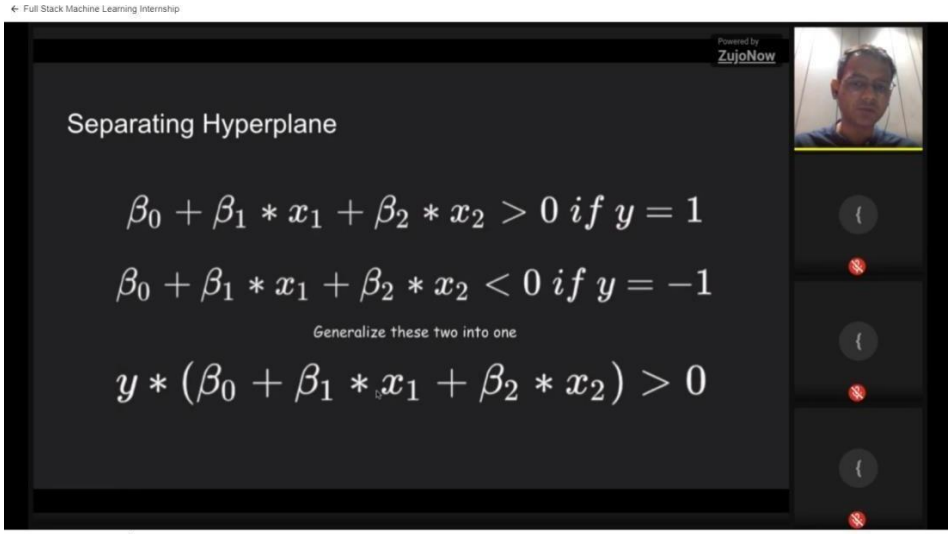
-The disadvantages of support vector machines include:

If the number of features is much greater than the number of samples, avoid over-fitting in choosing Kernel functions and regularization term is crucial.

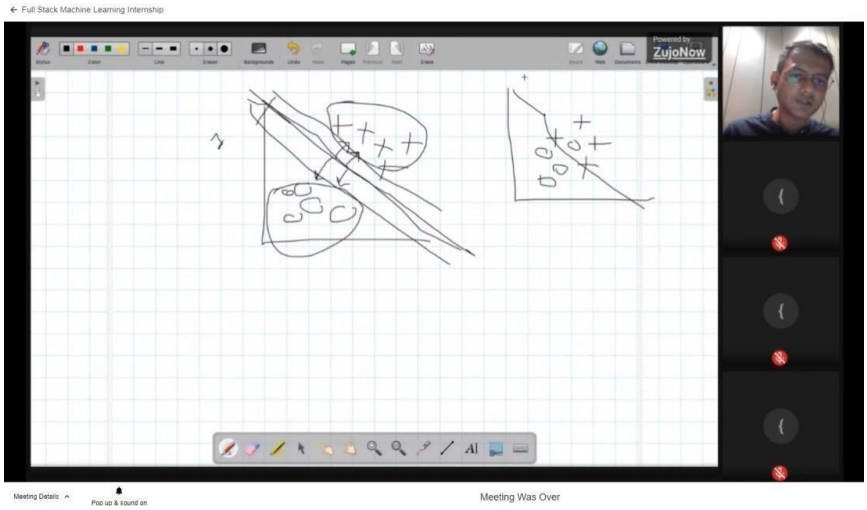
SVMs do not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation (see Scores and probabilities, below).



