SIGNIFICANT OF EARTHQUAKE HAPPENS USING DATASCIENCE

A PROJECT REPORT

Submitted by

BHARATHI S [REGISTER NO:211414104040]
VASUNDHARA G [REGISTER NO:211414104305]
GNANADEEPAM B [REGISTER NO:211418104065]

in partial fulfillment for the award of the degree

of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE AND ENGINEERING

PANIMALAR ENGINEERING COLLEGE

(An Autonomous Institution, Affiliated to Anna University, Chennai)

MAY 2022

PANIMALAR ENGINEERING COLLEGE

(An Autonomous Institution, Affiliated to Anna University, Chennai)

BONAFIDE CERTIFICATE

Certified that this project report **SIGNIFICANT OF EARTHQUAKE HAPPENS USING DATASCIENCE** is the bonafide work of **BHARATHI S [211418104040], VASUNDHARA G [211418104305], GNANADEEPAM B [211418104065]** who carried out the project work under my supervision.

SIGNATURE

Dr.S.MURUGAVALLI,M.E.,Ph.D .,
HEAD OF THE DEPARTMENT

DEPARTMENT OF CSE,
PANIMALAR ENGINEERING COLLEGE,
NASARATHPETTAI,
POONAMALLEE,
CHENNAI-600 123.

SIGNATURE

Mrs.D.JENNIFER,M.E.,
ASSISTANT PROFESSOR

DEPARTMENT OF CSE,
PANIMALAR ENGINEERING COLLEGE,
NASARATHPETTAI,
POONAMALLEE,
CHENNAI-600 123.

Certified that the above mentioned students were examined in End Semester project viva-voice held on ______________________.

INTERNAL EXAMINER

EXTERNAL EXAMINER

DECLARATION BY THE STUDENT

We **BHARATHI S[211418104040], VASUNDHARA G [211418104305], GNANADEEPAM B[211418104065]** hereby declare that this project report titled **SIGNIFICANT OF EARTHQUAKE HAPPENS USING DATASCIENCE**, under the guidance of **Mrs.D.JENNIFER,M.E.,** is the orginial work done by us and we have not plagiarized or submitted to any other degree in any university by us.

BHARATHI S

VASUNDHARA G

GNANADEEPAM B

ACKNOWLEDGEMENT

We would like to express our deep gratitude to our respected Secretary and Correspondent **Dr.P.CHINNADURAI, M.A., Ph.D.** for his kind words and enthusiastic motivation, which inspired us a lot in completing this project.

We express our sincere thanks to our Directors **Tmt.C.VIJAYARAJESWARI**, **Dr.C.SAKTHI KUMAR,M.E.,Ph.D** and **Dr.SARANYASREE SAKTHI KUMAR B.E.,M.B.A.,Ph.D.,** for providing us with the necessary facilities to undertake this project.

We also express our gratitude to our Principal **Dr.K.Mani, M.E., Ph.D.** who facilitated us in completing the project.

We thank the Head of the CSE Department, **Dr. S.MURUGAVALLI , M.E.,Ph.D.,** for the support extended throughout the project.

We would like to thank my **Project Guide Mrs.D.JENNIFER,M.E.,** and all the faculty members of the Department of CSE for their advice and encouragement for the successful completion of the project.

BHARATHI S [211418104040]
VASUNDHARA G[211418104305]
GNANADEEPAM B[211418104065]

ABSTRACT

An earthquake is the shaking of the surface of the Earth resulting from a sudden release of energy in the Earth's lithosphere that creates seismic waves. At the Earth's surface, earthquakes manifest themselves by shaking and displacing or disrupting the ground. So predicting the factors of an earthquake is a challenging job as an earthquake does not show specific patterns resulting in inaccurate predictions. Techniques based on machine learning are well known for their capability to find hidden patterns in data. The machine learning model is built based on the past data related to earthquakes where the model can learn the pattern from the data and takes the consideration of the factors. The factors which are taken into consideration are the pre-processed data that is the dependency of the factors are checked in accordance with earthquake. Comparison of machine learning algorithms are done for better prediction and performance metrics are also calculated an evaluated**.**

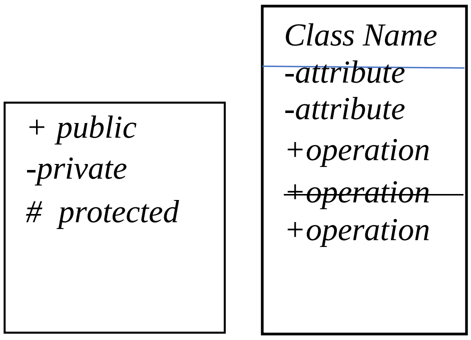
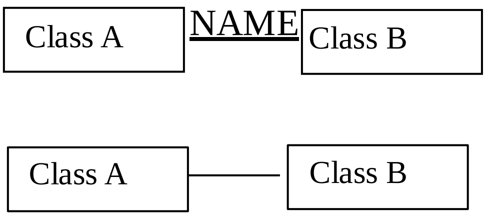
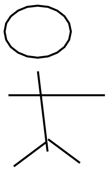
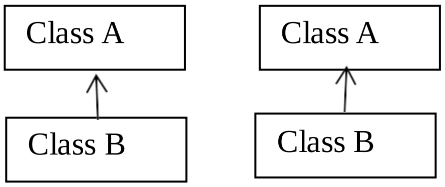
TABLE OF CONTENTS

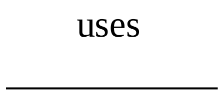
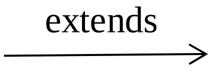
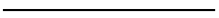
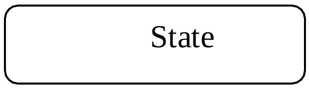
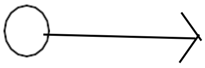
LIST OF FIGURES

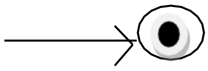
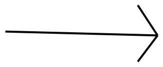
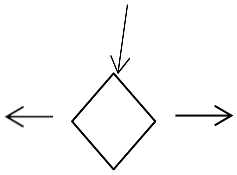
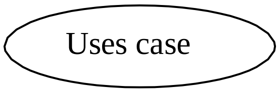
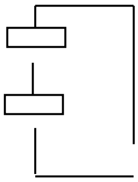
LIST OF ABBREVIATIONS

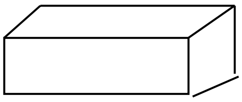
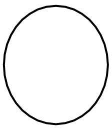
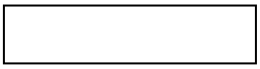
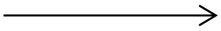
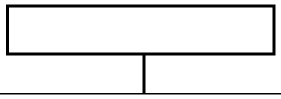
EEWS	Earthquake Early Warning Systems
PTWP	Picking Target Window Prediction
MTR	Multi-Target Regression
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative

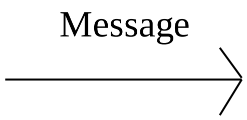
LIST OF SYMBOLS

S.NO	NOTATION NAME	NOTATION	DESCRIPTION
1.	Class	 + public -private # protected Class Name -attribute -attribute +operation +operation +operation	Represents a collection of similar entities grouped together.
2.	Association		Associations represents static relationships between classes. Roles represents the way the two classes see each other.
3.	Actor		It aggregates several classes into a single classes.
4.	Aggregation		Interaction between the system and

			external environment
5.	Relation(uses)		Used for additional process communication.
6.	Relation (extends)		Extends relationship is used when one use case is similar to another use case but does a bit more.
7.	Communication		Communication between various use cases.
8.	State		State of the process.
9.	Initial State		Initial state of the object

10.	Final state		Final state of the object
11.	Control flow		Represents various control flow between the states.
12.	Decision box		Represents decision making process from a constraint
13.	Use case		Interaction between the system and external environment.
14.	Component		Represents physical modules which is a collection of components.

15.	Node		Represents physical modules which are a collection of components
16.	Data Process/State		A circle in DFD represents a state or process which has been triggered due to some event or action.
17.	External entity		Represents external entities such as keyboard, sensors etc.
18.	Transition		Represents communication that occurs between processes.
19.	Object Lifeline		Represents the vertical dimensions

			that the object communications.
20.	Message		Represents the message exchanged.

CHAPTER 1

INTRODUCTION

1.1 OVERVIEW OF THE PROJECT

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains.

The term "data science" has been traced back to 1974, when Peter Naur proposed it as an alternative name for computer science. In 1996, the International Federation of Classification Societies became the first conference to specifically feature data science as a topic. However, the definition was still in flux.

The term “data science” was first coined in 2008 by D.J.Patil, and Jeff Hammerbacher, the pioneer leads of data and analytics efforts at LinkedIn and Facebook. In less than a decade, it has become one of the hottest and most trending professions in the market.

Data science is the field of study that combines domain expertise, programming skills, and knowledge of mathematics and statistics to extract meaningful insights from data.

Data science can be defined as a blend of mathematics, business acumen, tools, algorithms and machine learning techniques, all of which help us in finding out the hidden insights or patterns from raw data which can be of major use in the formation of big business decisions.

They have business acumen and analytical skills as well as the ability to mine, clean, and present data. Businesses use data scientists to source, manage, and analyze large amounts of unstructured data.

Artificial intelligence (AI) refers to the simulation of human intelligence in machines that are programmed to think like humans and mimic their actions. The term may also be applied to any machine that exhibits traits associated with a human mind such as learning and problem-solving.

Artificial intelligence (AI) is intelligence demonstrated by machines, as opposed to the natural intelligence displayed by humans or animals. Leading AI textbooks define the field as the study of "intelligent agents" any system that perceives its environment and takes actions that maximize its chance of achieving its goals. Some popular accounts use the term "artificial intelligence" to describe machines that mimic "cognitive" functions that humans associate with the human mind, such as "learning" and "problem solving", however this definition is rejected by major AI researchers.

Artificial intelligence is the simulation of human intelligence processes by machines, especially computer systems. Specific applications of AI include expert systems, natural language processing, speech recognition and machine vision.

AI applications include advanced web search engines, recommendation systems (used by Youtube, Amazon and Netflix), Understanding human speech (such as Siri or Alexa), self-driving cars (e.g. Tesla), and competing at the highest level in strategic game systems (such as chess and Go), As machines become increasingly capable, tasks considered to require "intelligence" are often removed from the definition of AI, a phenomenon known as the AI effect. For instance, optical character recognition is frequently excluded from things considered to be AI, having become a routine technology.

Artificial intelligence was founded as an academic discipline in 1956, and in the years since has experienced several waves of optimism, followed by disappointment and the loss of funding (known as an "AI winter"), followed by new approaches, success and renewed funding. AI research has tried and discarded many different approaches during its lifetime, including simulating the brain, modeling human problem solving, formal logic, large databases of knowledge and imitating animal behavior. In the first decades of the 21st century, highly mathematical statistical machine learning has dominated the field, and this technique has proved highly successful, helping to solve many challenging problems throughout industry and academia.

