Data Analyzer

A Report submitted in partial fulfilment of the requirement for the

degree of

B.Tech.

In

###### Computer Science & Engineering

By

##### Aakarshit Srivastava (2101641520001)

##### Dev Gupta (2101641520046)

##### Shashak Raj Singh (2101641520128)

##### Under the guidance of

##### (Dr. S.S. Parihar)

##### (Associate Professor)

##### Project Id: 22\_AI\_2A\_15



Pranveer Singh Institute of Technology, Kanpur

Dr A P J A K Technical University

Lucknow

## DECLARATION

This is to certify that Report entitled “ Data Analyzer” which is submitted by me in partial fulfilment of the requirement for the award of degree B.Tech. in Computer Science and Engineering to Pranveer Singh Institute of Technology, Kanpur Dr. A P J A K Technical University, Lucknow comprises only our own work and due acknowledgement has been made in the text to all other material used.

#### Date: 27/02/2023

Aakarshit Srivastava (2101641520001)

Dev Gupta(2101641520046)

Shashak Raj Singh(2101641520128)

## 

## 

## Certificate

This is to certify that Report entitled “ Data Analyzer ” which is submitted by Aakarshit Srivastava (2101641520001), Dev Gupta (2101641520046), Shashak Raj Singh (2101641520128) in partial fulfilment of the requirement for the award of degree B.Tech. in Computer Science & Engineering to Pranveer Singh Institute of Technology, Kanpur affiliated to Dr. A P J A K Technical University, Lucknow is a record of the candidate own work carried out by him under my supervision. The matter embodied in this thesis is original and has not been submitted for the award of any other degree.

|  |  |  |
| --- | --- | --- |
| Signature:     Dr. Vishal Nagar  Dean CSE Department,  PSIT, Kanpur |  | Signature:  Dr S.S. Parihar  Assistant Professor  CSE Department,  PSIT, Kanpur |

***ACKNOWLEDGEMENT***

*It gives us a great sense of pleasure to present the report of the B.Tech. Project undertaken during B.Tech. Third Year. We owe special debt of gratitude to our project supervisor Dr. S.S. Parihar, Department of Computer Science and Engineering, Pranveer Singh Institute of Technology, Kanpur, for his constant support and guidance throughout the course of our work. His sincerely, thoroughness and perseverance have been a constant source of inspiration for us. It is only his cognizant efforts that our endeavours have seen light of the day.*

*We also take the opportunity to acknowledge the contribution of Professor Dr. Vishal Nagar, Dean Computer Science & Engineering Department, Pranveer Singh Institute of Technology, Kanpur for his full support and assistance during the development of the project.*

*We also do not like to miss the opportunity to acknowledge the contribution of all faculty members of the department for their kind assistance and cooperation during the development of our project. Last but not the least, we acknowledge our friends for their contribution in the completion of the project.*

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |

***ABSTRACT***

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

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.

.

**Table of contents**

CHAPTER NO. TITLE PAGE NO.

DECLARATION 2

CERTIFICATE 3

ACKNOWLEDGEMENTS 4

ABSTRACT 5

CHAPTER I (INTRODUCTION) 8

CHAPTER II(METHODOLOGY/DESIGN) 17

CHAPTERIII(IMPLEMENTATION) 23

CHAPTERIV(TESTING) 29

CHAPTERV(RESULT) 31

CHAPTERVI(FUTURE ENHANCEMENTS) 34

**LIST OF SYMBOLS**

[x] Integer value of x.

≠ Not Equal

χ Belongs to

€ Euro- A Currency

\_ Optical distance

\_o Optical thickness or optical half thickness

**LIST OF ABBREVIATIONS**

AAM Active Appearance Model

ICA Independent Component Analysis

ISC Increment Sign Correlation

PCA Principal Component Analysis

ROC Receiver Operating Characteristics

Data Analyzer

**CHAPTER I. INTRODUCTION**

**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.

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

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

The process of systematically applying statistical and/or logical techniques to describe and illustrate, condense and recap, and evaluate data. According to Shamoo and Resnik (2003) various analytic procedures “provide a way of drawing inductive inferences from data and distinguishing the signal (the phenomenon of interest) from the noise (statistical fluctuations) present in the data”..

While data analysis in qualitative research can include statistical procedures, many times analysis becomes an ongoing iterative process where data is continuously collected and analyzed almost simultaneously. Indeed, researchers generally analyze for patterns in observations through the entire data collection phase (Savenye, Robinson, 2004). The form of the analysis is determined by the specific qualitative approach taken (field study, ethnography content analysis, oral history, biography, **unobtrusive** research) and the form of the data (field notes, documents, audiotape, videotape).

An essential component of ensuring data integrity is the accurate and appropriate analysis of research findings. Improper statistical analyses distort scientific findings, mislead casual readers (Shepard, 2002), and may negatively influence the public perception of research. Integrity issues are just as relevant to analysis of non-statistical data as well.

Considerations/issues in data analysis  
There are a number of issues that researchers should be cognizant of with respect to data analysis. These include:

* Having the necessary skills to analyze
* Concurrently selecting data collection methods and appropriate analysis
* Drawing unbiased inference
* Inappropriate subgroup analysis
* Following acceptable norms for disciplines
* Determining **statistical significance**
* Lack of clearly defined and objective **outcome measurements**
* Providing honest and accurate analysis
* Manner of presenting data
* Environmental/contextual issues
* Data recording method
* **Partitioning ‘text’** when analyzing qualitative data
* Training of staff conducting analyses
* Reliability and Validity
* **Extent of analysis**

Having necessary skills to analyzeA tacit assumption of investigators is that they have received training sufficient to demonstrate a high standard of research practice. Unintentional ‘scientific misconduct' is likely the result of poor instruction and follow-up. A number of studies suggest this may be the case more often than believed (Nowak, 1994; Silverman, Manson, 2003). For example, Sica found that adequate training of physicians in medical schools in the proper design, implementation and evaluation of clinical trials is “abysmally small” (Sica, cited in Nowak, 1994). Indeed, a single course in biostatistics is the most that is usually offered (Christopher Williams, cited in Nowak, 1994).

A common practice of investigators is to defer the selection of analytic procedure to a research team ‘statistician’. Ideally, investigators should have substantially more than a basic understanding of the rationale for selecting one method of analysis over another. This can allow investigators to better supervise staff who conduct the data analyses process and make informed decisions   
  
Concurrently selecting data collection methods and appropriate analysisWhile methods of analysis may differ by scientific discipline, the optimal stage for determining appropriate analytic procedures occurs early in the research process and should not be an afterthought. According to Smeeton and Goda (2003), “Statistical advice should be obtained at the stage of initial planning of an investigation so that, for example, the method of sampling and design of questionnaire are appropriate”.  
  
Drawing unbiased inferenceThe chief aim of analysis is to distinguish between an event occurring as either reflecting a true effect versus a false one. Any bias occurring in the collection of the data, or selection of method of analysis, will increase the likelihood of drawing a biased inference. Bias can occur when recruitment of study participants falls below minimum number required to demonstrate statistical power or failure to maintain a sufficient follow-up period needed to demonstrate an effect (Altman, 2001).   
  
