SYNOPSIS

ON

“DATA ANALYZER”

Submitted in

Partial Fulfillment of requirements for the Award of Degree

*of*

Bachelor of Technology

*In*

Computer Science and Engineering

Artificial Intelligence

By

**(Project Id:22\_AI\_2A\_15 )**

# Aakarshit Srivastava (2101641520001)

# Dev Gupta(2101641520046)

# Shashak Raj Singh(2101641520128)

Under the supervision of

**Dr. S.S.Parihar**

**Professor**



**Pranveer Singh Institute of Technology**.

 Kanpur - Agra - Delhi National Highway - 19

Bhauti -Kanpur - 209305.

(Affiliated to Dr. A.P.J. Abdul Kalam Technical University)

1. Introduction

Modern era of technology is moving forward towards high end components and devices,these devices generate a ton of data per user.This data can be collected, processed and cleaned to be utilized by certain architectures that can benefit a lot of people and industries.The data can be further analysed by deep learning and machine learning algorithms to come out with patterns and logics that can accurately predict outcomes and improve the operational efficiency further,creates a drive for innovations and maximizes profits.The Project works on the dataset analysis and training ai models as per the given dataset to provide output that the user can utilize for his gains and overcome situations.Data is a very important resource and without data you are just another person with opinions

“In god we trust,all other must bring data”-W.Edwards Deming

**Data analysis** is defined as a process of cleaning, transforming, and modeling data to discover useful information for business decision-making. The purpose of Data Analysis is to extract useful information from data and taking the decision based upon the data analysis. A simple example of Data analysis is whenever we take any decision in our day-to-day life is by thinking about what happened last time or what will happen by choosing that particular decision. This is nothing but analyzing our past or future and making decisions based on it. For that, we gather memories of our past or dreams of our future. So that is nothing but data analysis. Now same thing analyst does for business purposes, is called Data Analysis.

To grow your business even to grow in your life, sometimes all you need to do is Analysis.If your business is not growing, then you have to look back and acknowledge your mistakes and make a plan again without repeating those mistakes. And even if your business is growing, then you have to look forward to making the business to grow more. All you need to do is analyze your business data and business processes. Data analysis tools make it easier for users to process and manipulate data, analyze the relationships and correlations between data sets, and it also helps to identify patterns and trends for interpretation. Here is a complete list of tools used for data analysis in research. There are several **types of Data Analysis** techniques that exist based on business and technology. However, the major Data Analysis methods are:

* Text Analysis
* Statistical Analysis
* Diagnostic Analysis
* Predictive Analysis
* Prescriptive Analysis

Text Analysis is also referred to as Data Mining. It is one of the methods of data analysis to discover a pattern in large data sets using databases or data mining tools. It used to transform raw data into business information. Business Intelligence tools are present in the market which is used to take strategic business decisions. Overall it offers a way to extract and examine data and deriving patterns and finally interpretation of the data.

Statistical Analysis shows “What happen?” by using past data in the form of dashboards. Statistical Analysis includes collection, Analysis, interpretation, presentation, and modeling of data. It analyses a set of data or a sample of data. There are two categories of this type of Analysis – Descriptive Analysis and Inferential Analysis.

Descriptive Analysis, analyses complete data or a sample of summarized numerical data. It shows mean and deviation for continuous data whereas percentage and frequency for categorical data.Inferential Analysis, analyses sample from complete data. In this type of Analysis, you can find different conclusions from the same data by selecting different samples.

