## Image result for krct logo INSURANCE CLAIM FRAUD PREDICTION WEB APP

## USING MACHINE LEARNING-FLASK WEB FRAMEWORK

**A DESIGN PROJECT REPORT**

***Submitted by***

# MOHAMED THANISH.M(Reg No.811720243029)

# SAM ARAVIND.R(Reg No. 811720243036)

# BALAJI.S (Reg No. 811720243302)

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

## BACHELOR OF TECHNOLOGY

***in***

## ARTIFICIAL INTELLIGENGE AND DATA SCIENCE

**K. RAMAKRISHNAN COLLEGE OF TECHNOLOGY**

(An Autonomous Institution, affiliated to Anna University Chennai and Approved by AICTE, New Delhi)

## SAMAYAPURAM – 621 112

**DEC - 2022**

## K. RAMAKRISHNAN COLLEGE OFTECHNOLOGY

## (AUTONOMOUS)

**SAMAYAPURAM – 621 112**

## BONAFIDE CERTIFICATE

## Certified that this design project report titled “ INSURANCE CLAIM FRAUD PREDICTION WEB APP USING MACHINE LEARNING AND FLASK WEB FRAMEWORK” is the bonafide work of MOHAMED THANISH.M (811720243029), SAM ARAVIND.R (811720243036), BALAJI.S (811720243302) who carried out the project under my supervision. Certified further, that to the best of my knowledge the work reported here in does not form part of any other project report or dissertation based on which a degree or award was conferred on an earlier occasion on this or any other candidate.

|  |  |
| --- | --- |
| **SIGNATURE**  **Dr**.**T.AVUDAIAPPAN.,**M.E.,Ph.D.,  **HEAD OF THE DEPARTMENT**  Department of AI  K.Ramakrishnan College of Technology (Autonomous)  Samayapuram – 621 112 | **SIGNATURE**  **MS.S.MURUGAVALLI**.,B.TECH.,M.E.,  **ASST HEAD OF THE DEPARTMENT**  Department of AI  K.Ramakrishnan College of Technology (Autonomous)  Samayapuram – 621 112 |

Submitted for the viva-voce examination held on ………………

**INTERNAL EXAMINER**

## DECLARATION

We jointly declare that the project report on **“INSURANCE CLAIM FRAUD PREDICTION WEB APP USING MACHINE LEARNING AND FLASK WEB FRAMEWORK”** is the result of original work done by us and best of our knowledge, similar work has not been submitted to **“ANNA UNIVERSITY CHENNAI”** for the requirement of Degree of **BACHELOR OF TECHNOLOGY**. This design project report is submitted on the partial fulfilment of the requirement of the award of Degree of **BACHELOR OF TECHNOLOGY**.

|  |
| --- |
| **SIGNATURE** |
| **M.MOHAMED THANISH** |
| **S.BALAJI**  **R.SAM ARAVIND** |

**Place:**Samayapuram **Date:**

## ACKNOWLEDGEMENT

## 

It is with great pride that we express our gratitude and in - debt to our institution “**K.Ramakrishnan College of Technology (KRCT) (Autonomous)**”, for providing us with the opportunity to do this project.

We are glad to credit honourable chairman **Dr. K. RAMAKRISHNAN**, **B.E.,** for having provided for the facilities during the course of our study in college.

We would like to express our sincere thanks to our beloved Executive Director **Dr. S. KUPPUSAMY, MBA, Ph.D.,** for forwarding to our project and offering adequate duration in completing our project.

We would like to thank our principal **Dr. N.VASUDEVAN, M.E., Ph.D.,** , who gave opportunity to frame the project the full satisfaction.

We whole heartily thanks to **DR. T. AVUDAIAPPAN**, **M.E., Ph.D.,** Head of the department, **ARTIFICAL INTELLIGENCE** for providing his encourage pursuing this project.

I express my deep and sincere gratitude to my project guide **MS.S.MURUGAVALLI, B.TECH., M.E.,** ASST HOD of **ARTIFICIAL INTELLIGENCE**, for her incalculable suggestions, creativity. assistance, and patience which motivated me to carry

I render my sincere thanks to Course Coordinator and other staff members for providing valuable information during the course. I wish to express my special thanks to the officials and Lab Technicians of our departments who rendered their help during the period of the work progress.