Inappropriate subgroup analysis   
  
When failing to demonstrate statistically different levels between treatment groups, investigators may resort to breaking down the analysis to smaller and smaller subgroups in order to find a difference. Although this practice may not inherently be unethical, these analyses should be proposed before beginning the study even if the intent is exploratory in nature. If it the study is exploratory in nature, the investigator should make this explicit so that readers understand that the research is more of a hunting expedition rather than being primarily theory driven. Although a researcher may not have a theory-based hypothesis for testing relationships between previously untested variables, a theory will have to be developed to explain an unanticipated finding. Indeed, in exploratory science, there are no a priori hypotheses therefore there are no hypothetical tests. Although theories can often drive the processes used in the investigation of qualitative studies, many times patterns of behavior or occurrences derived from analyzed data can result in developing new theoretical frameworks rather than determined **a priori** (Savenye, Robinson, 2004).   
  
It is conceivable that multiple statistical tests could yield a significant finding by chance alone rather than reflecting a true effect. Integrity is compromised if the investigator only reports tests with significant findings, and neglects to mention a large number of tests failing to reach significance. While access to computer-based statistical packages can facilitate application of increasingly complex analytic procedures, inappropriate uses of these packages can result in abuses as well.   
  
Following acceptable norms for disciplines   
  
Every field of study has developed its accepted practices for data analysis. Resnik (2000) states that it is prudent for investigators to follow these accepted norms. Resnik further states that the norms are ‘…based on two factors:

(1) the nature of the variables used (i.e., quantitative, comparative, or qualitative),

(2) assumptions about the population from which the data are drawn (i.e., random distribution, independence, sample size, etc.). If one uses unconventional norms, it is crucial to clearly state this is being done, and to show how this new and possibly unaccepted method of analysis is being used, as well as how it differs from other more traditional methods. For example, Schroder, Carey, and Vanable (2003) juxtapose their identification of new and powerful data analytic solutions developed to count data in the area of HIV contraction risk with a discussion of the limitations of commonly applied methods.   
  
If one uses unconventional norms, it is crucial to clearly state this is being done, and to show how this new and possibly unaccepted method of analysis is being used, as well as how it differs from other more traditional methods. For example, Schroder, Carey, and Vanable (2003) juxtapose their identification of new and powerful data analytic solutions developed to count data in the area of HIV contraction risk with a discussion of the limitations of commonly applied methods.   
  
Determining significance   
  
While the conventional practice is to establish a standard of acceptability for statistical significance, with certain disciplines, it may also be appropriate to discuss whether attaining statistical significance has a true practical meaning, i.e., **‘clinical significance’**. Jeans (1992) defines ‘clinical significance’ as “the potential for research findings to make a real and important difference to clients or clinical practice, to health status or to any other problem identified as a relevant priority for the discipline”.

Kendall and Grove (1988) define clinical significance in terms of what happens when “… troubled and disordered clients are now, after treatment, not distinguishable from a meaningful and representative non-disturbed reference group”. Thompson and Noferi (2002) suggest that readers of counseling literature should expect authors to report either practical or clinical significance indices, or both, within their research reports. Shepard (2003) questions why some authors fail to point out that the magnitude of observed changes may too small to have any clinical or practical significance, “sometimes, a supposed change may be described in some detail, but the investigator fails to disclose that the trend is not statistically significant ”.  
  
Lack of clearly defined and objective outcome measurements No amount of statistical analysis, regardless of the level of the sophistication, will correct poorly defined objective outcome measurements. Whether done unintentionally or by design, this practice increases the likelihood of clouding the interpretation of findings, thus potentially misleading readers.  
  
Provide honest and accurate analysisThe basis for this issue is the urgency of reducing the likelihood of statistical error. Common challenges include the exclusion of **outliers**, filling in missing data, altering or otherwise changing data, data mining, and developing graphical representations of the data (Shamoo, Resnik, 2003).  
  
Manner of presenting data  
At times investigators may enhance the impression of a significant finding by determining how to present **derived data** (as opposed to data in its raw form), which portion of the data is shown, why, how and to whom (Shamoo, Resnik, 2003). Nowak (1994) notes that even experts do not agree in distinguishing between analyzing and massaging data. Shamoo (1989) recommends that investigators maintain a sufficient and accurate paper trail of how data was manipulated for future review.

Environmental/contextual issues   
  
The integrity of data analysis can be compromised by the environment or context in which data was collected i.e., face-to face interviews vs. focused group. The **interaction** occurring within a dyadic relationship (interviewer-interviewee) differs from the group dynamic occurring within a focus group because of the number of participants, and how they react to each other’s responses. Since the data collection process could be influenced by the environment/context, researchers should take this into account when conducting data analysis.

Data recording methodAnalyses could also be influenced by the method in which data was recorded. For example, research events could be documented by:

a. recording audio and/or video and transcribing later  
b. either a researcher or self-administered survey  
c. either **closed ended survey** or **open ended survey**  
d. preparing ethnographic field notes from a participant/observer  
e. requesting that participants themselves take notes, compile and submit them to researchers.

While each methodology employed has rationale and advantages, issues of objectivity and subjectivity may be raised when data is analyzed.

Partitioning the textDuring content analysis, staff researchers or ‘raters’ may use inconsistent strategies in analyzing text material. Some ‘raters’ may analyze comments as a whole while others may prefer to dissect text material by separating words, phrases, clauses, sentences or groups of sentences. Every effort should be made to reduce or eliminate inconsistencies between “raters” so that data integrity is not compromised.

Training of Staff conducting analyses A major challenge to data integrity could occur with the unmonitored supervision of inductive techniques. Content analysis requires raters to assign topics to text material (comments). The threat to integrity may arise when raters have received inconsistent training, or may have received previous training experience(s). Previous experience may affect how raters perceive the material or even perceive the nature of the analyses to be conducted. Thus one rater could assign topics or codes to material that is significantly different from another rater. Strategies to address this would include clearly stating a list of analyses procedures in the protocol manual, consistent training, and routine monitoring of raters.

Reliability and ValidityResearchers performing analysis on either quantitative or qualitative analyses should be aware of challenges to reliability and validity. For example, in the area of content analysis, Gottschalk (1995) identifies three factors that can affect the reliability of analyzed data:

* stability , or the tendency for coders to consistently re-code the same data in the same way over a period of time
* reproducibility , or the tendency for a group of coders to classify categories membership in the same way
* accuracy , or the extent to which the classification of a text corresponds to a standard or norm statistically

The potential for compromising data integrity arises when researchers cannot consistently demonstrate stability, reproducibility, or accuracy of data analysis

According Gottschalk, (1995), the validity of a content analysis study refers to the correspondence of the categories (the classification that raters’ assigned to text content) to the conclusions, and the generalizability of results to a theory (did the categories support the study’s conclusion, and was the finding adequately robust to support or be applied to a selected theoretical rationale?).

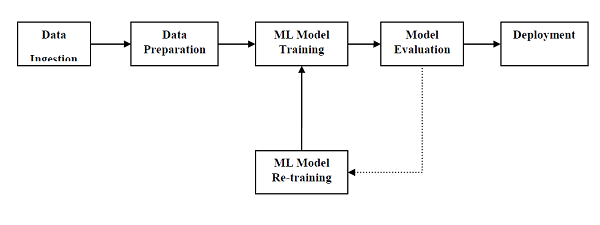
Extent of analysis  
  
Upon coding text material for content analysis, raters must classify each code into an appropriate category of a cross-reference matrix. Relying on computer software to determine a frequency or word count can lead to inaccuracies. “One may obtain an accurate count of that word's occurrence and frequency, but not have an accurate accounting of the meaning inherent in each particular usage” (Gottschalk, 1995). Further analyses might be appropriate to discover the dimensionality of the data set or identity new meaningful underlying variables.