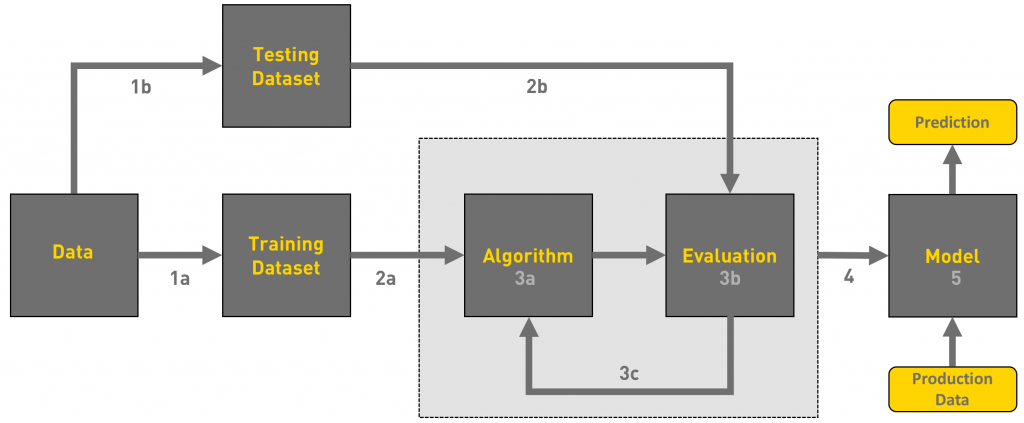
Diagnostic Analysis shows “Why did it happen?” by finding the cause from the insight found in Statistical Analysis. This Analysis is useful to identify behavior patterns of data. If a new problem arrives in your business process, then you can look into this Analysis to find similar patterns of that problem. And it may have chances to use similar prescriptions for the new problems.

Predictive Analysis shows “what is likely to happen” by using previous data. The simplest data analysis example is like if last year I bought two dresses based on my savings and if this year my salary is increasing double then I can buy four dresses. But of course it’s not easy like this because you have to think about other circumstances like chances of prices of clothes is increased this year or maybe instead of dresses you want to buy a new bike, or you need to buy a house.So here, this Analysis makes predictions about future outcomes based on current or past data. Forecasting is just an estimate. Its accuracy is based on how much detailed information you have and how much you dig in it.

Prescriptive Analysis combines the insight from all previous Analysis to determine which action to take in a current problem or decision. Most data-driven companies are utilizing Prescriptive Analysis because predictive and descriptive Analysis are not enough to improve data performance. Based on current situations and problems, they analyze the data and make decisions.

The **Data Analysis Process** is nothing but gathering information by using a proper application or tool which allows you to explore the data and find a pattern in it. Based on that information and data, you can make decisions, or you can get ultimate conclusions. Data Analysis consists of the following phases:

* Data Requirement Gathering
* Data Collection
* Data Cleaning
* Data Analysis
* Data Interpretation
* Data Visualization



2. Project Objective

* The end product will be a SaaS i.e. Software as a service. Plus after 2020 India is also in the phase of adapting to SaaS market(Projected to be valued at $50 billion by 2030, India's SaaS market has reached a critical inflection point as venture dollars deployed in the region reached $4.8 billion in 2021)
* A intuitive Webpage that will be accessed by the user for the inputs and the user satisfactory output will be given according to the trained algorithms.
* A graphical menu driven front end and model trained on datasets for backend.
* The Application will help the end user with staying updated with his requirements and work according without any extra effort and may even become a type of morning habit.
* The project can help many startups and businesses to grow and gain customers effectively.
* The milestone for the project will be to become a global phenomenon and gain as many users as possible from a 12 year old child to a 75 year old retired person.

This should give a clear picture of the project. Objectives should

be clearly specified. What the project ends up to and in what way

this is going to help the end user has to be mentioned.

1. Feasibility Study:

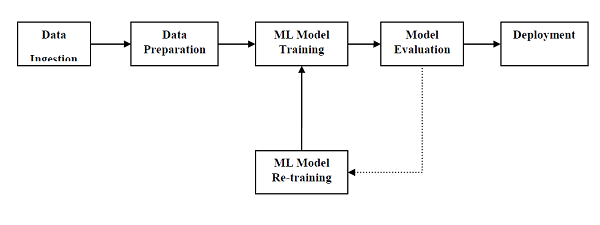
This project aims to utilize the data as a resource to generate predictions using the artificial engines and mechanisms. The project will provide solutions to many problems and enhance strategies and an overall information requirement of the user with satisfactory results.Data being the immediate need of the today’s world the product/service will be very optimal and will result in a very successful venture

Gantt Chart

Gantt chart version 2.1

* Economic Feasibilty: As the indian market will be growing in technology the demand and value of data will increase. The process is reliable and can be used for passive income.
* Technical Feasibilty: The service will be completely client server based with frequent need of updation and may need upgradation.
* Operational Feasibility: The service will be very effective in fulfilling user’s need of information and validation of information.
* Schedule Feasibilty: The service is very compatible and light to develop and optimization of the resources is also very simple and easy task.