The various sub-fields of AI research are centered around particular goals and the use of particular tools. The traditional goals of AI research include reasoning, knowledge representation, planning, learning, natural language processing, perception and the ability to move and manipulate objects. General intelligence (the ability to solve an arbitrary problem) is among the field's long-term goals. To solve these problems, AI researchers use versions of search and mathematical optimization, formal logic, artificial neural networks, and methods based on statistics, probability and economics. AI also draws upon computer science, psychology, linguistics, philosophy, and many other fields.

The field was founded on the assumption that human intelligence "can be so precisely described that a machine can be made to simulate it". This raises philosophical arguments about the mind and the ethics of creating artificial beings endowed with human-like intelligence. These issues have been explored by myth, fiction and philosophy since antiquity. Science fiction and futurology have also suggested that, with its enormous potential and power, AI may become an existential risk to humanity.

As the hype around AI has accelerated, vendors have been scrambling to promote how their products and services use AI. Often what they refer to as AI is simply one component of AI, such as machine learning. AI requires a foundation of specialized hardware and software for writing and training machine learning algorithms. No one programming language is synonymous with AI, but a few, including Python, R and Java, are popular.

In general, AI systems work by ingesting large amounts of labeled training data, analyzing the data for correlations and patterns, and using these patterns to make predictions about future states. In this way, a chatbot that is fed examples of text chats can learn to produce life like exchanges with people, or an image recognition tool can learn to identify and describe objects in images by reviewing millions of examples.

AI is important because it can give enterprises insights into their operations that they may not have been aware of previously and because, in some cases, AI can perform tasks better than humans. Particularly when it comes to repetitive, detail-oriented tasks like analyzing large numbers of legal documents to ensure relevant fields are filled in properly, AI tools often complete jobs quickly and with relatively few errors.

Artificial neural networks and deep learning artificial intelligence technologies are quickly evolving, primarily because AI processes large amounts of data much faster and makes predictions more accurately than humanly possible.

1.2 SCOPE OF THE PROJECT

While mining this data set through normal EDA process I came across the fact that not all earthquakes are natural and few are indeed caused by humans although very small in numbers.

At the end of the analysis I have tried to predict earthquakes and other quakes (seismicactivities related to explosion, quarry blast etc.). I have also tried to handle the class imbalance problem because the data set is 98:2.

The next steps are pretty usual ones with loading and probing the data. Let's get started with loading the libraries first.

1.3 OBJECTIVE OF THE PROJECT

Exploration data analysis of variable identification

- Loading the given dataset
- Import required libraries packages
- Analyze the general properties
- Find duplicate and missing values
- Checking unique and count value.
- Rename, add data and drop the data
- To specify data type
- Exploration data analysis of bi-variate and multi-variate
- Plot diagram of pairplot, heatmap, bar chart and Histogram
- Comparing algorithm to predict the result Based on the best accuracy

1.4 PROBLEM DEFINITION

The goal is to develop a machine learning model for Earth Quake prediction,to potentially replace the updatable supervised machine learning classification models by predicting results in the form of best accuracy by comparing supervised algorithm.

CHAPTER 2

LITERATURE SURVEY

INTRODUCTION

A literature review is a body of text that aims to review the critical points of current knowledge on and/or methodological approaches to a particular topic.

2.1 Earthquake Early Warning System using Ncheck and hard Orthogonal Multi Target Regression in Deeplearning

Author : Adi Wibowo , Member, IEEE, C. Pratama, David P. Sahara, L. S. Heliani, S. Rasyid,Zharfan Akbar, Faiz Muttaqy, and Ajat Sudrajat

Year : 2021

PTWP and MTR tasks for the EEWS to determine earthquake source parameters in real time. The N-check algorithm improves window prediction selection to reduce false alarms in multistation waveforms that have noise, and MTR with hard-shared orthogonal are proven to improve earthquake parameter determination performance. The response time that is the prediction need to be improved. Classification and confusion matrix are not calculated.

2.2 Earthquake predictions and scientific forecast: dangers and opportunities for a technical and anthropological perspective

Author : Edgar Tapia-Hernández, Elizabeth A. Reddy and Laura J. Oros-Avilés

Year :2019

Supporting earthquake risk management with clear seismic communication may necessitate encounters with various popular misapprehensions regarding earthquake prediction. Drawing on technical data as well as insights from anthropology and economics, this paper addresses common and scientifically-unsupported ideas about earthquake prediction, as well as the state of science-based studies regarding statistical forecasting and physical precursors.Low computation speed.

2.3 Automatic Arrival Time Detection for Earthquakes Based on Stacked Denoising Autoencoder

Author : O.M. Saad, K. Inoue, A. Shalaby, L. Samy, and M. S. Sayed,

Year : 2018

The accurate detection of P-wave arrival time is imperative for determining the hypocenter location of an earthquake. However, precise detection of onset time becomes more difficult when the signal-to-noise ratio (SNR) of the seismic data is low, such as during microearthquakes. The Algorithm can pick arrival times accurately for weak SNR seismic data with SNR higher than -14 dB .Reliable issuing of alarms of imminent large earthquakes appears to be effectively impossible.

2.4 Understanding post_earthquake decision on multi storey concrete building in Christ Church.

Author : F.Marquis, J.J.Kim, K.J.ElWood,S.E.Chang

Year : 2017

The earthquake sequence has revealed unique issues and complexities for the owners of commercial and multi-storey residential buildings in relation to unexpected technical, legal, and financial challenges when making decisions regarding the future of their buildings impacted by the earthquakes. The paper presents a framework to understand the factors influencing post-earthquake decisions (repair or demolish) on multi-storey concrete buildings in Christchurch. The study, conducted in 2014, includes in-depth investigations on 15 case-study buildings using 27 semi-structured interviews with various property owners, property managers, insurers, engineers, and government authorities in New Zealand. It does not handle a complex problem.

2.5 Simplified Seismic loss function for suspended ceilings and drywall partitions.

Author : R.P.Dhakal, A.Pourali. S.K.Sha.

Year : 2016

Generalized loss functions for two important NSCs commonly used namely suspended ceilings and drywall partitions are developed in this study. The methodology to develop the loss functions, in the form of engineering demand parameter vs. expected loss due to the considered components, is based on the existing framework for the storey level loss estimation. The results confirm the accuracy of the proposed generic seismic loss functions. Less accuracy with high-dimensional data.

2.6 Towards the "Ultimate Earthquake-proof" Building: Development of an Integrated low-Damage System.

Author :S.Pampanian

Year : 2015

The 2010–2011 Canterbury earthquake sequence has highlighted the severe mismatch between societal expectations over the reality of seismic performance of modern buildings. A paradigm shift in performance-based design criteria and objectives towards damage-control or low-damage design philosophy and technologies is urgently required. The wider acceptance and implementation of cost-efficient damage-control (or low- damage) technologies. Learners have to wait for a response.

2.7 A Review of Application of Data Mining in Earthquake Prediction

Author : G.V. Otari, Dr. R.V. Kulkarni

Year : 2015

Data mining techniques can also be used for prediction of these natural hazards. 16 journal articles on the subject published between 1989 and 2011 was analyzed. The main data mining techniques used for earthquake prediction are logistic models, neural networks, the Bayesian belief network, and decision trees, all of which provide primary solutions to the problems inherent in the prediction of earthquakes, tsunamis, landslides and other micro seismic activities. This paper also aims to encourage additional research on topics, and concludes with several suggestions for further research of checking missing values of data frame.

CHAPTER 3

SYSTEM ANALYSIS

3.1 EXISTING SYSTEM

They proposed two methods and named them as picking target window prediction PTWP and multitarget regression (MTR) tasks to determine earthquake source parameters. The N-check algorithm improves window prediction selection to reduce false alarms in multistation waveforms that have noise, and MTR with hard-shared orthogonal are proven to improve earthquake parameter determination performance. Our system can provide reliable earthquake parameters. The three stations with three-component seismogram traces are represented in red, green, and blue components to form pixels in a row in one frame in a 10 s window. The sampling rate at each station varied between 20 and 25 Hz and was then normalized to 20 Hz. They used a band pass filter to minimize noise and normalize each stream by dividing its absolute peak amplitude. The data set has high noise for 506 seismic events and has a peak SNR of less than 50 dB.

Disadvantages:

1. The response time that is the prediction need to be improved.
2. Classification and confusion matrix are not calculated.

3.2 PROPOSED SYSTEM

Earthquakes are a natural disaster that can cause a lot of damage to both lives and properties. The machine learning is applied to every field where the dataset can be used to learn patterns and then from that pattern the prediction can be done. Our objective is to build a machine learning model that uses the past earthquake related dataset the data is pre-processed by using variable identification that is finding the dependent and independent variables after that the data is used to train the model by using machine learning libraries. Different algorithms are used to compare the model and the performance metrics are calculated and evaluated.

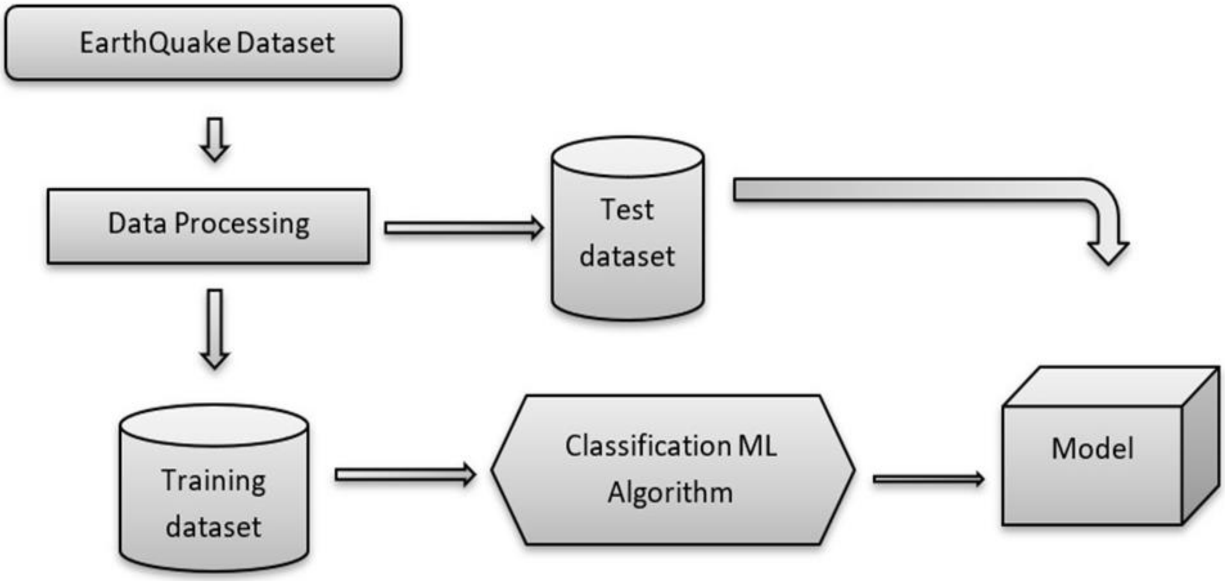


Fig 3.1 Architecture of proposed model

Advantages:

1. It handle complex problems, provide computational efficiency, propagate and treat uncertainties.

2. Damage can be controlled by early prediction of earthquakes

3.3 FEASIBILITY STUDY

3.3.1 Technical:

1. Software Requirements:

Operating System : Windows

Tool : Anaconda with Jupyter Notebook

2. Hardware requirements:

Processor : Intel Core i5

Hard disk : minimum 80 GB

RAM : 4 GB

3.3.2 Social:

1. Predicting the occurrence of Earthquake early can save the lives of many people and their valuable properties.

2. It also enables Government undertake precautionary measures so that the effects and consequences are minimized.

3.3.3 Economic:

Accuracy, Precision and Recall are calculated using the following predefined formulas:

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Precision = TP / (TP + FP)

Recall = TP / (TP + FN)

Where,

TP -> True Positives

TN ->True Negatives

FP ->False Positives

FN ->False Negatives

Score=2*(Recall*Precision)/(Recall + Precision)

Accuracy works best if false positives and false negatives have similar cost.