Whether statistical or non-statistical methods of analyses are used, researchers should be aware of the potential for compromising data integrity. While statistical analysis is typically performed on quantitative data, there are numerous analytic procedures specifically designed for qualitative material including content, thematic, and ethnographic analysis. Regardless of whether one studies quantitative or qualitative phenomena, researchers use a variety of tools to analyze data in order to test hypotheses, discern patterns of behavior, and ultimately answer research questions. Failure to understand or acknowledge data analysis issues presented can compromise data integrity.

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.

**CHAPTER II. METHODOLOGY/DESIGN**

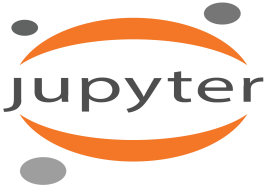
****

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



Machine learning is the ability of computers to learn without being explicitly programmed.

“**Without being explicitly programmed”**means, e.g. we’ve to predict the image of animals. So before machine learning, each image would be transformed to a vector by features then traditionally we’ve to write down a lot of rules or methods in order to get computers to be intelligent and detect the animals. Perhaps it would be the failure because its highly dependent upon current data sets.

So here comes the machine learning, using machine learning allows us to build a model that looks at all the feature sets, and their corresponding type of animals, and learn it learns the pattern of each animal. It is a model built by machine learning algorithms. It detects without explicitly being programmed to do so. In essence, machine learning follows the same process that a 4-year-old child uses to learn, understand, and differentiate animals.

So, machine learning algorithms, inspired by the human learning process, iteratively learn from data and allow computers to find hidden insights. These models help us in a variety of tasks, such as object recognition, summarization, recommendation, and so on.

Machine Learning impacts society in a very influential way. E.g.

* Paypal uses Machine Learning to detect fraud.
* Amazon uses Machine Learning to give you suggestion, what you can further buy.
* Banks also use Machine Learning to approve Loans.
* Telcos use customers data to segment them.

## Applications of Machine Learning;

There are many applications of machine learning like Search engine results, voice recognition, Number Plate Recognition, Dream Reader. This small sampling is just the beginning, from automatic cars to scientific discovery, any of these things are part of today’s world of machine learning.

If we talk about the search engine, Imagine if we’re on Google, we give very reliable information and speed, it’s automated and time goes on we got more information, the search engine returns better and better results.

Same with Voice Recognition, where its better and better voice recognizing what we’re saying and able to transcribe it for any of our Google commands or home devices where they recognized our voice, we can see that in a number of recognition apps.

So the use of machine learning is because it helps make life easier. It helps our processes be more consistent and reliable.

## Major Techniques of Machine Learning

So, let’s quickly examine a few of the more popular techniques.

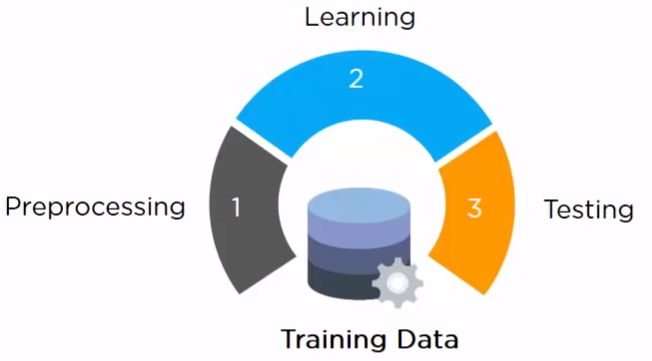
* **Regression / Estimation;** Predict Continous Values
  + This technique is used for predicting a **continuous value**;
    - E.g. predicting things like the price of a house based on its characteristics, or to estimate the CO2 emission from a car’s engine.
* **Classification;** Predicting the item class/category of a case.
  + A Classification technique is used for Predicting the class or category of a case.
    - E.g. if a cell is benign or malignant, or whether or not a customer will churn.
* **Clustering;** Finding the structure of data; summarization.
  + Clustering groups of similar cases.
    - E.g. Can find similar patients, or can be used for customer segmentation in the banking field.
* **Anomaly Detection;** Discovering abnormal and unusual cases.
* Anomaly detection is used to discover abnormal and unusual cases.
  + E.g. It is used for credit card fraud detection
* **Dimension Reduction;** Reducing the size of data (PCA).
  + Dimension reduction is used to reduce the size of data.
* **Recommendation Systems;** Recommending Items.
  + This associates people’s preferences with others who have similar tastes and recommends new items to them.
    - E.g. Recommended Books or Food.
* **Sequence mining;** Predicting next events, click-stream (Morkov Model, HMM).
  + Sequence mining is used for predicting the next event.
    - E.g. the click-stream in websites.

## 

## How does Machine Learning work?

### Phase#1: Learning

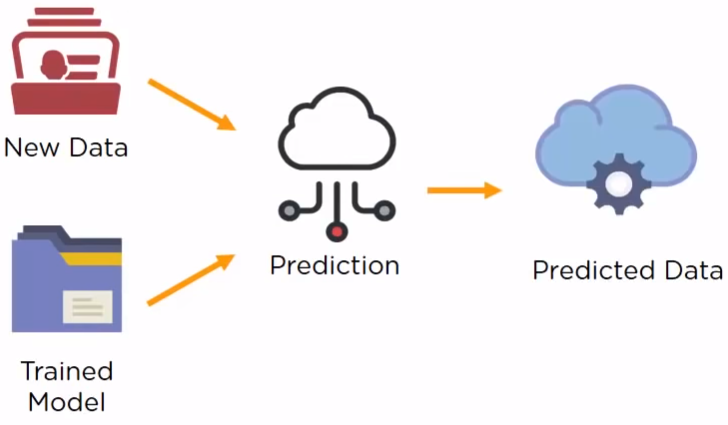
We’ve phase#1 which “Learning”, that broken up into three different steps;



* **Pre-Processing:** The first step is we need to clean and format the data. (That is because computers are not smart when it comes to figuring out the difference between a picture or text when we send it in), so the first thing we do is usually clean the data so all our pictures are in one file and text is being processed separately. Because if we would try to process text like we do a picture we’re not gonna get the right answer and vice-versa, once we pre-process the data and we’ve it nicely clean, we’re gonna go in and start learning.
* **Learning:** In this step, we take that data and learn from it. And here comes the supervised and unsupervised learning.
* **Testing:** In this step, we have it a test to make sure we are getting the right answer out of it.

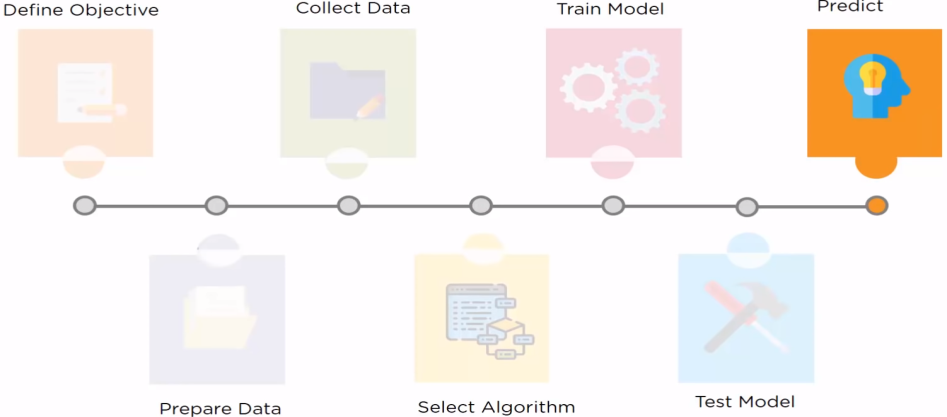
### Phase#2: Prediction

In this phase, we’re actually using it or putting it into commercial use and that is to do a prediction and on there now we have our train model and our new data come together and output is going to be a prediction of what we are looking for. We can see that in the form of predicted data.