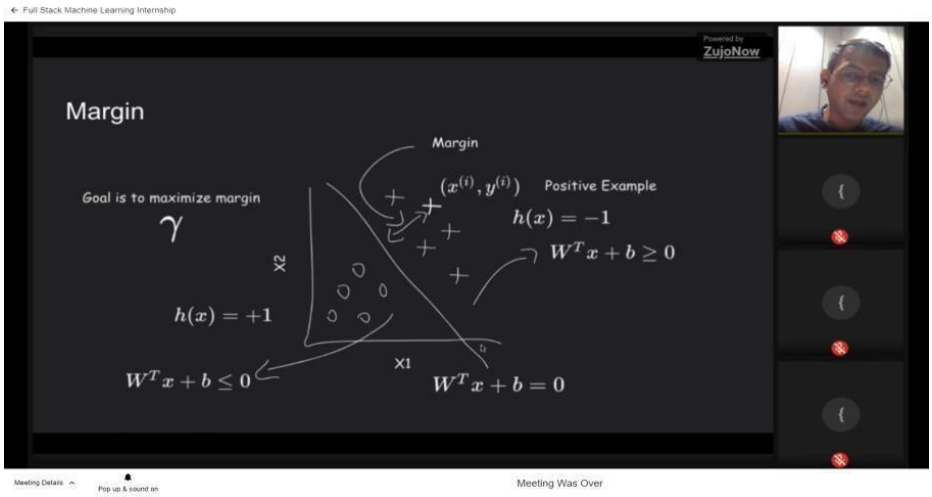
(Fig:-2.2.1):-Definition of SVM



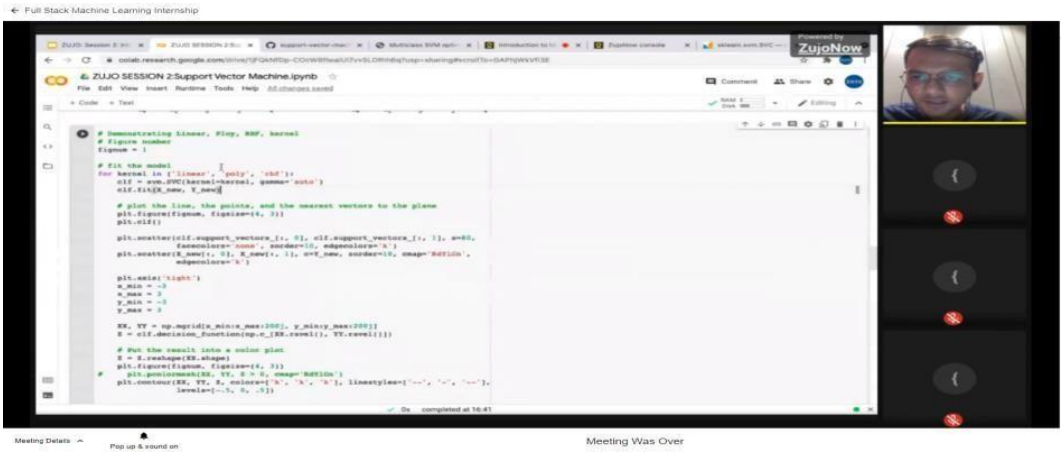
(Fig:-2.2.2):-Separating Hypothesis theorem



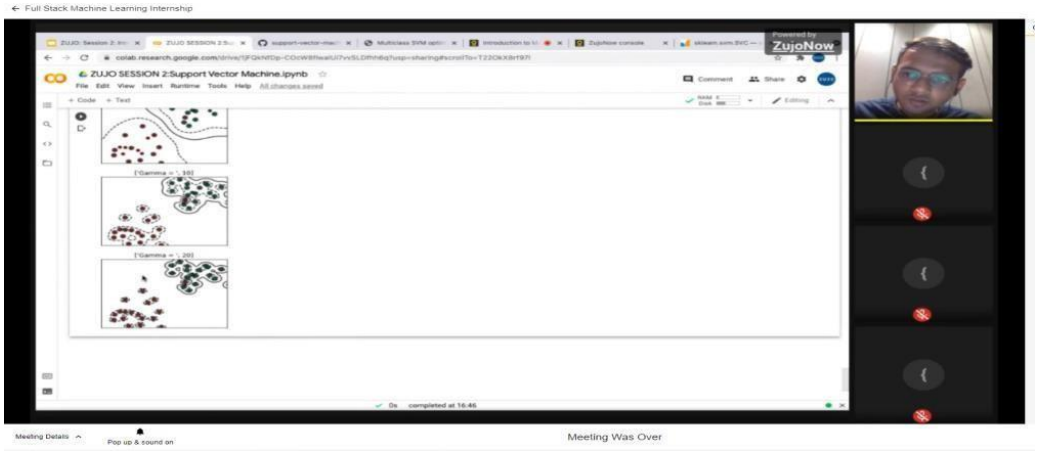
(Fig:-2.2.3):-explaination



(Fig:-2.2.4):-Margin explaination



(Fig:-2.2.5(a)):-Practical Session



(Fig:-2.2.5(b)):-Practical Session

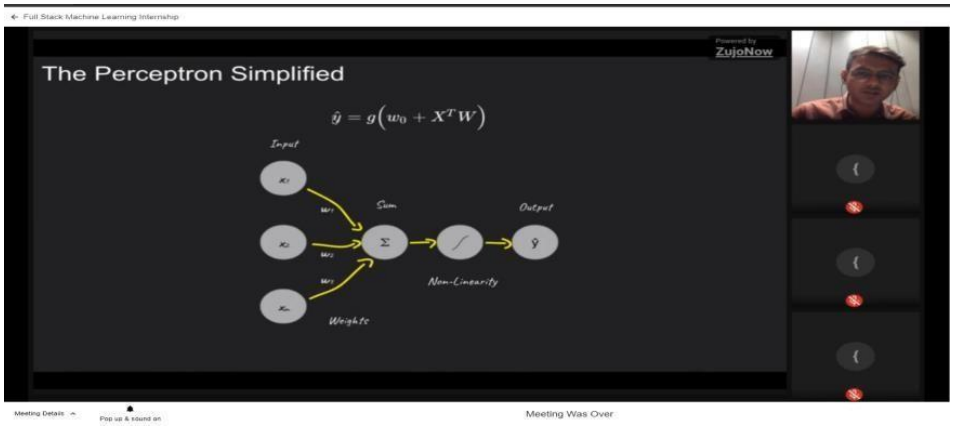
2.3 Artificial Neural Network

An artificial neural network (ANN) is the piece of a computing system designed to simulate the way the human brain analyzes and processes information. It is the foundation of AI and solves problems that would prove impossible or difficult by human or statistical standards. ANNs have self-learning capabilities that enable them to produce better results as more data becomes available.