3.4 SYSTEM REQUIREMENTS

Requirements are the basic constrains that are required to develop a system. Requirements are collected while designing the system. The following are the requirements that are to be discussed.

Functional requirements

Non-Functional requirements

Environment requirements

A. Hardware requirements

B. software requirements

3.4.1 Functional requirements:

The software requirements specification is a technical specification of requirements for the software product. It is the first step in the requirements analysis process.

It lists requirements of a particular software system. The following details to follow the special libraries like sk-learn, pandas, numpy, matplotlib and seaborn.

3.4.2 Non-Functional Requirements:

Process of functional steps,

1. Problem define
2. Preparing data
3. Evaluating algorithms
4. Improving results
5. Prediction the result

3.4.3 Environmental Requirements:

1. Software Requirements:

Operating System	: Windows
Tool	: Anaconda with Jupyter Notebook

2. Hardware requirements:

Processor	: Pentium IV/III
Hard disk	: minimum 80 GB
RAM	: minimum 2 GB

3.5 SOFTWARE DESCRIPTION

Anaconda is a free and open-source distribution of the Python and R programming languages for scientific computing (data science, machine learning applications, large-scale data processing, predictive analytics, etc.), that aims to simplify package management and deployment. Package versions are managed by the package management system “Conda”. The Anaconda distribution is used by over 12 million users and includes more than 1400 popular data-science packages suitable for Windows, Linux, and MacOS. So, Anaconda distribution comes with more than 1,400 packages as well as the Conda package and virtual environment manager called Anaconda Navigator and it eliminates the need to learn to install each library independently. The open source packages can be individually installed from the Anaconda repository with the conda install command or using the pip install command that is installed with Anaconda.

3.5.1 ANACONDA NAVIGATOR

Anaconda Navigator is a desktop graphical user interface (GUI) included in Anaconda distribution that allows you to launch applications and easily manage conda packages, environments, and channels without using command-line commands.

Navigator can search for packages on Anaconda.org or in a local Anaconda Repository.

Anaconda. Now, if you are primarily doing data science work, Anaconda is also a great option. Anaconda is created by Continuum Analytics, and it is a Python distribution that comes preinstalled with lots of usefulpython libraries for data science. Anaconda is a distribution of the Python and R programming languages for scientific computing (data science, machine learning applications, large-scale data processing, predictive analytics, etc.), that aims to simplify package management and deployment.

Navigator is an easy, point-and-click way to work with packages and environments without needing to type conda commands in a terminal window. You can use it to find the packages you want, install them in an environment, run the packages, and update them – all inside Navigator.

The following applications are available by default in Navigator:

- JupyterLab
- Jupyter Notebook
- Spyder
- PyCharm
- VSCode
- Glueviz
- Orange 3 App
- RStudio
- Anaconda Prompt (Windows only)
- Anaconda PowerShell (Windows)

Anaconda Navigator is a desktop graphical user interface (GUI) included in Anaconda distribution.Navigator allows you to launch common Python programs and easily manage conda packages, environments, and channels without using command-line commands. Navigator can search for packages on Anaconda Cloud or in a local Anaconda Repository.

Anaconda comes with many built-in packages that you can easily find with conda list on your anaconda prompt. As it has lots of packages (many of which are rarely used), it requires lots of space and time as well. If you have enough space, time and do not want to burden yourself to install small utilities like JSON, YAML, you better go for Anaconda.

3.5.2 JUPYTER NOTEBOOK

This website acts as “meta” documentation for the Jupyter ecosystem. It has a collection of resources to navigate the tools and communities in this ecosystem, and to help you get started.

Project Jupyter is a project and community whose goal is to "develop open-source software, open-standards, and services for interactive computing across dozens of programming languages". It was spun off from IPython in 2014 by Fernando.

Launching Jupyter Notebook App: The **Jupyter Notebook App** can be launched by clicking on the Jupyter Notebook icon installed by Anaconda in the start menu (Windows) or by typing in a terminal (cmd on Windows): “jupyter notebook”

This will launch a new browser window (or a new tab) showing the **Notebook Dashboard**, a sort of control panel that allows (among other things) to select which notebook to open.

When started, the **Jupyter Notebook App** can access only files within its start-up folder (including any sub-folder). No configuration is necessary if you place your notebooks in your home folder or subfolders. Otherwise, you need to choose a **Jupyter Notebook App s**tart-up folder which will contain all the notebooks**.**

Save notebooks: Modifications to the notebooks are automatically saved every few minutes. To avoid modifying the original notebook, make a copy of the notebook document (menu file -> make a copy…) and save the modifications on the copy.

Executing a notebook: Download the notebook you want to execute and put it in your notebook folder (or a sub-folder of it).

- Launch the jupyter notebook app
- In the Notebook Dashboard navigate to find the notebook: clicking on name will open it in a new browser tab.
- Click on the menu Help -> User Interface Tour for an overview of the Jupyter Notebook App user interface.
- You can run the notebook document step-by-step (one cell a time) by pressing shift + enter.

File Extension:

An IPYNB file is a notebook document created by Jupyter Notebook, an interactive computational environment that helps scientists manipulate and analyze data using Python**.**

The Jupyter Notebook App is a server-client application that allows editing and running notebook documents via a web browser.

Working Process:

Download and install anaconda and get the most useful package for machine learning in Python.

- Load a dataset and understand its structure using statistical summaries and data visualization.
- Machine learning models, pick the best and build confidence that the accuracy is reliable.

Python is a popular and powerful interpreted language. Unlike R, Python is a complete language and platform that you can use for both research and development and developing production systems. There are also a lot of modules and libraries to choose from, providing multiple ways to do each task. It can feel overwhelming.

The best way to get started using Python for machine learning is to complete a project.

- It will force you to install and start the Python interpreter (at the very least).
- It will give you a bird's eye view of how to step through a small project.
- It will give you confidence, maybe to go on to your own small projects.

When you are applying machine learning to your own datasets, you are working on a project. A machine learning project may not be linear, but it has a number of well-known steps:

- Define problem
- Prepare Data.
- Evaluate Algorithms.
- Improve Results.
- Present Results.

The best way to really come to terms with a new platform or tool is to work through a machine learning project end-to-end and cover the key steps. Namely, from loading data, summarizing data, evaluating algorithms and making some predictions.

Here is an overview of what we are going to cover:

1. Installing the Python anaconda platform.
2. Loading the dataset.
3. Summarizing the dataset.
4. Visualizing the dataset.
5. Evaluating some algorithms.
6. Making some predictions.

3.5.3 PYTHON

Introduction:

Python is an interpreted high-level general-purpose programming language. Its design philosophy emphasizes code readability with its use of significant indentation. Its language constructs as well as its object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects.

Python is dynamically-typed and garbage-collected. It supports multiple programming paradigms, including structured (particularly, procedural), object-oriented and functional programming. It is often described as a "batteries included" language due to its comprehensive library.

CHAPTER 4

SYSTEM DESIGN

4.1 ER DIAGRAM

An entity relationship diagram (ERD), also known as an entity relationship model, is a graphical representation of an information system that depicts the relationships among people, objects, places, concepts or events within that system. An ERD is a data modeling technique that can help define business processes and be used as the foundation for a relational database. Entity relationship diagrams provide a visual starting point for database design that can also be used to help determine information system requirements throughout an organization. After a relational database is rolled out, an ERD can serve as a referral point, should any debugging or business process re-engineering be needed later.

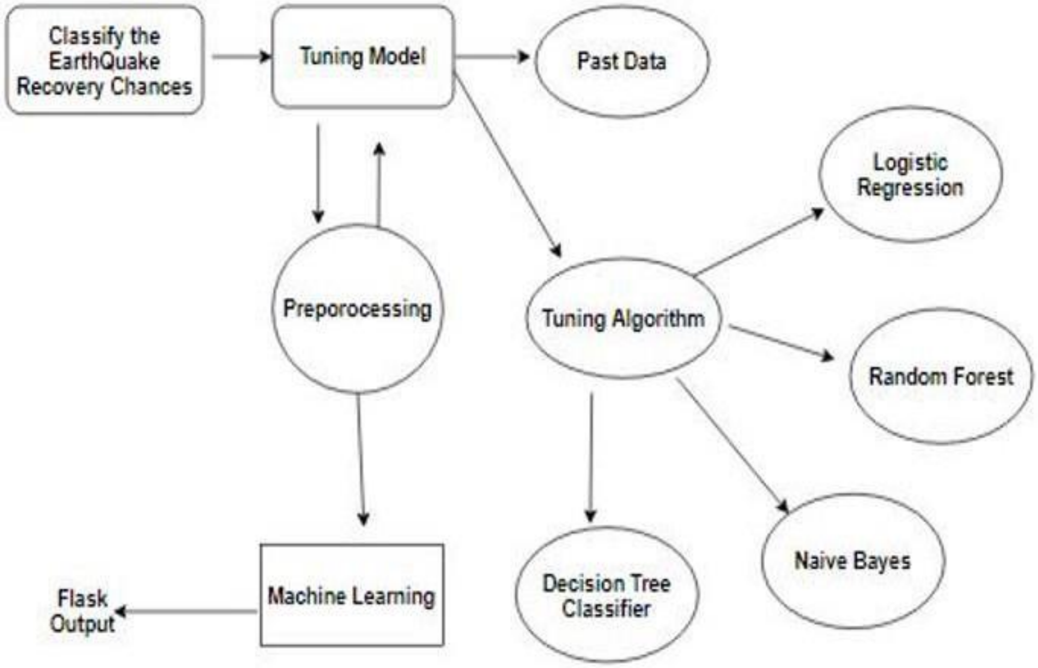


Fig 4.1 ER Diagram for earthquake prediction

4.2 DATAFLOW DIAGRAM

A data-flow diagram is a way of representing a flow of data through a process or a system (usually an information system). The DFD also provides information about the outputs and inputs of each entity and the process itself. A data-flow diagram has no control flow, there are no decision rules and no loops. There are several notations for displaying data-flow diagrams. For each data flow, at least one of the endpoints (source and / or destination) must exist in a process. The refined representation of a process can be done in another data-flow diagram, which subdivides this process into sub-processes. The 18 dataflow diagram is part of the structured-analysis modeling tools. When using UML, the activity diagram typically takes over the role of the data-flow diagram. A special form of data-flow plan is a site-oriented data-flow plan.