Machine Learning Workflow; It works iteratively;

* Define Objective
* Prepare the Data
* Collect Data
* Select Algorithm
* Train Model
* Test Model
* Predict



## Machine Learning with Python

Python is a preferred language among data scientists. We can write our machine learning algorithm using python, and it works very well. However, there are a lot of modules and libraries already implemented in python that can make our life much easier.

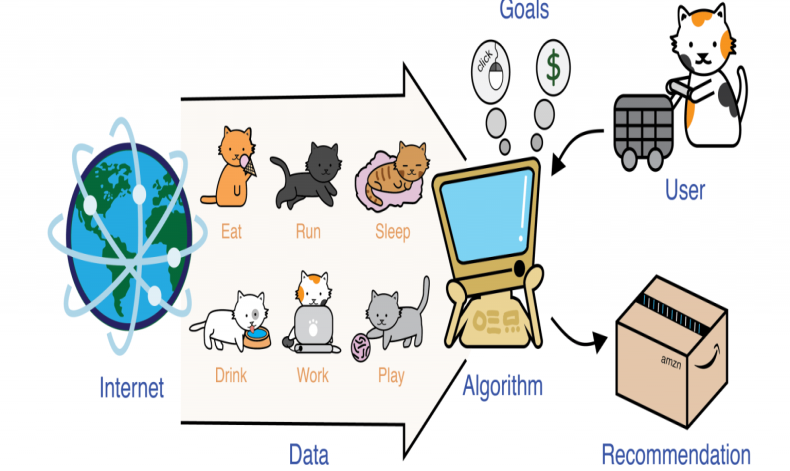
Numpy is a math library to work with n-dimensional arrays in Python. It enables you to do computation efficiently and effectively. It is better than regular python because of its amazing capabilities.

* E.g. for working with arrays, dictionaries, functions, datatypes, and working with images, we need to know Numpy.

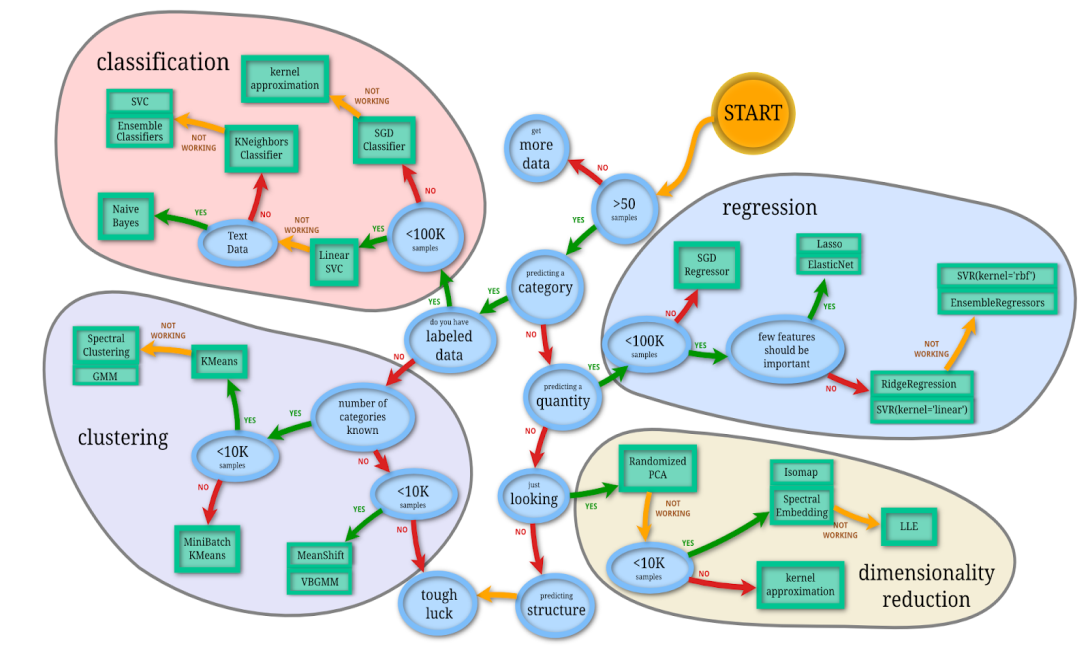
SciPy is a collection of numerical algorithms and domain-specific toolboxes, including signal processing, optimization, statistics and much more. SciPy is a good library for scientific and high-performance computation.

Matplotlib is a very popular plotting package that provides 2D plotting as well as 3D plotting.

Pandas library is a very high-level python library that provides high-performance, easy to use data structures. It has many functions for data importing, manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.



### Scikit Learn



Scikit-learn is a collection of algorithms and tools for machine learning. Scikit-learn is a free machine learning library for the Python programming language.

* It has most of the classification, regression and clustering algorithms.
* It’s designed to work with the Python numerical and scientific libraries
  + NumPy
  + SciPy.

Most of the tasks that need to be done in a machine learning pipeline are implemented already in scikit learn, including

* Pre-processing of data.
* Feature selection
* Feature extraction
* Train/Test splitting
* Defining the Algorithms.
* Fitting models
* Tuning parameters
* Prediction
* Evaluation, and
* Exporting the model.

**CHAPTER III. IMPLEMENTATION**

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.



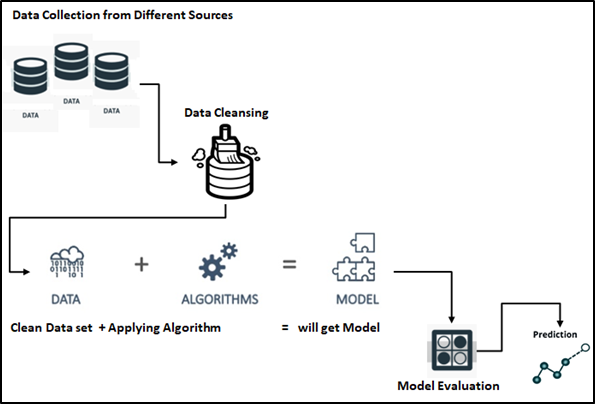
With the help of kaggle, we will obtain as many datasets as we require without any fear of data cleaning. Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many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