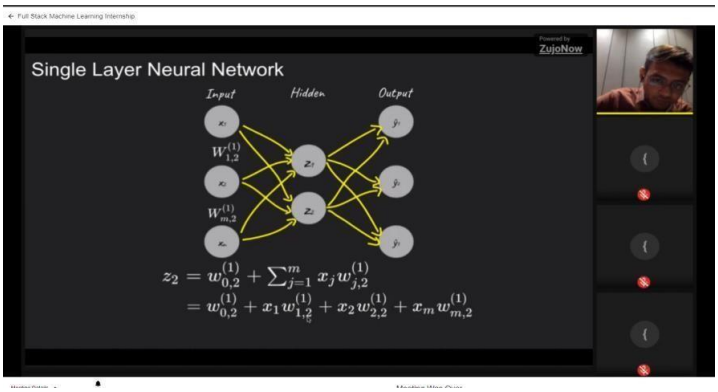
Artificial neural networks are built like the human brain, with neuron nodes interconnected like a web. The human brain has hundreds of billions of cells called neurons. Each neuron is made up of a cell body that is responsible for processing information by carrying information towards (inputs) and away (outputs) from the brain.

An ANN has hundreds or thousands of artificial neurons called processing units, which are interconnected by nodes. These processing units are made up of input and output units. The input units receive various forms and structures of information based on an internal weighting system, and the neural network attempts to learn about the information presented to produce one output report. Just like humans need rules and guidelines to come up with a result or output, ANNs also use a set of learning rules called backpropagation, an abbreviation for backward propagation of error, to perfect their output results.

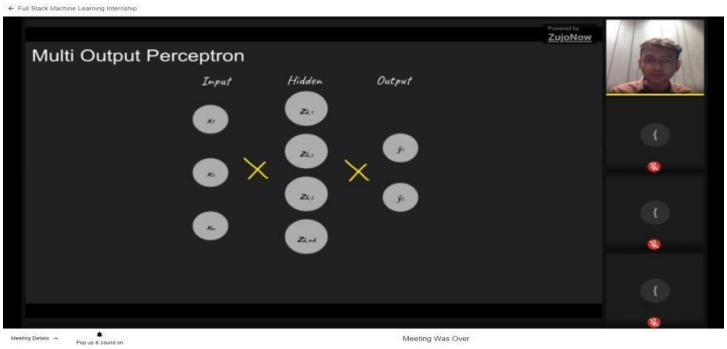
An ANN initially goes through a training phase where it learns to recognize patterns in data, whether visually, aurally, or textually. During this supervised phase, the network compares its actual output produced with what it was meant to produce—the desired output. The difference between both outcomes is adjusted using backpropagation. This means that the network works backward, going from the output unit to the input units to adjust the weight of its connections between the units until the difference between the actual and desired outcome produces the lowest possible error.