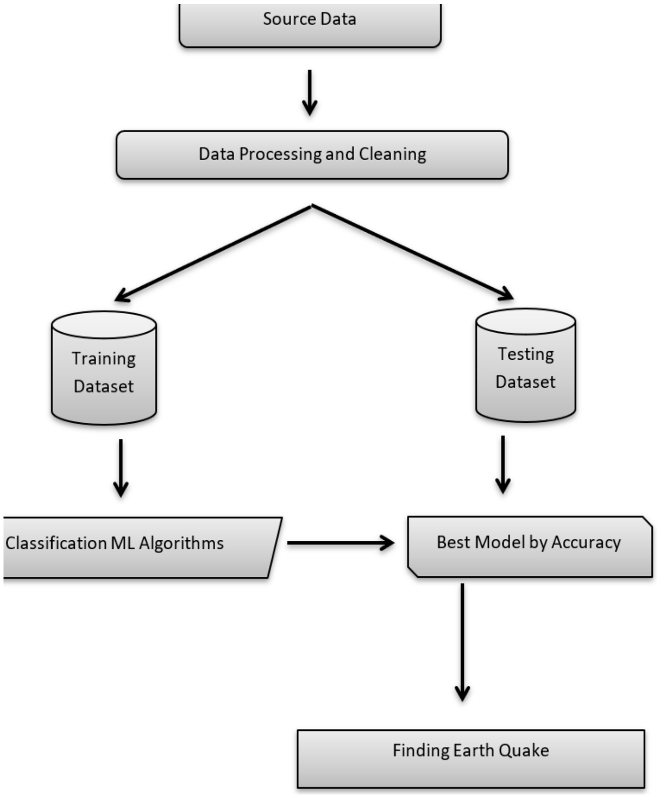


Fig 4.2 DFD Diagram for earthquake prediction

4.3 USECASE DIAGRAM

Use case diagrams are considered for high level requirement analysis of a system. So when the requirements of a system are analyzed the functionalities are captured in use cases. So, it can say that uses cases are nothing but the system functionalities written in an organized manner

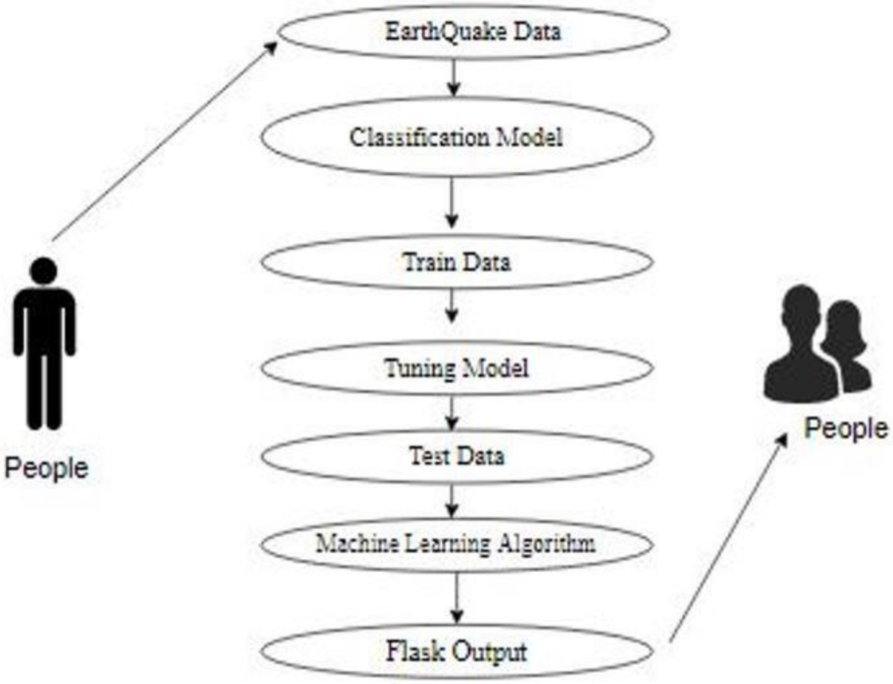


Fig 4.3 Use case Diagram for earthquake prediction

4.4 CLASS DIAGRAM

Class diagram is basically a graphical representation of the static view of the system and represents different aspects of the application. So a collection of class diagrams represent the whole system. The name of the class diagram should be meaningful to describe the aspect of the system. Each element and their relationships should be identified in advance Responsibility (attributes and methods) of each class should be clearly identified for each class minimum number of properties should be specified and because, unnecessary properties will make the diagram complicated. Use notes whenever required to describe some aspect of the diagram and at the end of the drawing it should be understandable to the developer/coder. Finally, before making the final version, the diagram should be drawn on plain paper and rework as many times as possible to make it correct.

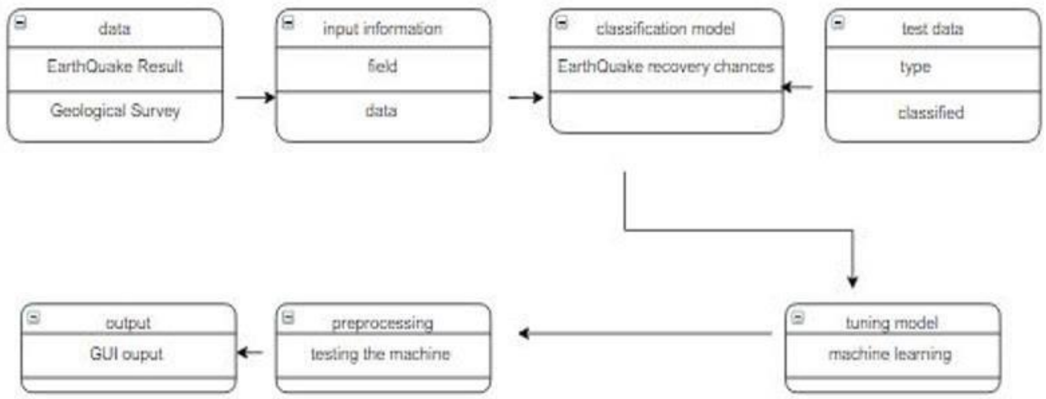


Fig 4.4 Class diagram for earthquake prediction

4.5 ACTIVITY DIAGRAM

Activity is a particular operation of the system. Activity diagrams are not only used for visualizing dynamic nature of a system but they are also used to construct the executable system by using forward and reverse engineering techniques. The only missing thing in activity diagram is the message part. It does not show any message flow from one activity to another. Activity diagram is some time considered as the flow chart. Although the diagrams looks like a flow chart but it is not. It shows different flow like parallel, branched, concurrent and single.

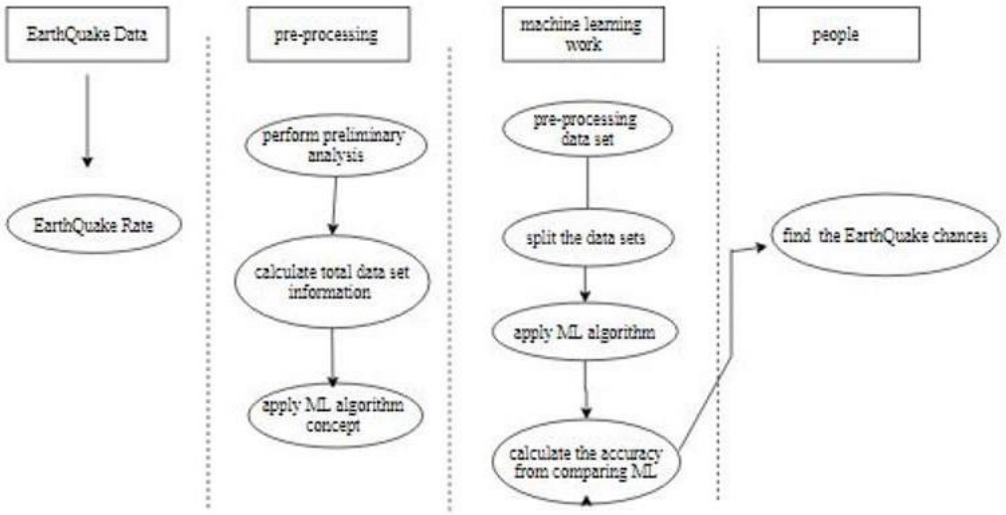

EarthQuake Data
pre-processing
machine learning work
people
EarthQuake Rate
perform preliminary analysis
calculate total data set information
apply ML algorithm concept
pre-processing data set
split the data sets
apply ML algorithm
calculate the accuracy from comparing ML
find the EarthQuake chances

Fig 4.5 Activity Diagram for earthquake prediction

4.6 SEQUENCE DIAGRAM

Sequence diagrams model the flow of logic within your system in a visual manner, enabling you both to document and validate your logic, and are commonly used for both analysis and design purposes. Sequence diagrams are the most popular UML artifact for dynamic modelling, which focuses on identifying the behaviour within your system. Other dynamic modelling techniques include activity diagramming, communication diagramming, timing diagramming, and interaction overview diagramming. Sequence diagrams, along with class diagrams and physical data models are in my opinion the most important design-level models for modern business application development