4. Methodology/ Planning of work

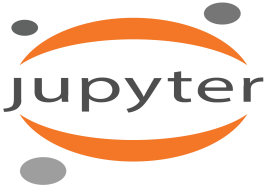


The project Data Analyzer is simply a machine learning based model that is being fed with upto date data and the trained algorithm is expected to predict outcomes, more the data input and regularization more will be the efficiency of the model. The trained model will then be connected and synced up with a web app builder called anvil that will be used to make a user interface or front end of the web service.

Keras is a high-level API for building and training deep learning models.TFKeras is TensorFlow’s implementation of this API. Keras is an open-source software library that provides a Python interface for artificial neural networks. Keras acts as an interface for the TensorFlow library. Up until version 2.3, Keras supported multiple backends, including TensorFlow, Microsoft Cognitive Toolkit, Theano, and PlaidML.

Language used for project development is python with the environment of Jupyter which is a notebook, JupyterLab is the latest web-based interactive development environment for notebooks, code, and data. Its flexible interface allows users to configure and arrange workflows in data science, scientific computing, computational journalism, and machine learning.

Jupyter notebooks basically provides an interactive computational environment for developing Python based Data Science applications. They are formerly known as ipython notebooks. The following are some of the features of Jupyter notebooks that makes it one of the best components of Python ML ecosystem.Jupyter notebooks can illustrate the analysis process step by step by arranging the stuff like code, images, text, output etc. in a step by step manner.It helps a data scientist to document the thought process while developing the analysis process.One can also capture the result as the part of the notebook.With the help of jupyter notebooks, we can share our work with a peer also



Dataset

The process of data collection can be very hectic and time taking and requires tons and tons of resources. Before you get your hands dirty with data, you first need to identify why do you need it in the first place. The identification is the stage in which you establish the questions you will need to answer. For example, what is the customer's perception of our brand? Or what type of need the customer require or what will be the next turning curve of the trend? Once the questions are outlined you are ready for the next step. the stage where you start collecting the needed data. Here, you define which sources of information you will use and how you will use them. The collection of data can come in different forms such as internal or external sources, surveys, interviews, questionnaires, focus groups, among others.  An important note here is that the way you collect the information will be different in a quantitative and qualitative scenario. Once you have the necessary data it is time to clean it and leave it ready for analysis. Not all the data you collect will be useful, when collecting big amounts of information in different formats it is very likely that you will find yourself with duplicate or badly formatted data. To avoid this, before you start working with your data you need to make sure to erase any white spaces, duplicate records, or formatting errors. This way you avoid hurting your analysis with incorrect data.

As a freshy in artificial intelligence department and the service being a project.I will be using Kaggle

It is an online community platform for data scientists and machine learning enthusiasts. Kaggle allows users to collaborate with other users, find and publish datasets, use GPU integrated notebooks, and compete with other data scientists to solve data science challenges.



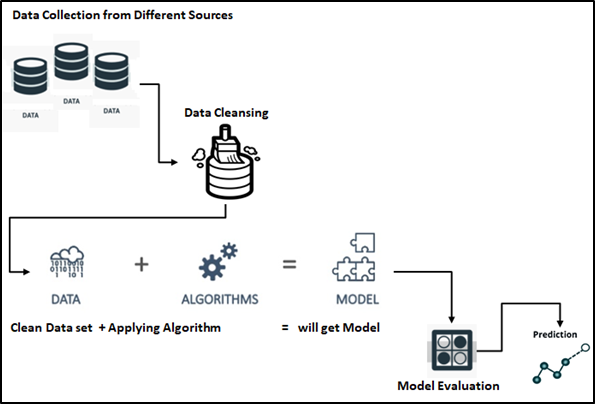
With the help of kaggle, we will obtain as many datasets as we require without any fear of data cleaning. Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled.