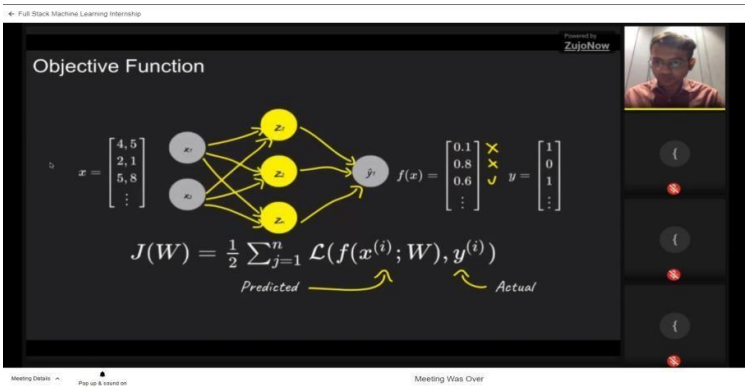
(Fig:-2.3.1):-Simplify perceptron



(Fig:-2.3.2):-Single layer neural network

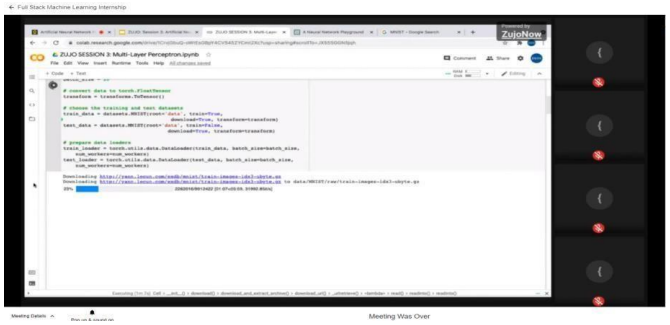


(Fig:-2.3.3):-Multi Output Percepton



(Fig:-2.3.4):-Explain the Objective function

Hands on session

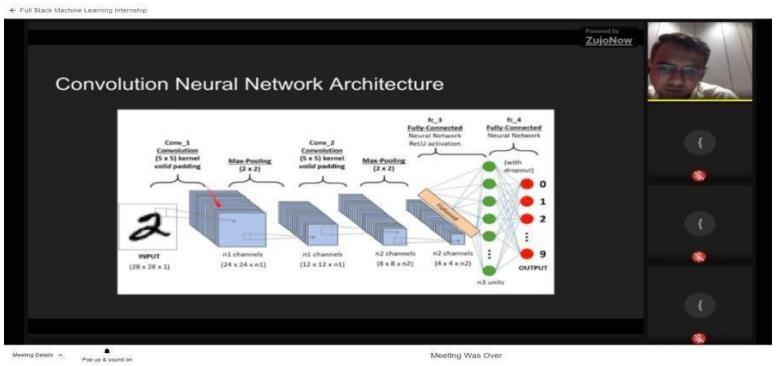


(Fig:-2.3.5)Practical session

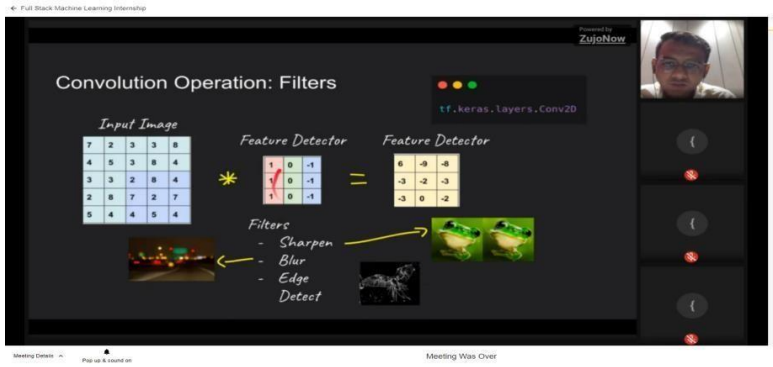
2.4 Convolution Neural Network

In the past few decades, Deep Learning has proved to be a very powerful tool because of its ability to handle large amounts of data. The interest to use hidden layers has surpassed traditional techniques, especially in pattern recognition. One of the most popular deep neural networks is Convolutional Neural Networks.

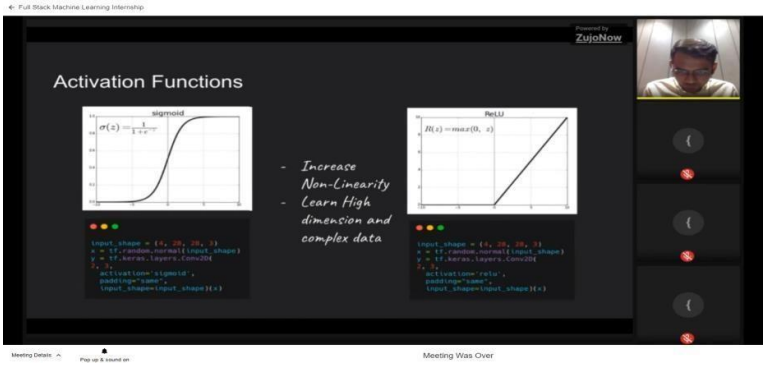
Convolutional neural networks are composed of multiple layers of artificial neurons. Artificial neurons, a rough imitation of their biological counterparts, are mathematical functions that calculate the weighted sum of multiple inputs and outputs an activation value. When you input an image in a ConvNet, each layer generates several activation functions that are passed on to the next layer.