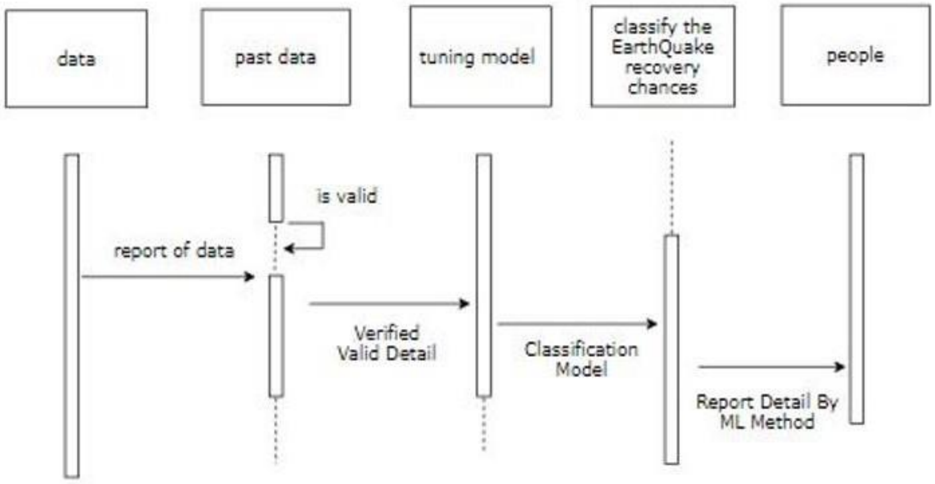


Fig 4.6 Sequence Diagram for earthquake prediction

CHAPTER 5

SYSTEM ARCHITECTURE

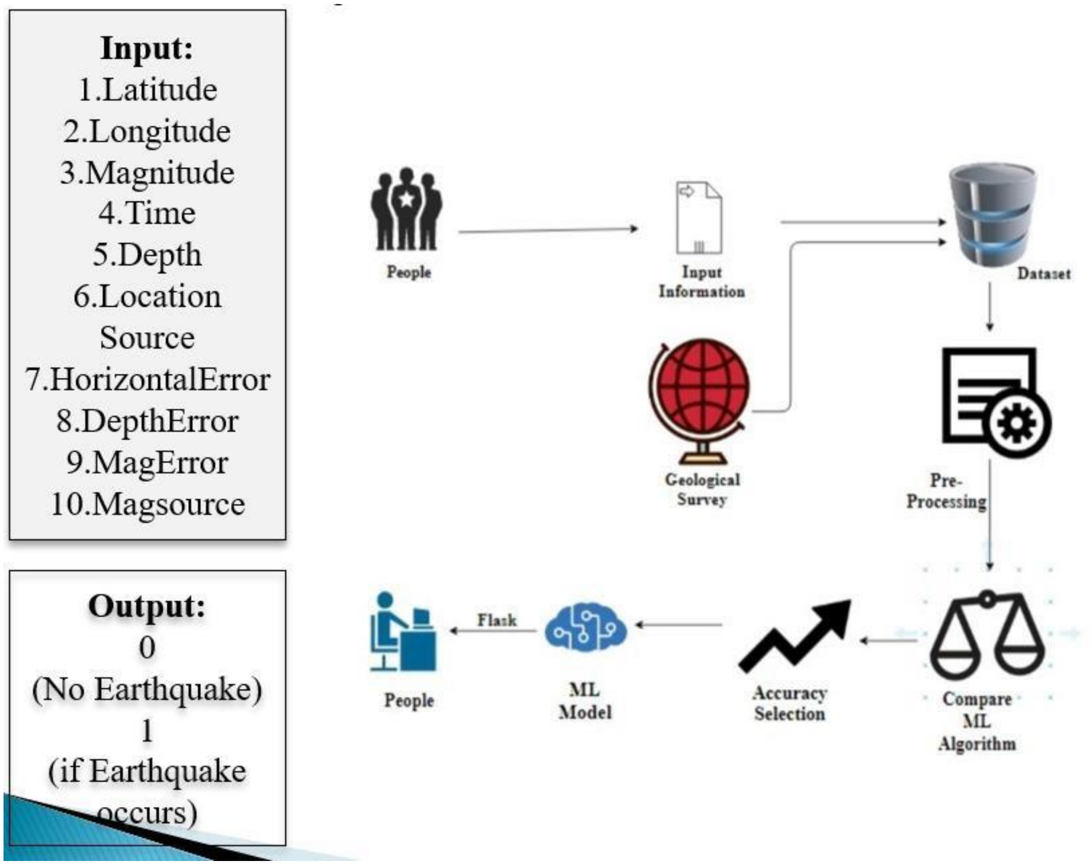


Fig 5.1 Architecture Diagram for earthquake prediction

5.1 MODULE DESCRIPTION

Data Pre-processing

Validation techniques in machine learning are used to get the error rate of the Machine Learning (ML) model, which can be considered as close to the true error rate of the dataset. If the data volume is large enough to be representative of the population, you may not need the validation techniques. However, in real-world scenarios, to work with samples of data that may not be a true representative of the population of given dataset. To finding the missing value, duplicate value and description of data type whether it is float variable or integer. The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyper parameters.

A number of different **data cleaning** tasks using Python's Pandas library and specifically, it focus on probably the biggest data cleaning task, **missing values** and it able to **more quickly clean data**. It wants to **spend less time cleaning data**, and more time exploring and modeling.

Some of these sources are just simple random mistakes. Other times, there can be a deeper reason why data is missing. It's important to understand these different types of missing data from a statistics point of view. The type of missing data will influence how to deal with filling in the missing values and to detect missing values, and do some basic imputation and detailed statistical approach for dealing with missing data. Before, joint into code, it's important to understand the sources of missing data. Here are some typical reasons why data is missing:

- User forgot to fill in a field.
- Data was lost while transferring manually from a legacy database.
- There was a programming error.
- Users chose not to fill out a field tied to their beliefs about how the results would be used or interpreted.

Variable identification with Uni-variate, Bi-variate and Multi-variate analysis:

- import libraries for access and functional purpose and read the given dataset
- General Properties of Analyzing the given dataset
- Display the given dataset in the form of data frame
- show columns
- shape of the data frame
- To describe the data frame

- Checking data type and information about dataset
- To specify the type of values
- To create extra columns

Data Validation/ Cleaning/Preparing Process

Importing the library packages with loading given dataset. To analyzing the variable identification by data shape, data type and evaluating the missing values, duplicate values. A validation dataset is a sample of data held back from training your model that is used to give an estimate of model skill while tuning model's and procedures that you can use to make the best use of validation and test datasets when evaluating your models. Data cleaning / preparing by rename the given dataset and drop the columnetc. to analyze the uni-variate, bi-variate and multi-variate process. The steps and techniques for data cleaning will vary from dataset to dataset. The primary goal of data cleaning is to detect and remove errors and anomalies to increase the value of data in analytics and decision making.

MODULE DIAGRAM

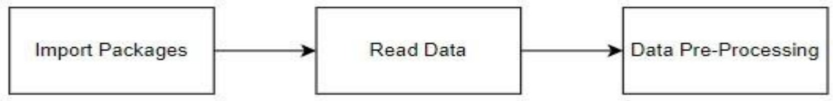


Fig 5.2 Module diagram for data pre-processing

GIVEN INPUT EXPECTED OUTPUT

input : data
output:removing noisy data.

Exploration data analysis of visualization

Data visualization is an important skill in applied statistics and machine learning. Statistics does indeed focus on quantitative descriptions and estimations of data. Data visualization provides an important suite of tools for gaining a qualitative understanding. This can be helpful when exploring and getting to know a dataset and can help with identifying patterns, corrupt data, outliers, and much more. With a little domain knowledge, data visualizations can be used to express and demonstrate key relationships in plots and charts that are more visceral and stakeholders than measures of association or significance. Data visualization and exploratory data analysis are whole fields themselves and it will recommend a deeper dive into some the books mentioned at the end.

MODULE DIAGRAM

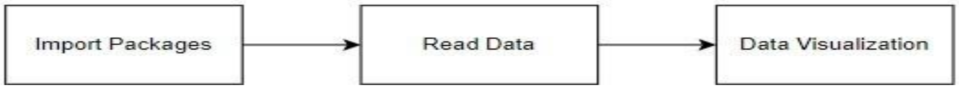


Fig 5.4 Module diagram for data visualization

GIVEN INPUT EXPECTED OUTPUT

input : data
output : visualized data

Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. Data Preprocessing is a technique that is used to convert the raw data into a clean data set Some specified Machine Learning model needs information in a specified format, for example, Random Forest algorithm does not support null values. Therefore, to execute random forest algorithm null values have to be managed from the original raw data set. And another aspect is that data set should be formatted in such a way that more than one Machine Learning and Deep Learning algorithms are executed in given dataset.

False Positives (FP): A person who will pay predicted as defaulter. When actual class is no and predicted class is yes. E.g. if actual class says this passenger did not survive but predicted class tells you that this passenger will survive.

False Negatives (FN): A person who default predicted as payer. When actual class is yes but predicted class in no. E.g. if actual class value indicates that this passenger survived and predicted class tells you that passenger will die.

True Positives (TP): A person who will not pay predicted as defaulter. These are the correctly predicted positive values which means that the value of actual class is yes and the value of predicted class is also yes. E.g. if actual class value indicates that this passenger survived and predicted class tells you the same thing.

True Negatives (TN): A person who default predicted as payer. These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no. E.g. if actual class says this passenger did not survive and predicted class tells you the same thing.

Comparing Algorithm with prediction in the form of best accuracy result

It is important to compare the performance of multiple different machine learning algorithms consistently and it will discover to create a test harness to compare multiple different machine learning algorithms in Python with scikit-learn. It can use this test harness as a template on your own machine learning problems and add more and different algorithms to compare. Each model will have different performance characteristics. Using resampling methods like cross validation, you can get an estimate for how accurate each model may be on unseen data. It needs to be able to use these estimates to choose one or two best models from the suite of models that you have created. When have a new dataset, it is a good idea to visualize the data using different techniques in order to look at the data from different perspectives. The same idea applies to model selection. You should use a number of different ways of looking at the estimated accuracy of your machine learning algorithms in order to choose the one or two to finalize. A way to do this is to use different visualization methods to show the average accuracy, variance and other properties of the distribution of model accuracies.

In the example below 4 different algorithms are compared:

- Random Forest
- Logistic Regression
- Naïve Bayes
- Decision Tree Classifier

The K-fold cross validation procedure is used to evaluate each algorithm, importantly configured with the same random seed to ensure that the same splits to the training data are performed and that each algorithm is evaluated in precisely the same way. Before that comparing algorithm, Building a Machine Learning Model using install Scikit-Learn libraries. In this library package have to done preprocessing, linear model with logistic regression method, cross validating by KFold method, ensemble with random forest method and tree with decision tree classifier. Additionally, splitting the train set and test set. To predicting the result by comparing accuracy.

Prediction result by accuracy:

Logistic regression algorithm also uses a linear equation with independent predictors to predict a value. The predicted value can be anywhere between negative infinity to positive infinity. It need the output of the algorithm to be classified variable data. Higher accuracy predicting result is logistic regression model by comparing the best accuracy.

True Positive Rate(TPR) = TP / (TP + FN)

False Positive rate(FPR) = FP / (FP +TN)

Accuracy calculation:

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. One may think that, if we have high accuracy then our model is best. Yes, accuracy is a great measure but only when you have symmetric datasets where values of false positive and false negatives are almost same.

Precision: The proportion of positive predictions that are actually correct.

Precision = TP / (TP + FP)

Recall: The proportion of positive observed values correctly predicted.(The proportion of actual defaulters that the model will correctly predict)

Recall = TP / (TP + FN)

F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it’s better to look at both Precision and Recall.