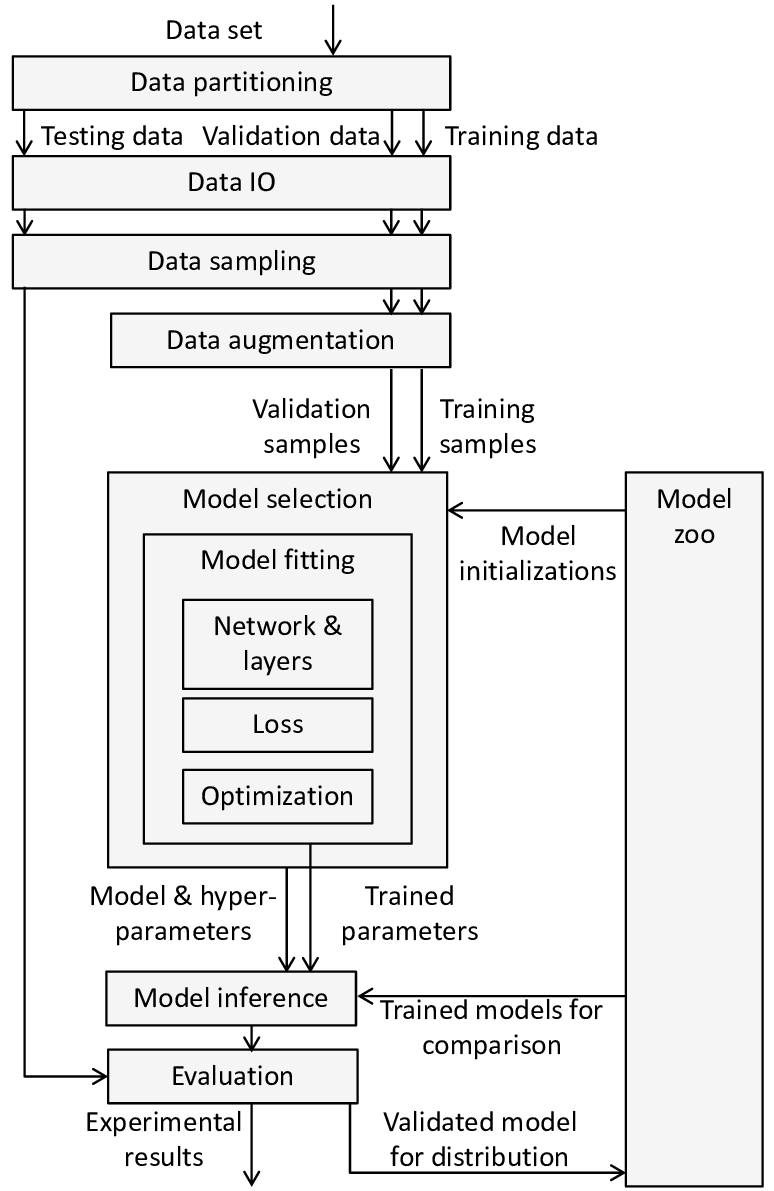


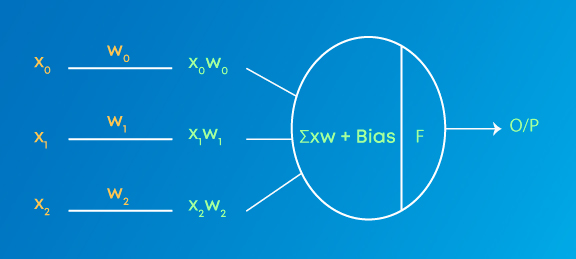
Fig. 6

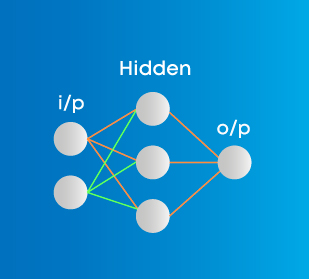
Preparing the data

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.

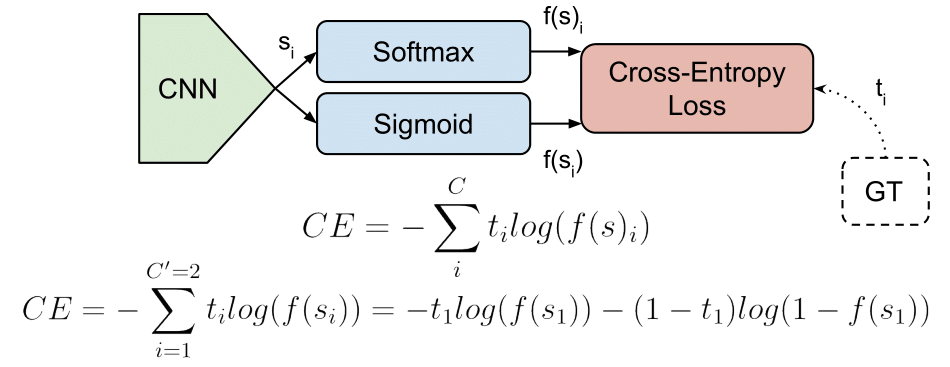
Optimization technique used are-Adam Optimization and cross entropy optimization

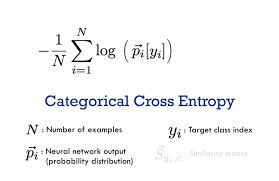
Adam Optimization

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.

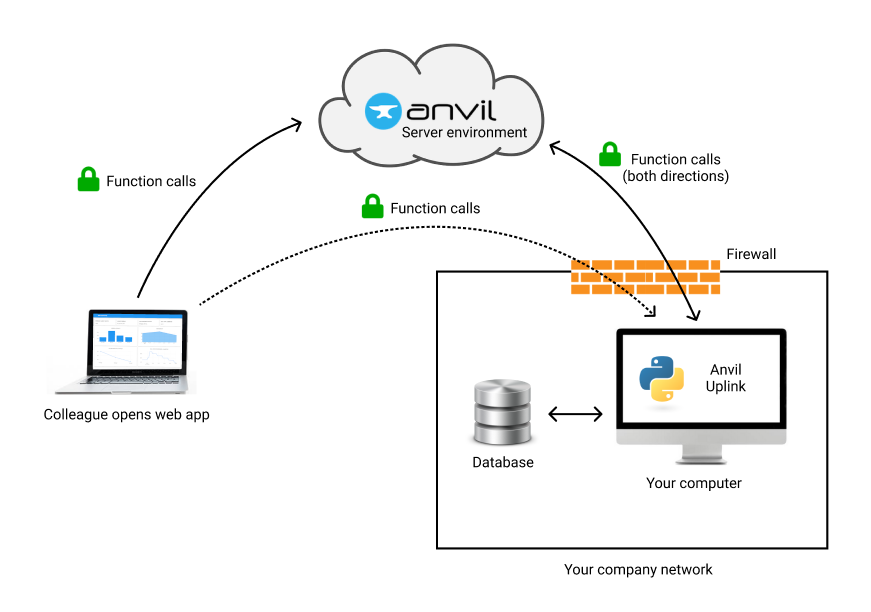




Web Application

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.

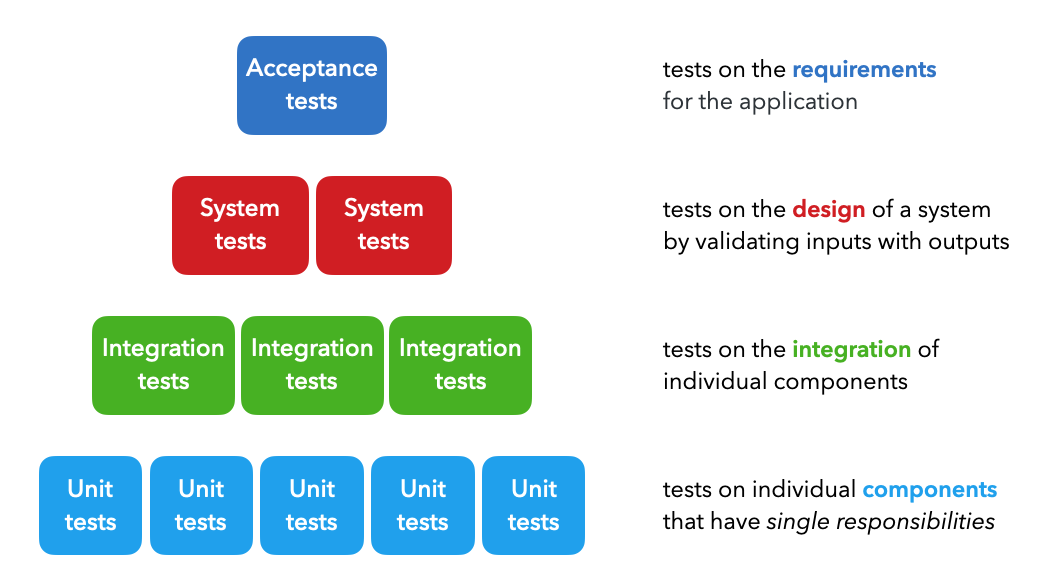
**CHAPTER IV. TESTING**

### Types of tests

There are four majors types of tests which are utilized at different points in the development cycle:

1. Unit tests: tests on individual components that each have a [single responsibility](https://en.wikipedia.org/wiki/Single-responsibility_principle) (ex. function that filters a list).
2. Integration tests: tests on the combined functionality of individual components (ex. data processing).
3. System tests: tests on the design of a system for expected outputs given inputs (ex. training, inference, etc.).
4. Acceptance tests: tests to verify that requirements have been met, usually referred to as User Acceptance Testing (UAT).
5. Regression tests: tests based on errors we've seen before to ensure new changes don't reintroduce them.

While ML systems are probabilistic in nature, they are composed of many deterministic components that can be tested in a similar manner as traditional software systems. The distinction between testing ML systems begins when we move from testing code to testing the [data](https://madewithml.com/courses/mlops/testing/#data) and [models](https://madewithml.com/courses/mlops/testing/#models).



There are many other types of functional and non-functional tests as well, such as smoke tests (quick health checks), performance tests (load, stress), security tests, etc. but we can generalize all of these under the system tests above.

### How should we test?

The framework to use when composing tests is the [Arrange Act Assert](http://wiki.c2.com/?ArrangeActAssert) methodology.

* Arrange: set up the different inputs to test on.
* Act: apply the inputs on the component we want to test.
* Assert: confirm that we received the expected output.

Cleaning is an unofficial fourth step to this methodology because it's important to not leave remnants of a previous test which may affect subsequent tests. We can use packages such as [pytest-randomly](https://github.com/pytest-dev/pytest-randomly) to test against state dependency by executing tests randomly.

In Python, there are many tools, such as [unittest](https://docs.python.org/3/library/unittest.html), [pytest](https://docs.pytest.org/en/stable/), etc. that allow us to easily implement our tests while adhering to the Arrange Act Assert framework. These tools come with powerful built-in functionality such as parametrization, filters, and more, to test many conditions at scale.

### What should we test?

When arranging our inputs and asserting our expected outputs, what are some aspects of our inputs and outputs that we should be testing for?

* **inputs**: data types, format, length, edge cases (min/max, small/large, etc.)
* **outputs**: data types, formats, exceptions, intermediary and final outputs

 We'll cover specific details pertaining to what to test for regarding our [data](https://madewithml.com/courses/mlops/testing/#data) and [models](https://madewithml.com/courses/mlops/testing/#models) below.

## Best practices

Regardless of the framework we use, it's important to strongly tie testing into the development process.

* atomic: when creating functions and classes, we need to ensure that they have a [single responsibility](https://en.wikipedia.org/wiki/Single-responsibility_principle) so that we can easily test them. If not, we'll need to split them into more granular components.
* compose: when we create new components, we want to compose tests to validate their functionality. It's a great way to ensure reliability and catch errors early on.
* reuse: we should maintain central repositories where core functionality is tested at the source and reused across many projects. This significantly reduces testing efforts for each new project's code base.
* regression: we want to account for new errors we come across with a regression test so we can ensure we don't reintroduce the same errors in the future.
* coverage: we want to ensure [100% coverage](https://madewithml.com/courses/mlops/testing/#coverage) for our codebase. This doesn't mean writing a test for every single line of code but rather accounting for every single line.
* automate: in the event we forget to run our tests before committing to a repository, we want to auto run tests when we make changes to our codebase. We'll learn how to do this locally using [pre-commit hooks](https://madewithml.com/courses/mlops/pre-commit/) and remotely via [GitHub actions](https://madewithml.com/courses/mlops/cicd/#github-actions) in subsequent lessons.

## Test-driven development

[Test-driven development (TDD)](https://en.wikipedia.org/wiki/Test-driven_development) is the process of writing a test before writing the functionality to ensure that tests are always written. This is in contrast to writing functionality first and then composing tests afterwards. Here are our thoughts on this:

* good to write tests as we progress, but it does signify 100% correctness.
* initial time should be spent on design before ever getting into the code or tests.

Perfect coverage doesn't mean that our application is error free if those tests aren't meaningful and don't encompass the field of possible inputs, intermediates and outputs. Therefore, we should work towards better design and agility when facing errors, quickly resolving them and writing test cases around them to avoid next time.

## Application

In our [application](https://github.com/GokuMohandas/mlops-course), we'll be testing the code, data and models. We'll start by creating a separate tests directory with code subdirectory for testing our tagifai scripts. We'll create subdirectories for testing [data](https://madewithml.com/courses/mlops/testing/#🔢nbsp-data) and [models](https://madewithml.com/courses/mlops/testing/#🤖nbsp-models) soon below.

mkdir tests

cd tests

mkdir app config model tagifai

touch <SCRIPTS>

cd ../

tests/

└── code/

│ ├── test\_data.py

│ ├── test\_evaluate.py

│ ├── test\_main.py

│ ├── test\_predict.py

│ └── test\_utils.py

Feel free to write the tests and organize them in these scripts after learning about all the concepts in this lesson. We suggest using our [tests](https://github.com/GokuMohandas/mlops-course/tree/main/tests) directory on GitHub as a reference.

Notice that our tagifai/train.py script does not have it's respective tests/code/test\_train.py. Some scripts have large functions (ex. train.train(), train.optimize(), predict.predict(), etc.) with dependencies (ex. artifacts) and it makes sense to test them via tests/code/test\_main.py.

## 💻  Code

We'll start by testing our code and we'll use [pytest](https://docs.pytest.org/en/stable/) as our testing framework for it's powerful builtin features such as [parametrization](https://madewithml.com/courses/mlops/testing/#parametrize), [fixtures](https://madewithml.com/courses/mlops/testing/#fixtures), [markers](https://madewithml.com/courses/mlops/testing/#markers) and more.

pip install pytest==7.1.2

Since pytest is not integral to the core machine learning operations (ie. only a developer would need to run tests), let's create a separate list in our setup.py and add it to our extras\_require:

|  |
| --- |
|  |

|  |
| --- |
| # setup.py  test\_packages = [  "pytest==7.1.2",  ]  # Define our package  setup(  ...  extras\_require={  "dev": docs\_packages + style\_packages + test\_packages,  "docs": docs\_packages,  "test": test\_packages,  },  ) |