(Fig:-2.4.1):-Architecture of CNN

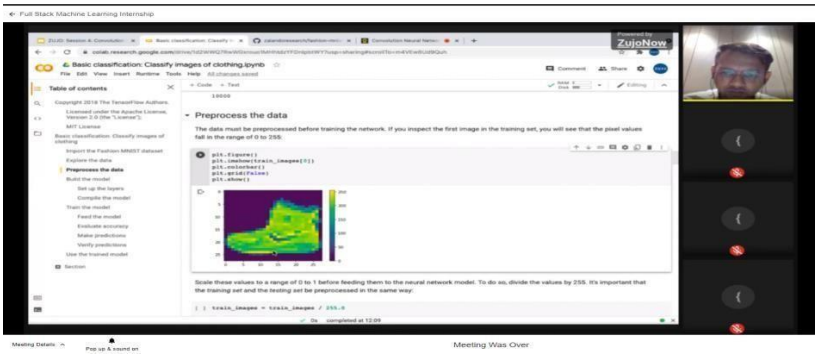


(Fig:-2.4.2):-Convolution operation

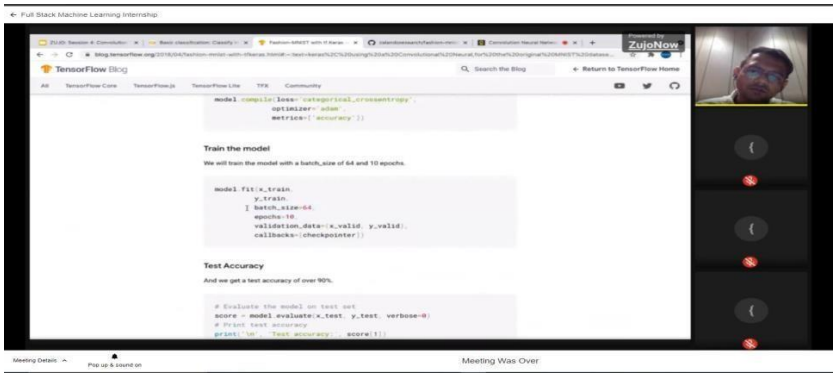


(Fig:-2.4.3):-Explanation of Activation Function

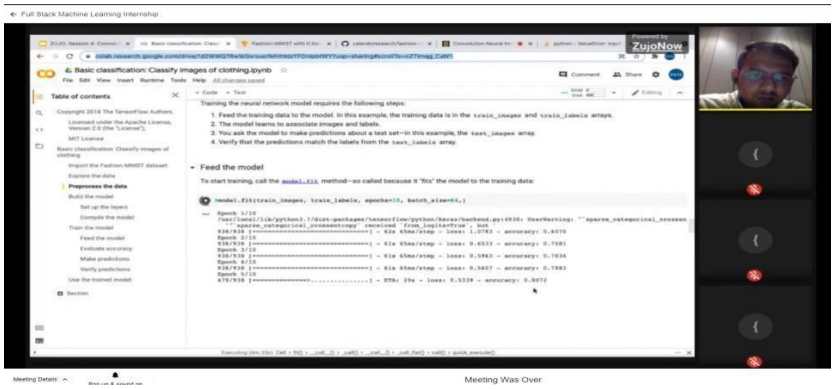
Hands on session



(Fig:-2.4.4(a)):-Practical session 1



(Fig:-2.4.4(b)):-Practical session 2



(Fig:-2.4.4(c)):-Practical session 3

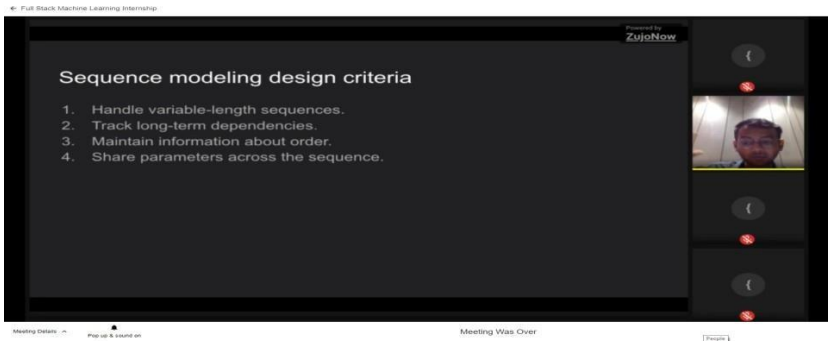
2.5 Natural Language Processing

Natural language processing (NLP) refers to the branch of computer science—and more specifically, the branch of artificial intelligence or AI—concerned with giving computers the ability to understand text and spoken words in much the same way human beings can.

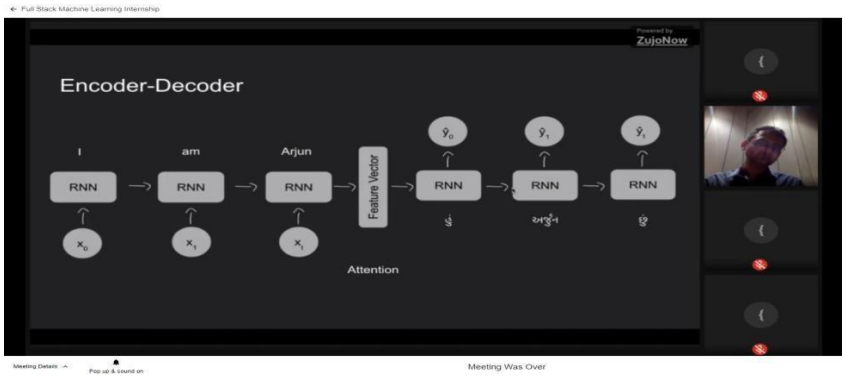
NLP combines computational linguistics—rule-based modeling of human language—with statistical, machine learning, and deep learning models. Together, these technologies enable computers to process human language in the form of text or voice data and to 'understand' its full meaning, complete with the speaker or writer's intent and sentiment.

NLP drives computer programs that translate text from one language to another, respond to spoken commands, and summarize large volumes of text rapidly—even in real time. There's a good chance you've interacted with NLP in the form of voice-operated GPS systems, digital assistants, speech-to-text dictation software, customer service chatbots, and other consumer conveniences. But NLP also plays a growing role in enterprise solutions that help streamline business operations, increase employee productivity, and simplify mission-critical business processes.