General Formula:

F- Measure = 2TP / (2TP + FP + FN)

F1-Score Formula:

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

5.2ALGORITHM AND TECHNIQUES

Algorithm Explanation

In machine learning and statistics, classification is a supervised learning approach in which the computer program learns from the data input given to it and then uses this learning to classify new observation. This data set may simply be bi-class (like identifying whether the person is male or female or that the mail is spam or non-spam) or it may be multi-class too. Some examples of classification problems are: speech recognition,

Used Python Packages:

sklearn:

- In python, sklearn is a machine learning package which include a lot of ML algorithms.
- Here, we are using some of its modules like train_test_split, DecisionTreeClassifier or Logistic Regression and accuracy_score.

NumPy:

- It is a numeric python module which provides fast maths functions for calculations.
- It is used to read data in numpy arrays and for manipulation purpose.

Pandas:

- Used to read and write different files.
- Data manipulation can be done easily with data frames.

Matplotlib:

- Data visualization is a useful way to help with identify the patterns from given dataset.
- Data manipulation can be done easily with data frames.

Logistic Regression

It is a statistical method for analysing a data set in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables. Logistic regression is a Machine

Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.).

In other words, the logistic regression model predicts P(Y=1) as a function of X. Logistic regression Assumptions:

- Binary logistic regression requires the dependent variable to be binary.For a binary regression, the factor level 1 of the dependent variable should repre- sent the desired outcome.
- The independent variables are linearly related to the log odds.
- Logistic regression requires quite large sample sizes.

MODULE DIAGRAM

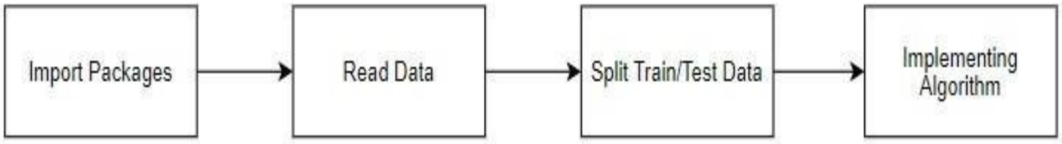


Fig 5.6 Module Diagram for Logistic Regression Algorithm implementation

GIVEN INPUT EXPECTED OUTPUT

input : data
output : getting accuracy

Random Forest Classifier

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set. Random forest is a type of supervised machine learning algorithm based on ensemble learning. Ensemble learning is a type of learning where you join different types of algorithms or same algorithm multiple times to form a more powerful prediction model.

The following are the basic steps involved in performing the random forest algorithm:

- Pick N random records from the dataset.
- Build a decision tree based on these N records.
- Choose the number of trees you want in your algorithm
- Repeat steps 1 and 2.
- In case of a regression problem, for a new record, each tree in the forest predicts a value for Y (output).
- The final value can be calculated by taking the average of all the values predicted by all the trees in forest.

MODULE DIAGRAM

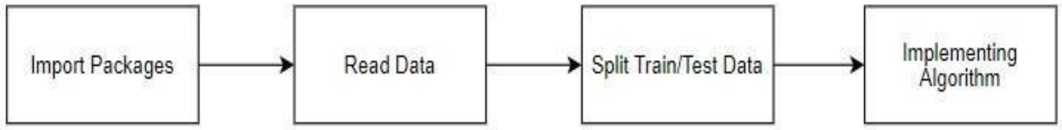


Fig 5.8 Module diagram for Random Forest Classifier Algorithm implementation

GIVEN INPUT EXPECTED OUTPUT

input : data
output : getting accuracy

Naive Bayes algorithm:

- The Naive Bayes algorithm is an intuitive method that uses the probabilities of each attribute belonging to each class to make a prediction. It is the supervised learning approach you would come up with if you wanted to model a predictive modeling problem probabilistically.
- Naive bayes simplifies the calculation of probabilities by assuming that the prob- ability of each attribute belonging to a given class value is independent of all other attributes. This is a strong assumption but results in a fast and effective method.
- The probability of a class value given a value of an attribute is called the condi- tional probability. By multiplying the conditional probabilities together for each attribute for a given class value, we have a probability of a data instance belonging to that class. To make a prediction we can calculate probabilities of the instance belonging to each class and select the class value with the highest probability.

MODULE DIAGRAM

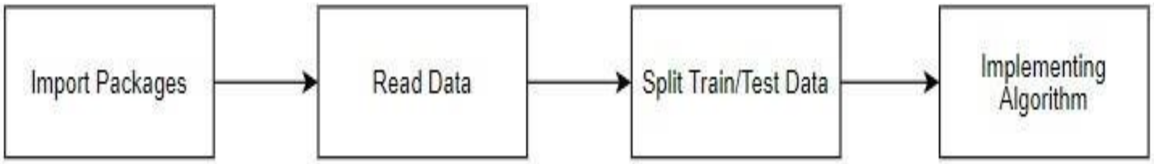


Fig 5.10 Module Diagram for Naive Bayes Algorithm implementation

GIVEN INPUT EXPECTED OUTPUT

input : data
output : getting accuracy

Decision Tree

It is one of the most powerful and popular algorithm. Decision-tree algorithm falls under the category of supervised learning algorithms

- At the beginning, we consider the whole training set as the root.
- Attributes are assumed to be categorical for information gain, attributes are as- sumed to be continuous.
- On the basis of attribute values records are distributed recursively.It utilizes an if-then rule set which is mutually exclusive and exhaustive for classification.

MODULE DIAGRAM

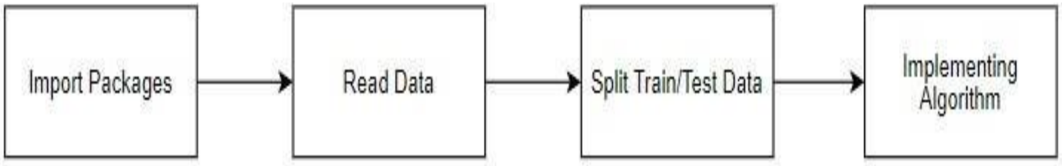


Fig 5.13 Module Diagram for Decision Tree algorithm implementation

GIVEN INPUT EXPECTED OUTPUT

input : data
output : getting accuracy

CHAPTER 6

SYSTEM IMPLEMENTATION

6.1 SERVER SIDE

Module - 1
Data-Validation & Pre-processing

```
import pandas as p
import numpy as n
import warnings
warnings.filterwarnings('ignore')
data = p.read_csv('demo1.csv')
data.head()
data.shape
data.columns
data.isnull().sum()
df = data.dropna()
df.shape
df.isnull().sum()
df.describe()
df.columns
df.magType.unique()
df.status.unique()
df.locationSource.unique()
df.magSource.unique()
df.net.unique()
p.Categorical(df['locationSource'])
p.Categorical(df['net']).describe()
p.Categorical(df['magSource']).describe()
p.Categorical(df['locationSource']).describe()
df.gap.unique()
print("Gap: ", sorted(df['gap'].unique()))
df.info()
df.duplicated()
sum(df.duplicated())
p.crosstab(df.horizontalError,df.depthError)
df.columns
print("Minimum value of Depth is:", df.depth.min())
print("Maximum value of Depth is:",
df.depth.max())print("Minimum value of Gap is:",
df.gap.min()) print("Maximum value of Gap is:",
df.gap.max()) print("Minimum value of Latitude is:",
df.latitude.min()) print("Maximum value of Latitude is:",
df.latitude.max()) print("Minimum value of Longitude is:",
df.longitude.min()) print("Maximum value of Longitude
is:", df.longitude.max()) print("Minimum value of Gap is:",
df.gap.min()) print("Maximum value of Gap is:",
df.gap.max()) p.Categorical(df['magType']).describe()
p.Categorical(df['status']).describe()
p.Categorical(df['locationSource']).describe()
df['mag'].value_counts()
df.corr()
from sklearn.preprocessing import LabelEncoder
var_mod = ['magType', 'status', 'locationSource']
le = LabelEncoder()
for i in var_mod:
df[i] = le.fit_transform(df[i]).astype(int)
df.head()
```

Module - 2
Visualization

```
import pandas as p
import matplotlib.pyplot as plt
import seaborn as s
import numpy as n
import warnings
warnings.filterwarnings('ignore')
data = p.read_csv('demo1.csv')
df = data.dropna()
from sklearn.preprocessing import LabelEncoder
var_mod = ['magType','status', 'locationSource']
le = LabelEncoder()
for i in var_mod:
    df[i] = le.fit_transform(df[i]).astype(str)
df.head()
df.columns
df['gap'].hist(figsize=(5,5), color='orange', alpha=1)
plt.xlabel('gap')
plt.ylabel('depth') plt.title('gap & depth')
df['magError'].hist(figsize=(5,5), color='r', alpha=1)
plt.xlabel('magError')
plt.ylabel('depthError')
plt.title('magError & depthError')
plt.boxplot(df['latitude'])
plt.show()
import seaborn as s
s.boxplot(df['latitude'], color='m')
fig, ax = plt.subplots(figsize=(16,8))
ax.scatter(df['gap'],df['depth'])
ax.set_xlabel('Gap')
ax.set_ylabel('Depth')
plt.show()
def PropByVar(df, variable):
dataframe_pie = df[variable].value_counts()
ax = dataframe_pie.plot.pie(figsize=(10,10), autopct='%1.2f%%', fontsize = 12)
ax.set_title(variable + ' \n', fontsize = 15)
return n.round(dataframe_pie/df.shape[0]*100,2)
PropByVar(df, 'nst')
fig, ax = plt.subplots(figsize=(15,10))
s.heatmap(df.corr(), ax=ax, annot=True)
plt.plot(df["gap"], df["depth"], color='g')
plt.xlabel('gap')
plt.ylabel('depth')
plt.title('EarthQuake')
plt.show()
df.columns
X = df.drop(labels='status', axis=1)
y = df.loc[:,'status']
print("Number of training dataset: ", len(X_train))
print("Number of test dataset: ", len(X_test))
print("Total number of dataset: ", len(X_train)+len(X_test))
```

Module - 3
Random Forest Algorithm

```
import pandas as p
import matplotlib.pyplot as plt
import seaborn as s
import numpy as n
import warnings
warnings.filterwarnings('ignore')
data=p.read_csv('demo1.csv')
df = data.dropna()
df.tail(10)
del df['time']
del df['updated']
del df['place']
del df['net']
del df['magSource']
del df['type']
del df['id']
df.status.unique()
from sklearn.preprocessing import LabelEncoder
var_mod = ['magType','status', 'locationSource']
le = LabelEncoder()
for i in var_mod:
df[i] = le.fit_transform(df[i]).astype(int)
df.status.unique()
X = df.drop(labels='status', axis=1)
y = df.loc[:,'status']
print("Number of training dataset: ", len(X_train))
print("Number of test dataset: ", len(X_test))
print("Total number of dataset: ", len(X_train)+len(X_test))
from sklearn.metrics import confusion_matrix, classification_report, matthews_corrcoef, cohen_kappa_score, accuracy_score, average_precision_score, roc_auc_score
from sklearn.metrics import accuracy_score, confusion_matrix
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import cross_val_score
rfc= RandomForestClassifier()
rfc.fit(X_train,y_train)
predictRF = rfc.predict(X_test)
```