We will be using numerous libraries that are present in python that provide a great aid in machine learning and data science, we can use simple statements like pip-install-<name>

Numpy is another useful component that makes Python as one of the favorite languages for Data Science. It basically stands for Numerical Python and consists of multidimensional array objects. By using NumPy, we can perform the following important operations −Mathematical and logical operations on arrays, Fourier transformation, Operations associated with linear algebra.We can also see NumPy as the replacement of MatLab because NumPy is mostly used along with Scipy (Scientific Python) and Mat-plotlib (plotting library).

Pandas is another useful Python library that makes Python one of the favorite languages for Data Science. Pandas is basically used for data manipulation, wrangling and analysis. It was developed by Wes McKinney in 2008. With the help of Pandas, in data processing we can accomplish the following five steps –Load, Prepare, Manipulate, Model, Analyze.

Another useful and most important python library for Data Science and machine learning in Python is *Scikit-learn*. The following are some features of *Scikit-learn* that makes it so useful ,It is built on NumPy, SciPy, and Matplotlib, It is an open source and can be reused under BSD license, It is accessible to everybody and can be reused in various contexts, Wide range of machine learning algorithms covering major areas of ML like classification, clustering, regression, dimensionality reduction, model selection etc. can be implemented with the help of it.



Model Training

A training model is a dataset that is used to train an ML algorithm. It consists of the sample output data and the corresponding sets of input data that have an influence on the output. The training model is used to run the input data through the algorithm to correlate the processed output against the sample output. The result from this correlation is used to modify the model. This iterative process is called “model fitting”. The accuracy of the training dataset or the validation dataset is critical for the precision of the model. Model training in machine language is the process of feeding an ML algorithm with data to help identify and learn good values for all attributes involved. There are several types of machine learning models, of which the most common ones are supervised and unsupervised learning. Supervised learning is possible when the training data contains both the input and output values. Each set of data that has the inputs and the expected output is called a supervisory signal. The training is done based on the deviation of the processed result from the documented result when the inputs are fed into the model. Unsupervised learning involves determining patterns in the data. Additional data is then used to fit patterns or clusters. This is also an iterative process that improves the accuracy based on the correlation to the expected patterns or clusters. There is no reference output dataset in this method.