(Fig:-2.5.1):-Introduction of sequence

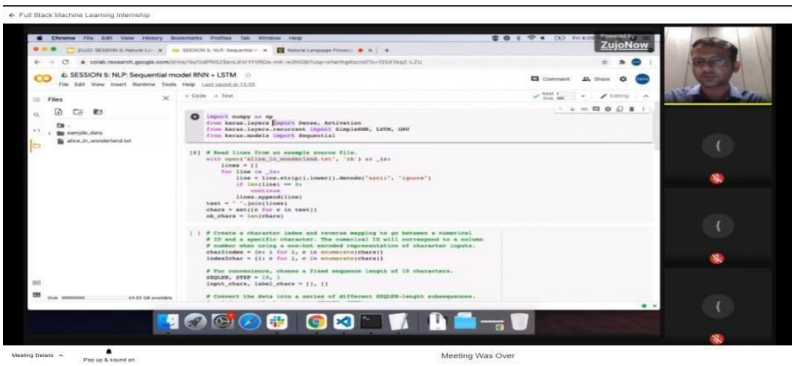


(Fig:-2.5.2):-Sequence modeling design

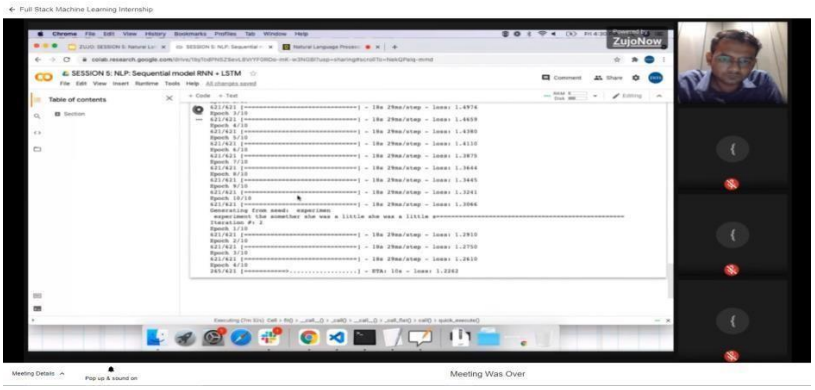


(Fig:-2.5.3):-Encoder-Decoder process

Hands on session



(Fig:-2.5.4(a)):-Practical session 1



(Fig:-2.5.4(b)):-Practical session 2

2.6 Flask : Setup, Create API & Inference a model

Flask is a web application framework written in Python. It has multiple modules that make it easier for a web developer to write applications without having to worry about the details like protocol management, thread management, etc.

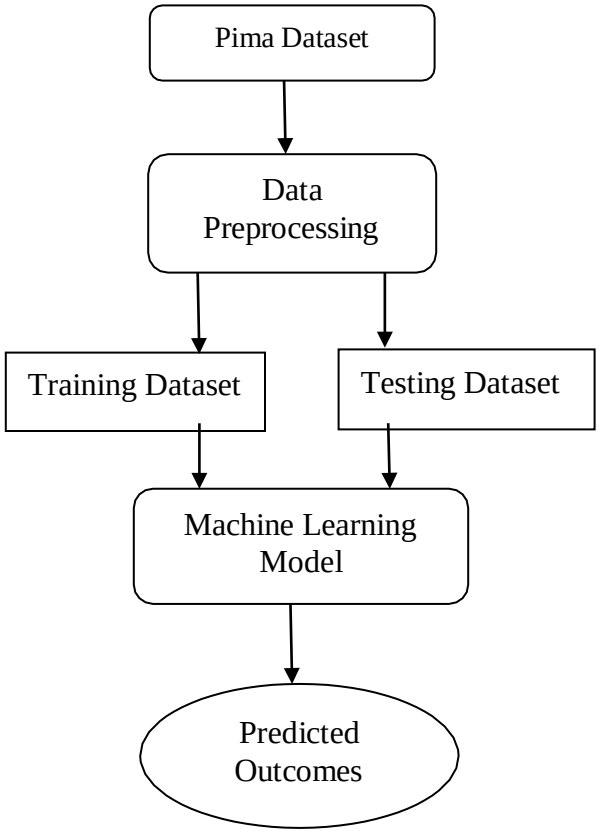
Flask gives is a variety of choices for developing web applications and it gives us the necessary tools and libraries that allow us to build a web application

Installing Flask on your Machine
Installing Flask is simple and straightforward. Here, I am assuming you already have Python 3 and pip installed. To install Flask, you need to run the following command:

sudo apt-get install python3-flask

3: PROJECT DETAILS

3.1 Project Design Diagram – Proposed Solution



3.2 Implementation Tools

- Numpy
- Pandas
- Seaborn
- matplotlib.pyplot
- XgBoost
- Random Forest

3.3 Functionalities of Project and Snapshot

- **Functionalities**
  Pregnant Women Diabetes Prediction is predict the Diabetes level of Women according to diagonist Parameters. It also show the value of diabetes count in the body and give the result. [3]

- **Working of Model**

```
[ ]  import pandas as pd
     import matplotlib.pyplot as plt
     import numpy as np

     %matplotlib inline
```

(Fig:-3.3.1):-Import Require Libraries

```
data = pd.read_csv("pima-data.csv")
```

```
[ ]  data.shape
```

```
(768, 10)
```

```
[ ]  data.head(5)
```

	num_preg	glucose_conc	diastolic_bp	thickness	insulin	bmi	diab_pred	age	skin	diabetes
0	6	148	72	35	0	33.6	0.627	50	1.3790	True
1	1	85	66	29	0	26.6	0.351	31	1.1426	False
2	8	183	64	0	0	23.3	0.672	32	0.0000	True
3	1	89	66	23	94	28.1	0.167	21	0.9062	False
4	0	137	40	35	168	43.1	2.288	33	1.3790	True