CLIENT SIDE

HTML:

```
<!DOCTYPE html>
<html >
```

```
<!--From https://codepen.io/frytyler/pen/EGdtg-->
<head>
  <meta charset="UTF-8">
  <title>TITLE</title>
<link rel="stylesheet" href="{{ url_for('static',
filename='css/bootstrap.min.css') }}">
  <link href='https://fonts.googleapis.com/css?family=Pacifico' rel='stylesheet'
type='text/css'>
<link href='https://fonts.googleapis.com/css?family=Arimo' rel='stylesheet'
type='text/css'>
<link href='https://fonts.googleapis.com/css?family=Hind:300' rel='stylesheet'
type='text/css'>
<link href='https://fonts.googleapis.com/css?family=Open+Sans+Condensed:300'
rel='stylesheet' type='text/css'>
<style>
.back{
 background-image: url("{{ url_for('static', filename='image/EarthQuake.jpg') }}");
background-repeat:no-repeat;
background-size:cover;
}
.white{
color:white;
}
.nspace{
margin:15px 15px 30px 30px;
padding:9px 10px;
background: palegreen;
width:500px
}
.space{
margin:10px 30px;
padding:10px 10px;
background: palegreen;
width:500px
}
.gap{
padding:10px 20px;
}
</style>
</head>
<body >
 <div>
<div class="jumbotron">
<h1 style="text-align:center">PREDICTION OF EARTH QUAKE </h1>
</div>
<div class="back">
<!-- Main Input For Receiving Query to our ML  -->
<form class="form-group"
action="{{ url_for('predict')}}"method="post">
<div class="row">
<div class="gap col-md-6 ">
<label class="white" for="">LATITUDE</label>
<input type="number" class="space form-control" step="0.01" name="LATITUDE"
placeholder="LATITUDE" required="required"
/><br>              <label class="white"  for="">LONGITUDE</label>
<input type="number" class="space form-control" step="0.01" name="LONGITUDE"
placeholder="LONGITUDE" required="required" /><br>
<label class="white" for="">DEPTH</label>
<input type="number" class="space form-control" step="0.01" name="DEPTH"
placeholder="DEPTH" required="required" /><br>
<label class="white" for="">MAG</label>
<input type="number" class="space form-control" step="0.01" name="MAG"
placeholder="MAG" required="required" /><br>
<label class="white" for="">MAG_TYPE</label>
<select class="nspace form-control" name="MagType" id="MagType">
```

```
  <option value=1>ML</option>
  <option value=0>MD</option>
  <option value=2>MLR</option>
</select>
<label class="white" for="">NST</label>
<input type="number" class="space form-control" step="0.01" name="NST"
placeholder="NST" required="required" /><br>
<label class="white" for="">GAP</label>
<input type="number" class="space form-control" step="0.01" name="GAP"
placeholder="GAP" required="required" /><br>
</div>
<div class="gap col-md-6">
<label class="white" for="">DMIN</label>
<input type="number" class="space form-control" step="0.01" name="DMIN"
placeholder="DMIN" required="required" /><br>
<label class="white" for="">RMS</label>
<input type="number" class="space form-control" step="0.01" name="RMS"
placeholder="RMS" required="required" /><br>
<label class="white" for="">HORIZANTAL</label>
<input type="number" class="space form-control" step="0.01"
name="HORIZANTAL" placeholder="HORIZANTAL" required="required" /><br>
<label class="white" for="">DEPTH_ERROR</label>
<input type="number" class="space form-control" step="0.01"
name="DEPTH_ERROR" placeholder="DEPTH_ERROR" required="required"
/><br>
<label class="white" for="">MAG_ERROR</label>
<input type="number" class="space form-control" step="0.01" name="MAG_ERROR"
placeholder="MAG_ERROR" required="required" /><br>
<label class="white" for="">MAG_NST</label>
<input type="number" class="space form-control" step="0.01" name="MAG_NST"
placeholder="MAG_NST" required="required" /><br>
<label class="white" for="">LOCATION_SOURCE</label>
<select class="nspace form-control" name="LocationSource" id="LocationSource">
```

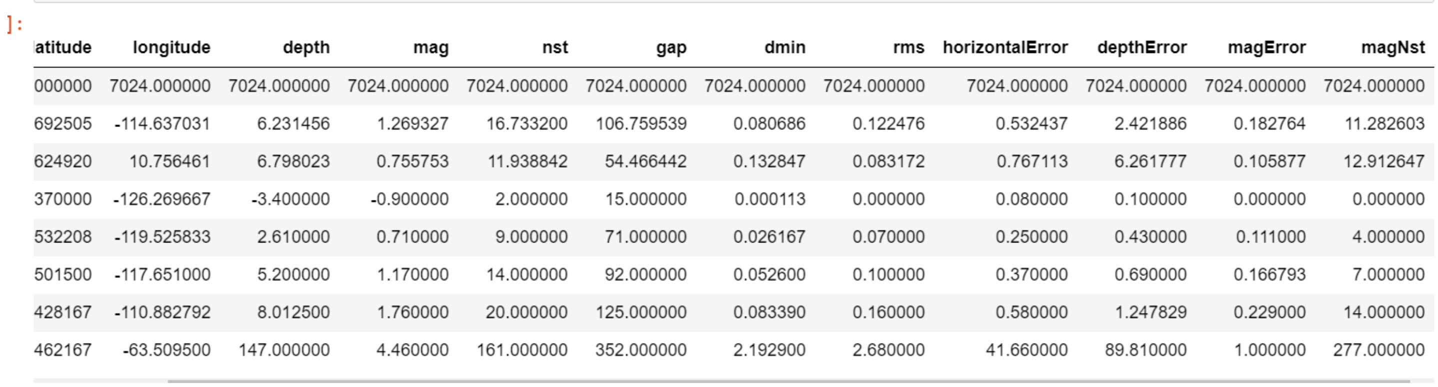
```
  <option value=0>CI</option>
  <option value=2>NC</option>
  <option value=8>UW</option>
  <option value=4>PR</option>
  <option value=6>TX</option>
  <option value=7>UU</option>
  <option value=1>MB</option>
  <option value=5>SE</option>
<option value=3>NM</option>
</select>
</div>
</div>
<div style="padding:2% 35%">
<button type="submit" class="btn btn-success btn-block"
style="width:350px;padding:20px">Predict</button>
</div>
</form>
</div>
<br>
<br>
<div style="background:skyblue;padding:2% 40%">
   {{ prediction_text }}
</div>
 </div>
</body>
</html>
```

FLASK DEPLOY:

```
import numpy as np
from flask import Flask, request, jsonify,  render_template
import pickle
import joblib
app = Flask(__name__)
model = joblib.load('rf.pkl')
@app.route('/')
def home():
return render_template('index.html')
@app.route('/predict',methods=['POST'])
def predict():
    '''
    For rendering results on HTML GUI
    '''
    int_features = [(x) for x in request.form.values()]
    final_features = [np.array(int_features)]
    print(final_features)
    prediction = model.predict(final_features)
    output = prediction[0]
    if output==1:
        output='happened'
    else:
        output="not happened"
    return render_template('index.html', prediction_text=' Earth Quake is
{}'.format(output))
if __name__ == "__main__":
app.run(host="localhost", port=5000)
```

CHAPTER 7
PERFORMANCE AND ANALYSIS

7.1 TEST CASE AND REPORT



]:

atitude	longitude	depth	mag	nst	gap	dmin	rms	horizontalError	depthError	magError	magNst
000000	7024.000000	7024.000000	7024.000000	7024.000000	7024.000000	7024.000000	7024.000000	7024.000000	7024.000000	7024.000000	7024.000000
692505	-114.637031	6.231456	1.269327	16.733200	106.759539	0.080686	0.122476	0.532437	2.421886	0.182764	11.282603
624920	10.756461	6.798023	0.755753	11.938842	54.466442	0.132847	0.083172	0.767113	6.261777	0.105877	12.912647
370000	-126.269667	-3.400000	-0.900000	2.000000	15.000000	0.000113	0.000000	0.080000	0.100000	0.000000	0.000000
532208	-119.525833	2.610000	0.710000	9.000000	71.000000	0.026167	0.070000	0.250000	0.430000	0.111000	4.000000
501500	-117.651000	5.200000	1.170000	14.000000	92.000000	0.052600	0.100000	0.370000	0.690000	0.166793	7.000000
428167	-110.882792	8.012500	1.760000	20.000000	125.000000	0.083390	0.160000	0.580000	1.247829	0.229000	14.000000
462167	-63.509500	147.000000	4.460000	161.000000	352.000000	2.192900	2.680000	41.660000	89.810000	1.000000	277.000000

Fig 7.1 Performance and Analysis for earthquake prediction

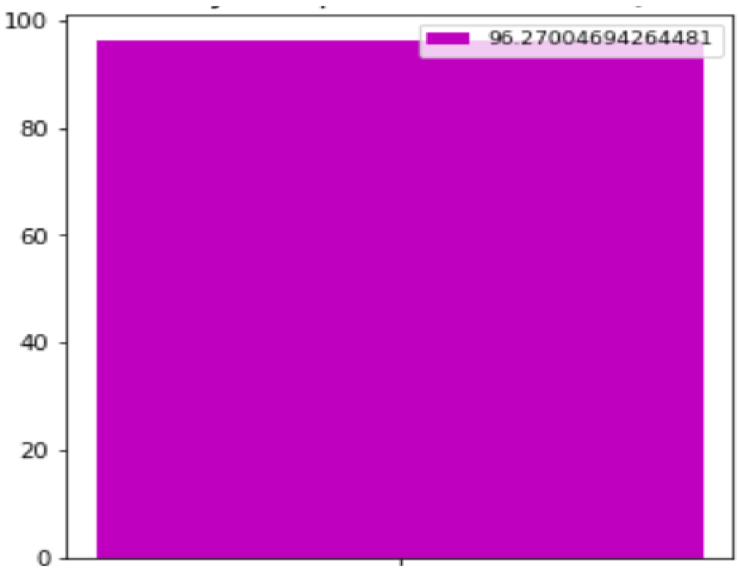


Fig 7.2 High accuracy algorithm to predict the earthquake

CHAPTER 8
CONCLUSION

8.1 RESULTS AND DISCUSSION

The analytical process started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation. The best accuracy on public test set is higher accuracy score is will be find out. This application can help to find the Prediction of Earth Quake.