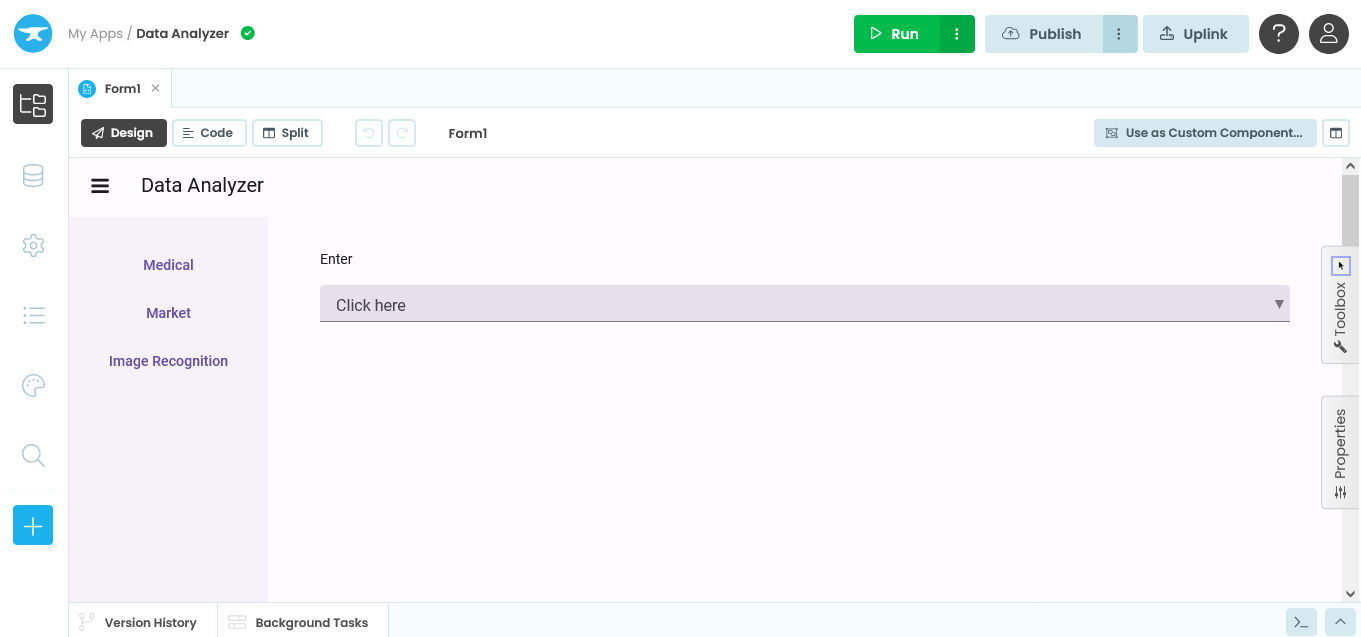
We created an explicit test option because a user will want to only download the testing packages. We'll see this in action when we use [CI/CD workflows](https://madewithml.com/courses/mlops/cicd/) to run tests via GitHub Actions.

**CHAPTER V. RESULT AND ANALYSIS**

Three versions of the project were developed to overcome different problems and development errors.

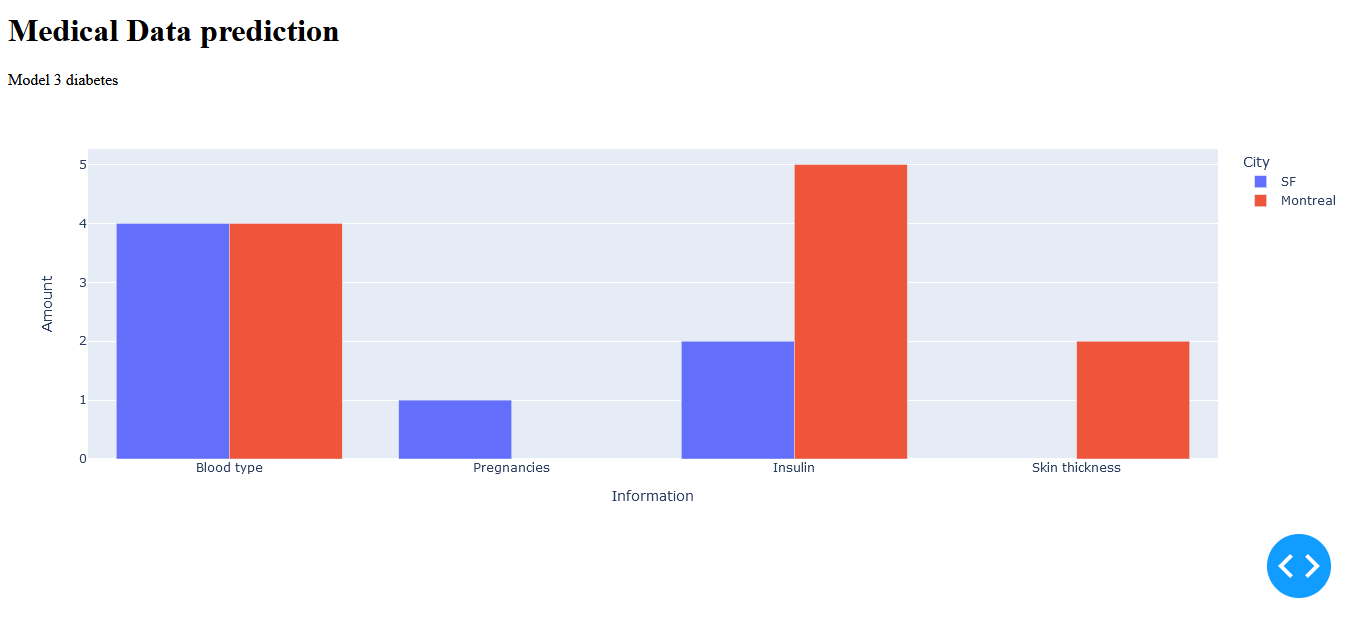
V1.0

This version is using the same machine learning models with the web application built on anvil app buider.With the help of anvil app builder the website and code can be uploaded and the product can be used by anyone on any device at any part of the world and at any time. The problem faced by us was the interface generation was a bit advanced and many python libraries required the premium for the anvil app.



V2.0

This version is using dash library to generate the output.the drawback with this interface is that it cannot generate interactive outputs.It can only showcase preprocessed data.This version can be enhanced with more libraries usage to make is user friendly and more usable



**V3.0**

This version uses the streamlit library to create a simple user friendly interface with simple few lines of code.No need to make an html page nor link it with javascript.