(Fig:-3.3.2):-Import The Dataset

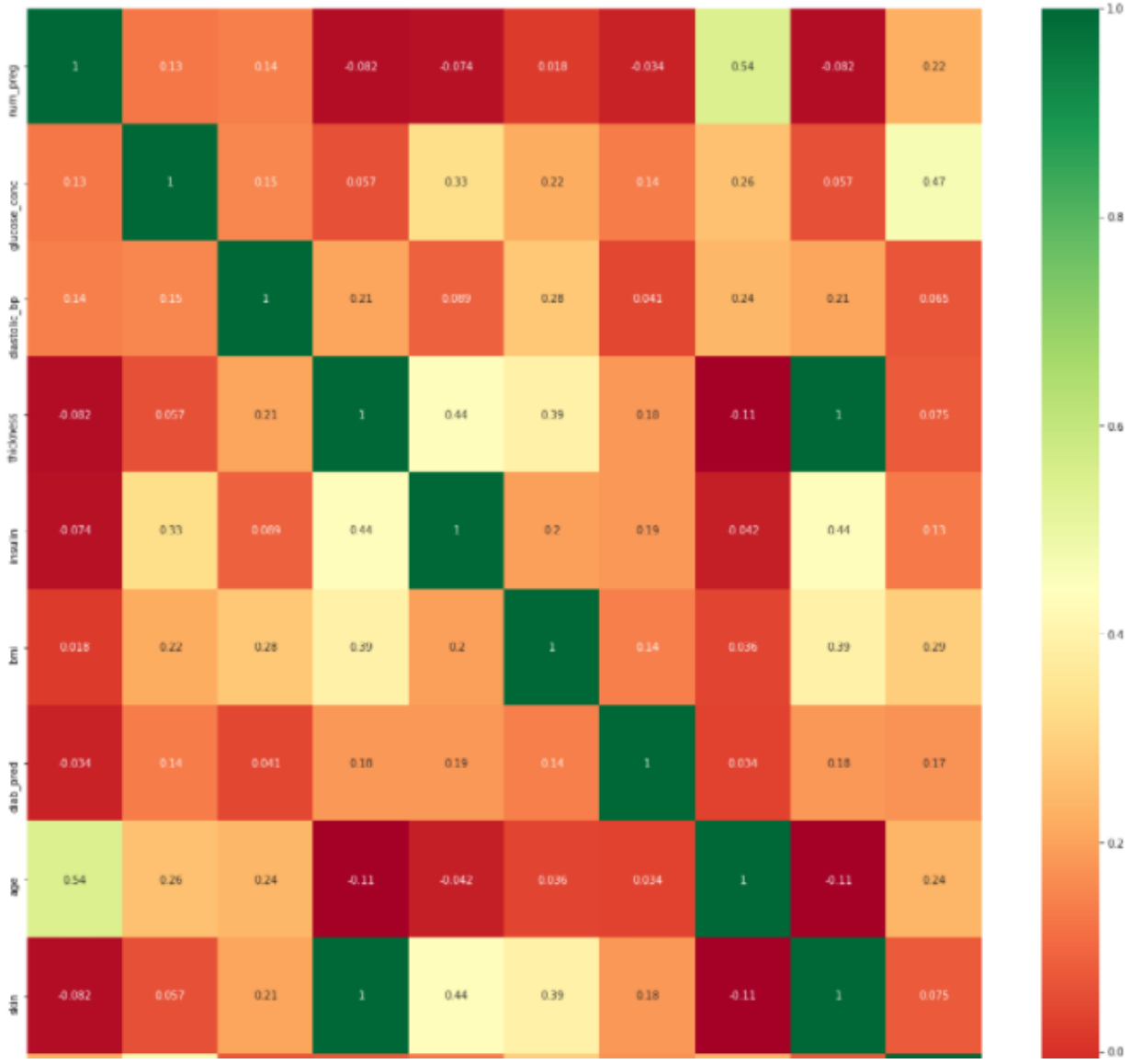
```
[ ]  # check if any null value is present
     data.isnull().values.any()
```

```
False
```

(Fig:-3.3.3):-Checking the null values

```
[ ]  ## Correlation
     import seaborn as sns
     import matplotlib.pyplot as plt
     #get correlations of each features in dataset
     corrmat = data.corr()
     top_corr_features = corrmat.index
     plt.figure(figsize=(20,20))
     #plot heat map
     g=sns.heatmap(data[top_corr_features].corr(),annot=True,cmap="RdYlGn")
```

(Fig:-3.3.4):-Splitting the Data



(Fig:-3.3.5):-Correlation between Diagonist parameter

- Apply RandomForest Algorithm:

Random forest is a commonly-used machine learning algorithm trademarked by Leo Breiman and Adele Cutler, which combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fueled its adoption, as it handles both classification and regression problems. [4]

```
[ ]  ## Apply Algorithm

     from sklearn.ensemble import RandomForestClassifier
     random_forest_model = RandomForestClassifier(random_state=10)

     random_forest_model.fit(X_train, y_train.ravel())
```

```
     RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None,
                            criterion='gini', max_depth=None, max_features='auto',
                            max_leaf_nodes=None, max_samples=None,
                            min_impurity_decrease=0.0, min_impurity_split=None,
                            min_samples_leaf=1, min_samples_split=2,
                            min_weight_fraction_leaf=0.0, n_estimators=100,
                            n_jobs=None, oob_score=False, random_state=10, verbose=0,
                            warm_start=False)
```

```
[ ]  predict_train_data = random_forest_model.predict(X_test)

     from sklearn import metrics

     print("Accuracy = {0:.3f}".format(metrics.accuracy_score(y_test, predict_train_data)))
```

```
     Accuracy = 0.736
```