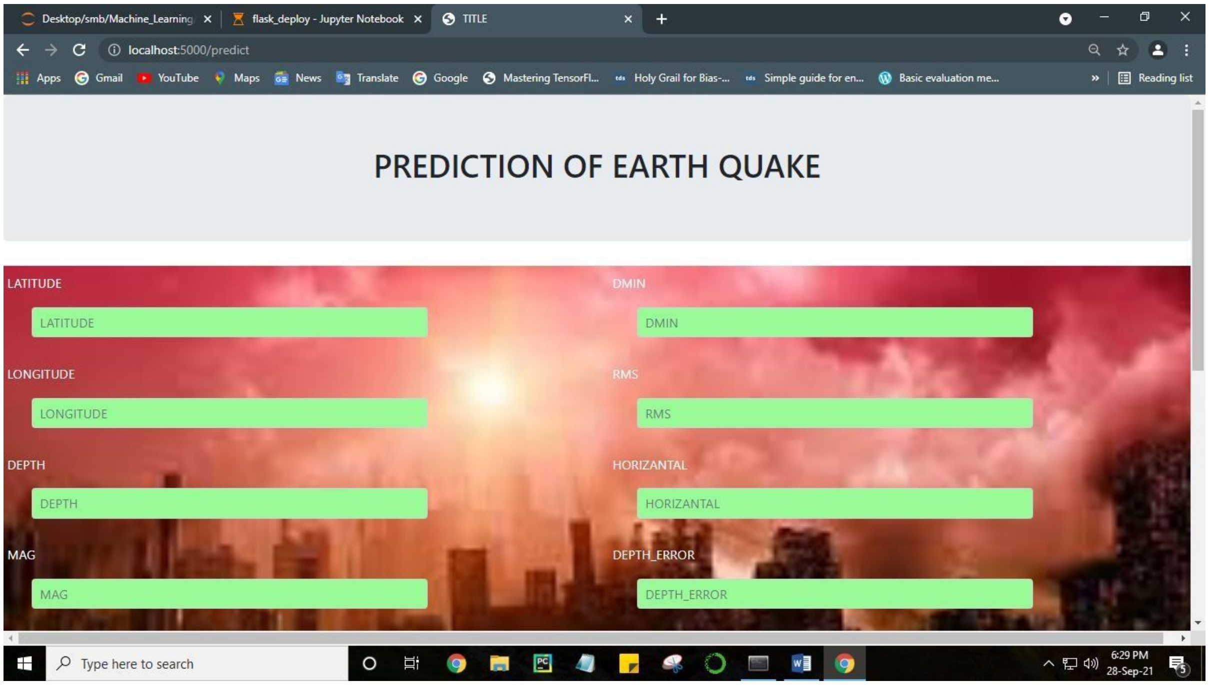
8.2 CONCLUSION AND FUTURE ENHANCEMENT

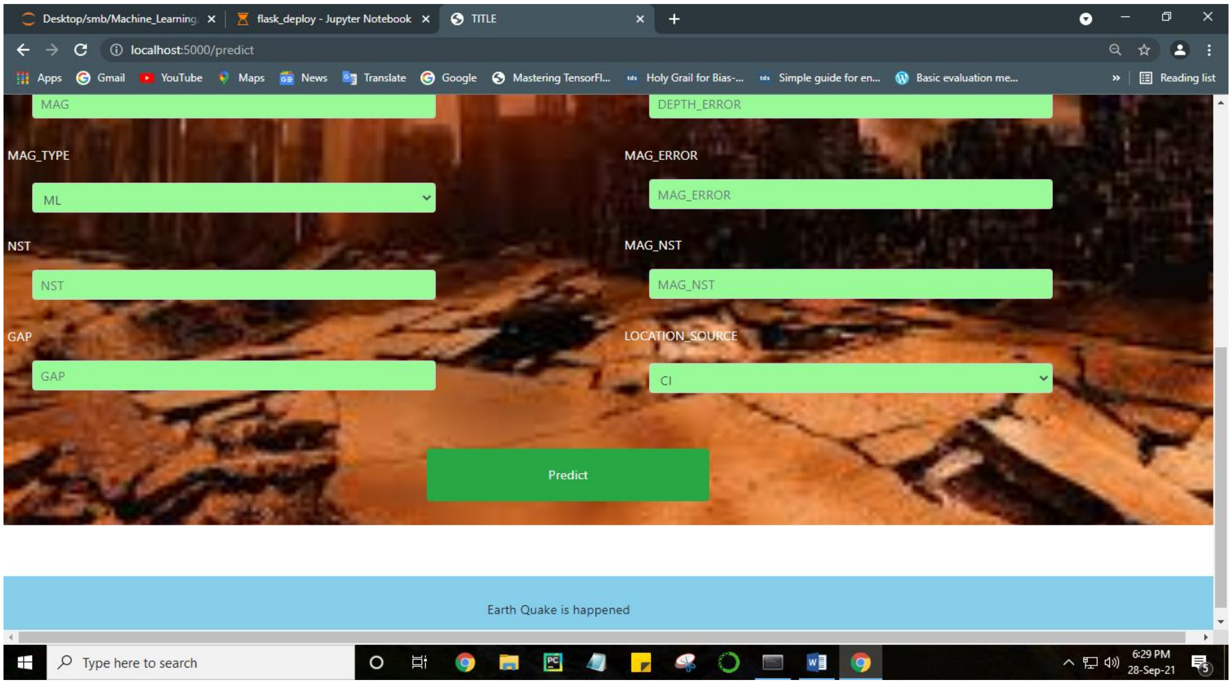
Therefore, with further optimization of the earthquake dataset and improvements in the algorithm,a computer-aided detection system can be expected to become an effective and efficient method of significant of Earthquake happens using Datascience concept the second module lenet gave the more accuracy compared to other modules.So by using lenet we train the machine for more accuracy.Hence we demonstrated that the DataScience and Machine learning Algorithm was useful for assessing the diagnosis and predictability.

- Earth Quake prediction to connect with AI model.
- To automate this process by show the prediction result in web application.
- To optimize the work to implement in Artificial Intelligence environment

APPENDICES

A.1 SAMPLE SCREENS





MAG_TYPE: ML

LOCATION_SOURCE: CI

Earth Quake is happened

Fig A.1 Sample Screens for earthquake prediction

REFRENCES

[1] C. Pratama, David P. Sahara, L. S. Heliani, S. Rasyid,Zharfan Akbar, Faiz Muttaqy, and Ajat Sudrajat(2021)," Earthquake Early Warning System Using Ncheck and Hard-Shared Orthogonal Multitarget Regression on Deep Learning"doi: https://doi.org/10.1109/LGRS.2021.3066346.

[2]S. Widiyantoro et al., (2020)"Implications for megathrust earthquakes and tsunamis from seismic gaps south of Java Indonesia," Sci. Rep., vol. 10,no. 1, pp. 1-11Dec.2020,doi:10.1038/s41598-020-72142-z.

[3]X. Zhang, J. Zhang, C. Yuan, S. Liu, Z. Chen, and W. Li,(2020) "Locating induced earthquakes with a network of seismic stations in Oklahoma via a deep learning method," Sci. Rep., vol. 10, no. 1, pp. 1—12, Dec. 2020,doi: 10.1038/s41598-020-58908-5.

[4]D. Jozinovi ́c, A. Lomax, I. Štajduhar, and A. Michelini, "Rapid predic-tion of earthquake ground shaking intensity using raw waveform data and a convolutional neural network," 2020, arXiv:2002.06893. [Online].Available: https://arxiv.org/abs/2002.06893

[5]M. P. A. van den Ende and J. P. Ampuero, "Automated seismic source characterization using deep graph neural networks," Geophys. Res. Lett., vol. 47, no. 17, pp. 1—11, Sep. 2020, doi: 10.1029/2020GL088690

[6]O. M. Saad and Y. Chen,(2020) "Earthquake detection and P-wave arrival time picking using capsule neural network," IEEE Trans. Geosci.Remote Sens., early access, Sep. 28, 2020, doi: 10.1109/tgrs.2020.3019520.

[7] S. M. Mousavi, W. L. Ellsworth, W. Zhu, L. Y. Chuang, andG. C Beroza,(2020)"Earthquake transformer—An attentive deep-learningmodel for simultaneous earthquake detection and phase picking," NatureCommun., vol. 11, no. 1, pp. 1—12, Dec. 2020, doi: 10.1038/s41467-020-17591-w.

[8] Z. E. Ross, Y. Yue, M. Meier, E. Hauksson, and T. H. Heaton,(2019)"PhaseLink: A deep learning approach to seismic phase association,"J. Geophys. Res., Solid Earth, vol. 124, no. 1, pp.856—869,Jan.2019,doi:10.1029/2018JB016674.

[9]E. Pardo, C. Garfias, and N. Malpica,(2019) "Seismic phase picking using convolutional networks," IEEE Trans. Geosci. Remote Sens., vol. 57,no. 9, pp. 7086—7092, Sep. 2019, doi: 10.1109/TGRS.2019.2911402

[10]A. Lomax, A. Michelini, and D. Jozinovi ́c,(2019) “An investigation of rapid earthquake characterization using single-station waveforms and a con-volutional neural network,” Seismolog. Res. Lett., vol. 90, no. 2A,pp. 517—529, Mar. 2019, doi: 10.1785/0220180311

[11]M. Kriegerowski, G. M. Petersen, H. Vasyura-Bathke, and M. Ohrnberger, (2019)“A deep convolutional neural network for localization of clustered earthquakes based on multistation full waveforms,” Seismolog. Res. Lett., vol. 90, no. 2A, pp. 510—516, Mar. 2019, doi: 10.1785/0220180320

[12] A. Howard et al.,(2019) “Searching for mobileNetV3,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 1314—1324, doi:10.1109/ICCV.2019.00140.

[13]M. Tan and Q. V. Le,(2019) “EfficientNet: Rethinking model scaling for convolutional neural networks,” in Proc. 36th Int. Conf. Mach. Learn (ICML), Jun. 2019, pp. 10691—10700.

[14]O. Reyes and S. Ventura, (2019)“Performing multi-target regression via aparameter sharing-based deep network,” Int. J. Neural Syst., vol. 29,no. 9, Nov. 2019, Art.no.1950014, doi: 10.1142/S012906571950014X

[15] O. M. Saad, K. Inoue, A. Shalaby, L. Samy, and M. S. Sayed, (2018)“Automatic arrival time detection for earthquakes based on stacked denoising autoen-coder,” IEEE Geosci. Remote Sens. Lett., vol. 15, no. 11, pp. 1687—1691,Nov. 2018, doi: 10.1109/LGRS.2018.2861218

[16]A. D. Nugraha et al., “Hypocenter relocation along the Sunda arc in Indonesia, using a 3D seismic-velocity model,” Seismolog. Res. Lett.,vol. 89, no. 2A, pp. 603—612, Mar. 2018, doi: 10.1785/0220170107.