## Structure Of The Project

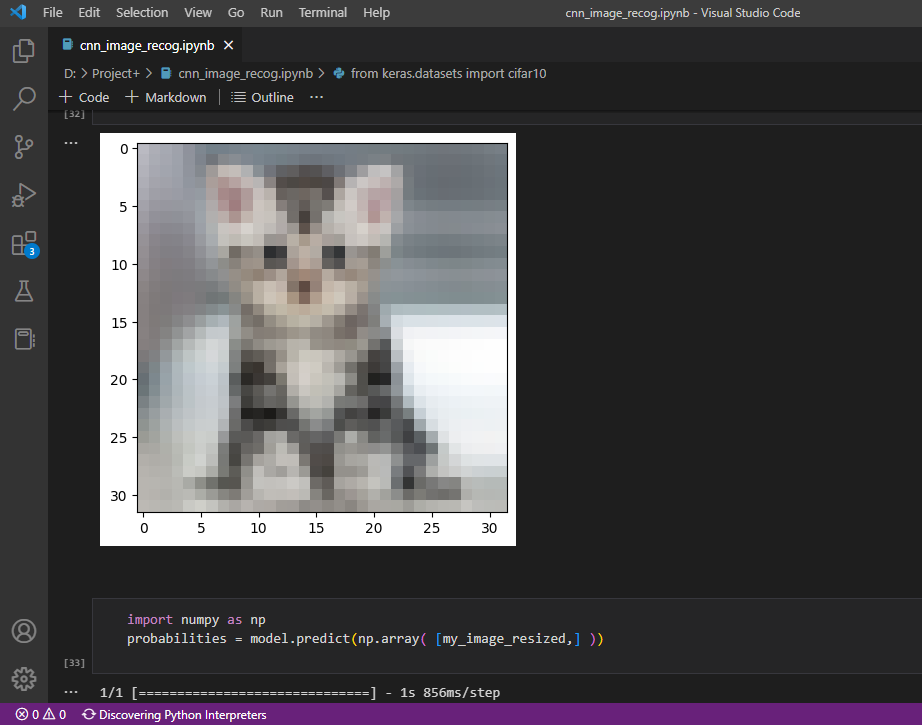
* Each prediction page is conneceted with a Machine Learning Model which uses Random Forest Classifier to predict the results.
* Also we have 3 different datasets used for each prediction.
* We can land into each prediction site of the web app from the options in the Navigation Menu.
* Each prediction is done with the help of 4 features which will be taken as input from the user.
* The most relevant features are taken into consideration for prediction also these features can be found out with simple tests or analysis without visiting any doctor.
* So the victim can get a broad overview of their health condition.

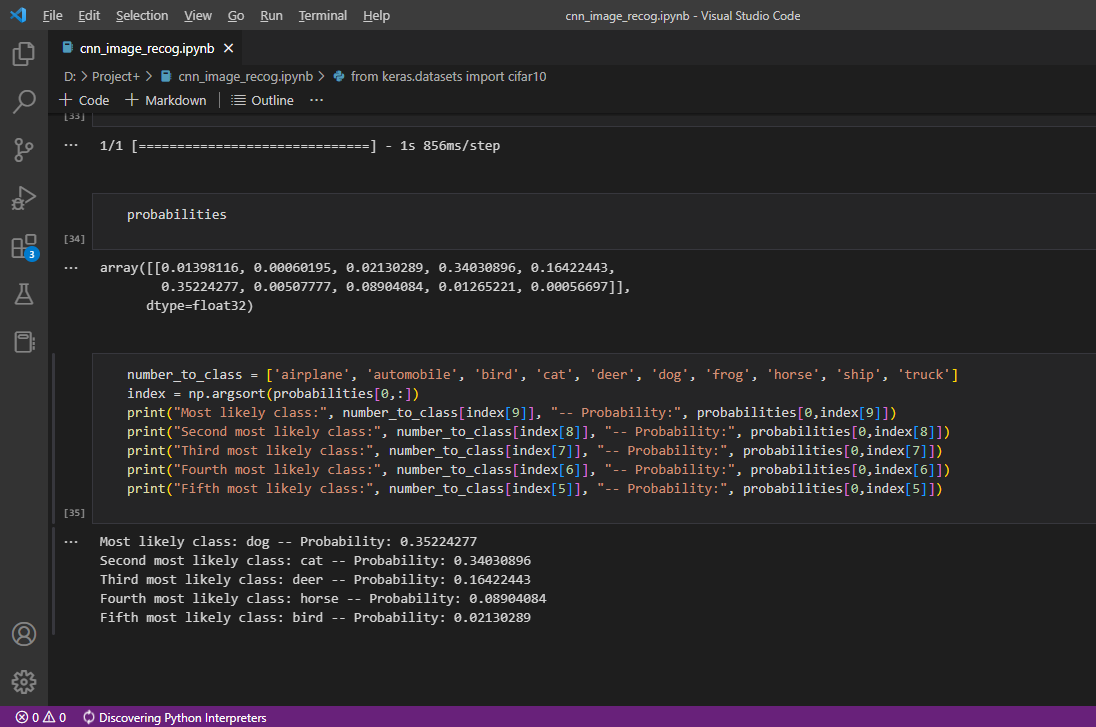
## The features taken into consideration

| **Disease** | **Features** |
| --- | --- |
| Covid-19 | Dry Cough, Fever, Sore Throat, Breathing Problem |
| Diabetes | Glucose, Insulin, Body Mass Index(BMI), Age |
| Heart Disease | Chest Pain, Blood Pressure(BP), Cholestrol, Max Heart Rate(HR) |

## Deployment Of The Project

After the modeling part the model is deployed using Streamlit library on Streamlit Share so that the app is available for usage for everyone.





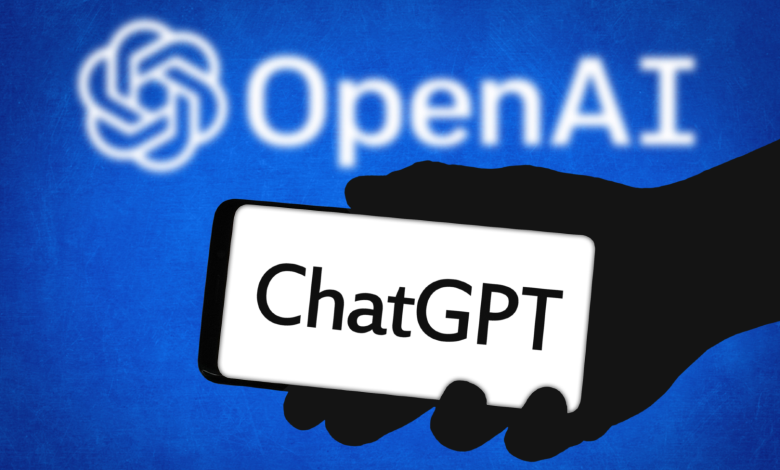
**CHAPTER VI. FUTURE ENHANCEMENTS**

A web development machine such as this, as with all things AI, will continue learning with the aim of delivering better results in the future. However, until there’s an AI web development program, we need to enhance,add more features,Increase precision of every single working model to get more and more accurate.

The scope of web development in the future extends to using AI in different ways.

* Web development companies can use AI-based algorithms to get real-time design/layout suggestions.
* Web developers may enhance customer engagement by using AI-powered chatbots and other communication tools.
* AI-based coding brings with it the potential to reduce turnaround times significantly and create more intuitive applications.
* Website owners may use AI systems to analyze user behavior and provide personalized [user experiences](https://www.topnotchdezigns.com/why-is-user-experience-really-important-for-a-website/).
* E-commerce web development can benefit through AI image recognition to detect counterfeit products and the technology can also help identify fake reviews.
* AI algorithms can help manage monotonous tasks such as website maintenance and collection of data.

A number of corporations such as Microsoft, Google, and Facebook have already embraced artificial intelligence and make use of it in different areas of computer programming. This technology is being used by Google to improve search engine efficiency and by Wikipedia to identify articles that are damaged or carry inaccurate information.

****

**Conclusion**

As in the past, one will witness various web development trends in the future. If you work in this field, you need to keep your eyes wide open because some of these trends will play a major role in what you do. If you’re looking at creating a new website for your business or updating an existing one, one or more of the aspects mentioned here might have a bearing on how you approach the process.

Bear in mind that just how well your website or app is developed plays an important role in how your target audience interacts with your business in the online world. As a result, it’s important for you to work with [a web agency](https://www.topnotchdezigns.com/) that specializes in popular web technologies of the day and also keeps up to date with web development trends of the future. Only then will you get a website or an app that is prepared to meet your customers’ ever-evolving needs.