(Fig:-3.3.6):-Applying Random Forest Algorithm

A

- Apply the XgBoost Algorithm

XGBoost is an algorithm that has recently been dominating applied machine learning and Kaggle competitions for structured or tabular data. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance. [4]

```
[ ]  ## Hyperparameter optimization using RandomizedSearchCV
     from sklearn.model_selection import RandomizedSearchCV
     import xgboost
```

```
[ ]  classifier=xgboost.XGBClassifier()
```

```
[ ]  random_search=RandomizedSearchCV(classifier,param_distributions=params,n_iter=5,scoring='roc_auc',n_jobs=-1,cv=5,verbose=3)
```

```
[ ]  def timer(start_time=None):
         if not start_time:
             start_time = datetime.now()
             return start_time
         elif start_time:
             thour, temp_sec = divmod((datetime.now() - start_time).total_seconds(), 3600)
             tmin, tsec = divmod(temp_sec, 60)
             print('\n Time taken: %i hours %i minutes and %s seconds.' % (thour, tmin, round(tsec, 2)))
```

```
[ ]  from datetime import datetime
     # Here we go
     start_time = timer(None) # timing starts from this point for "start_time" variable
     random_search.fit(X,y.ravel())
     timer(start_time) # timing ends here for "start_time" variable
```

```
     Fitting 5 folds for each of 5 candidates, totalling 25 fits
     [Parallel(n_jobs=-1)]: Using backend LokyBackend with 2 concurrent workers.

      Time taken: 0 hours 0 minutes and 2.53 seconds.
     [Parallel(n_jobs=-1)]: Done  25 out of  25 | elapsed:    2.4s finished
```

(Fig:-3.3.7(a)):-Applying XGBoost algorithm

```
[ ] random_search.best_estimator_

    XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1,
                  colsample_bynode=1, colsample_bytree=0.3, gamma=0.1,
                  learning_rate=0.05, max_delta_step=0, max_depth=10,
                  min_child_weight=1, missing=None, n_estimators=100, n_jobs=1,
                  nthread=None, objective='binary:logistic', random_state=0,
                  reg_alpha=0, reg_lambda=1, scale_pos_weight=1, seed=None,
                  silent=None, subsample=1, verbosity=1)

[ ] classifier=xgboost.XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1,
           colsample_bytree=0.3, gamma=0.0, learning_rate=0.25,
           max_delta_step=0, max_depth=3, min_child_weight=7, missing=None,
           n_estimators=100, n_jobs=1, nthread=None,
           objective='binary:logistic', random_state=0, reg_alpha=0,
           reg_lambda=1, scale_pos_weight=1, seed=None, silent=True,
           subsample=1)

[ ] from sklearn.model_selection import cross_val_score
    score=cross_val_score(classifier,X,y.ravel(),cv=10)

[ ] score

    array([0.7012987 , 0.77922078, 0.67532468, 0.67532468, 0.7012987 ,
           0.76623377, 0.75324675, 0.81818182, 0.75      , 0.81578947])

    score.mean()

    0.7435919343814081
```

(Fig:-3.3.7(b)):-Applying XGBoost Algorithm

3.4 Future Work

However, in reality, the patient must predict either to be in diabetic category or non-diabetic category. Such errors in diagnosis may lead to unnecessary treatments or no treatments at all when required. In order to avoid or reduce severity of such impact, there is a need to create a system using machine learning algorithm which will provide accurate results and reduce human efforts. Being able to made early Prediction of diabetes and start treatment of it before late.

3.5 Summary

The main aim of this project was to design and implement Diabetes Prediction Using Machine Learning Methods and Performance Analysis of that methods and it has been achieved successfully. [4] The proposed approach uses various classification and ensemble learning method in which Random Forest, Decision Tree, Xgboost are used. And 74% classification accuracy has been achieved. The Experimental results can be asst health care to take early prediction and make early decision to cure diabetes and save humans life.

4: LEARNING FROM THE INTERNSHIP PROGRAM

What we learn
Introduction to Machine Learning
Supervised Machine Learning Algorithm
- Linear Regression
- Gradient Descent

Supervised Machine Learning Algorithm
- Classification Algorithm
- Support Vector Machine

Artificial Neural Network (ANN):
- Introduction
- Mathematics
- Backpropagation
- SGD

Convolution Neural Network (CNN):
- Introduction
- Mathematics
- Adam and RMS prop

Natural Language Processing
- RNN
- LSTM
- Encoder / Decoder

Introduction to transformers
- Introduction
- BERT,

Introduction to flask
- Setup
- Create API
- Inference a model

5: REFERENCES

[1] "Center Of Disease control and Preventation.," [Online]. Available: https://www.cdc.gov/diabetes/basics/symptoms.html.

[2] "Mayo Clinic," [Online]. Available: https://www.mayoclinic.org/diseases-conditions/diabetes/symptoms-causes/syc-20371444.

[3] P. Institute, "Pima Indian Diabetes Database," 2017. [Online]. Available: https://www.kaggle.com/uciml/pima-indians-diabetes-database.

[4] J. Brownley, "Machine Learning Mastery," 2016. [Online]. Available: https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning.

[5] "BMC Endocrine Disorder," BMC: Part Of Springer Nature, 2019. [Online]. Available: https://bmcendocrdisord.biomedcentral.com/articles/10.1186/s12902-019-0436-6.

[6] B. Metzger, "National Institute of Diabetes and Digestive and Kidney Disease," 2017. [Online]. Available: https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes/gestational/symptoms-causes.