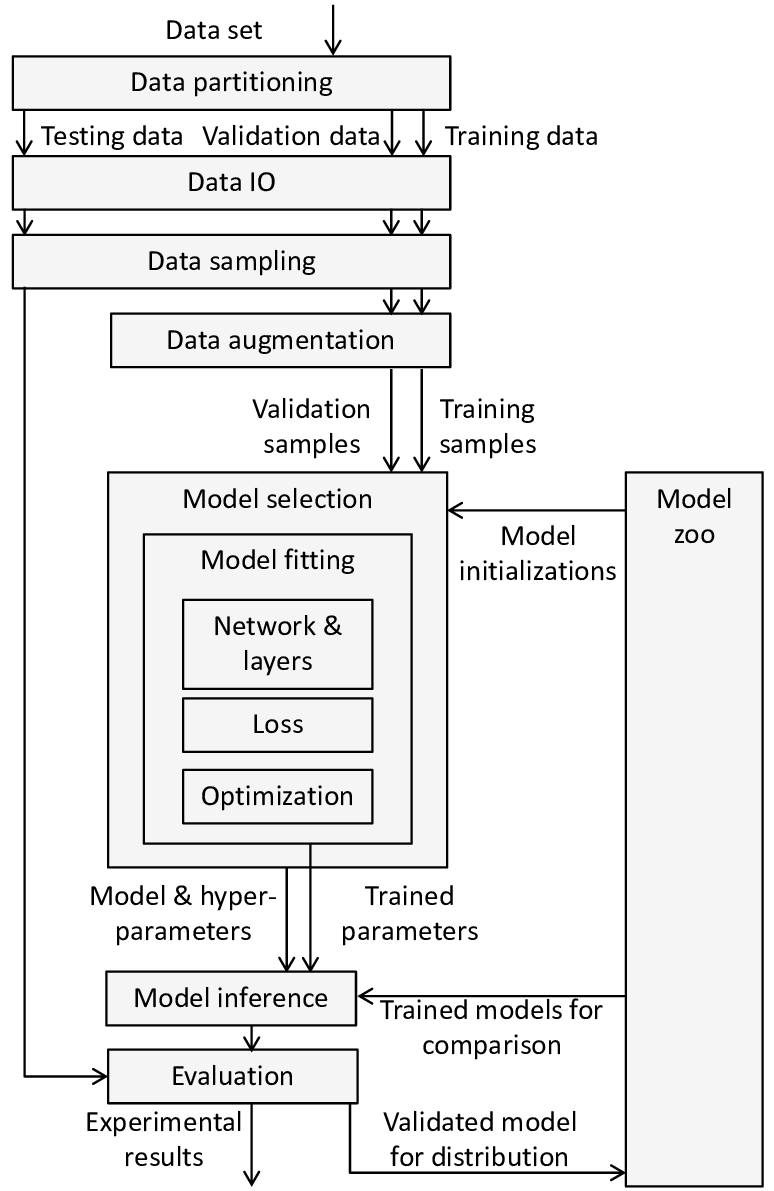


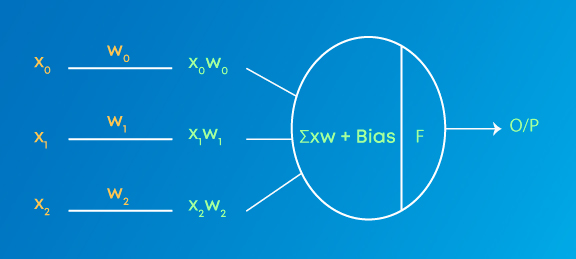
Fig. 6

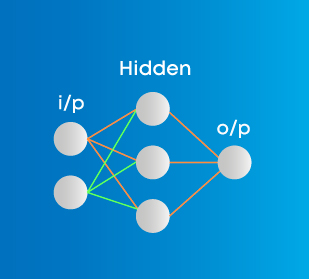
Preparing the data

The data preparation stage is when data is profiled, formatted and structured as needed to make it ready for training the model. This is the stage where the appropriate characteristics and attributes of data are selected. This stage is likely to have a direct impact on the execution time and results. This is also at the stage where data is categorized into two groups – one for training the ML model and the other for evaluating the model. Pre-processing of data by normalizing, eliminating duplicates and making error corrections is also carried out at this stage.

Preparing a neural network

Artificial neural networks are inspired by the biological neurons within the human body which activate under certain circumstances resulting in a related action performed by the body in response. Artificial neural nets consist of various layers of interconnected artificial neurons powered by activation functions that help in switching them ON/OFF. Like traditional machine learning, here too, there are certain values that neural nets learn in the training phase.





Different type of neural networks will be used for different type of datasets and complexities

The Densities of neuron will define the training complexity and evaluation

After Validation of the predicting model optimization of the model will be required to reduce loss and increase efficiency of the model. That will provide more accurate results and will be more stable with real world scenarios.

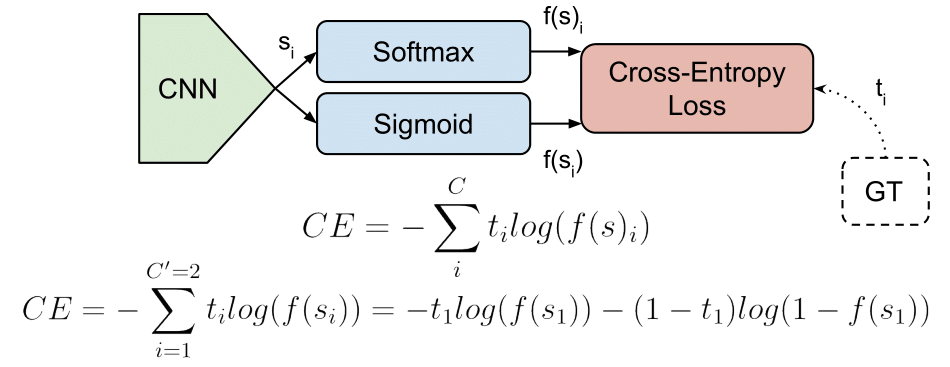
Optimization technique used are-Adam Optimization and cross entropy optimization

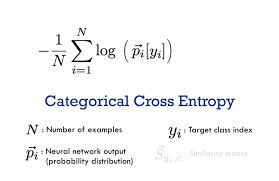
Adam Optimization

Adam is a replacement optimization algorithm for stochastic gradient descent for training deep learning models. Adam combines the best properties of the AdaGrad and RMSProp algorithms to provide an optimization algorithm that can handle sparse gradients on noisy problems

Crossentropy Optimization

The cross-entropy method is a versatile heuristic tool for solving difficult estimation and optimization problems, based on Kullback–Leibler (or cross-entropy) minimization. As an optimization method it unifies many existing population-based optimization heuristics.

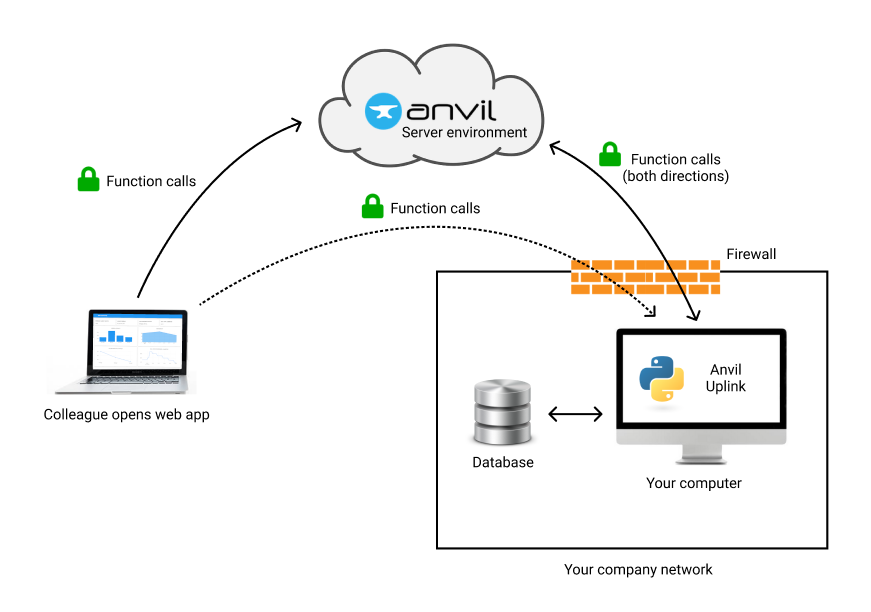




Web Application

Last and final step will be to create a user interface, and also a site that can be accessed globally and 247, for which I found a free web app builder anvil works, that is also a python based service. Anvil is a platform for building full-stack web apps with nothing but Python. No need to wrestle with JS, HTML, CSS, Python, SQL and all their frameworks.We can just pip install it and connect our code with the uplink and create a web appication





A link will be generated by anvil services which can access the project by anyone with the link and utilized, with proper and validated data we can publish the project and add some revenue options for some earnings to donation funds.

5. Tools/Technology Used:

5.1 Hardware Requirements

* 64 bit, 4 core, with minimum 3.0 Ghz each core clock frequency will be needed,8+ core is recommended for larger datasets
* 16 Gb RAM is recommended for efficient working
* Cloud computing services

5.2 Software Requirements

* **OS:** Windows 10/11
* **Language:** Python
* **IDE:** Visual Studio Code extension Jupyter Notebook
* **Plugins:** Tensorflow, Streamlit, Anvil + webservices

6. References:

* Infosys Springboard Deep learning and regression analysis https://infyspringboard.onwingspan.com
* Udemy Web Development course
* Udemy Python Bootcamp 2022
* Deep Learning for developers
* Next Gen Technologies
* YouTube tutorials and lectures

https://www.youtube.com/playlist?list=PLb0Z48JmZ1a-YhddZaIpp1YHuXejD-alz