## ABSTRACT

In recent years fraudulent insurance claims is the problem faced by many of the insurance companies which leads to huge financial loss yearly and insurance fraud has been since the beginning of the insurance organization. When a person makes a false claim in order to get benefits to which they are not entitled is known as an insurance fraud. So detection of an insurance fraud is a challenging problem for the insurance industry.so we chose to do a project based on the insurance claims fraud prediction using machine learning. These frauds have adverse consequences on society as the losses are settled down by increasing the premium cost of policy holders. Also the traditional claim investigation process being time consuming and tedious that generally leads to inaccurate results has been identified as main culprit. These project are going to identify potential fraud insurance claims and help insurance company to make more secured claim authentication & settlement by early identification of probable fraud claims. Solution of these problem is to Build a classification methodology AI based App to determine whether a customer is placing a fraudulent insurance claim using machine learning(ML) algorithms such as random forest classifier, XGBoost classifier, logistic regression, ensemble methods (namely bagging and boosting) with Support Vector Machines,  K-nearest neighbors and using HTML,FLASK for front end development. The aim is to identify fraud claims accurately within shorter period of time. Throughout the process data analysis is used to validate, clean and extract the relevant data. Hence, by using this framework insurance company can maintain its respectability in outside world and can also share trustworthy relationship with customers.

## TABLE OF CONTENTS

|  |  |  |
| --- | --- | --- |
| **CHAPTER** | **TITLE** | **PAGE NO.** |
|  | **ABSTRACT** | 5 |
| **1** | **INTRODUCTION** | 9 |
|  | 1.1 Problem Statement | 9 |
|  | 1.2 Motivation | 10 |
|  | 1.3 Purpose | 10 |
|  | 1.4 Data Analysis | 10 |
|  | 1.5 Scope | 11 |
|  | 1.6 Machine Learning | 12 |
| **2** | **LITERATURE SURVEY** | 13 |
|  | 2.1 Detecting Insurance Claims Fraud Using Machine Learning Techniques | 13 |
|  | 2.2 Robust Fuzzy Rule Based Technique To Detect Frauds In Vehicle Insurance | 14 |
|  | 2.3 Insurance Fraud Detection Using Machine Learning | 15 |
|  | 2.4 Enhancing Claims Handling Process With Insurance Based Language Models | 16 |
|  | 2.5 Detectin Fraudlent Motor Insurance Claims Using SVM(Support Vector Machines) With Adaptive Synthetic Sampling Method | 17 |
|  | 2.6 Automobile Insurance Fraud Detection Using Supervised Classifiers | 18 |
|  | 2.7 Analysis Accuracy Of XGboost Model For Multiclass Classification -A Case Study Of Applicant Level Risk Prediction For Life Insurance | 19 |
|  | 2.8 A Data Mining Based Target Regression- Oriented Approach To Modelling Of Health Insurance Claims | 20 |

|  |  |  |
| --- | --- | --- |
| **CHAPTER** | **TITLE** | **PAGE NO.** |
| **3** | **SYSTEM SPECIFICATION** | 21 |
|  | 3.1 Hardware Configuration | 21 |
|  | 3.2 Software Configuration | 21 |
|  | 3.3 Software Description | 21 |
|  | 3.4 Libraries | 22 |
|  | 3.5 Development Environment | 22 |
| **4** | **IMPLEMENTATION** | 23 |
|  | 4.1 Reference Implementation | 23 |
|  | 4.2 Other Implementation | 23 |
|  | 4.2.1 Unsupported Implementation | 23 |
|  | 4.2.2 Cross- Complier To Other Languages | 24 |
| **5** | **SYSTEM ANALYSIS** | 25 |
|  | 5.1 Existing System | 25 |
|  | 5.2 Proposed System | 27 |

|  |  |  |
| --- | --- | --- |
| **CHAPTER** | **TITLE** | **PAGE NO.** |
| **6** | **ARCHITECTURAL DESIGN** | 35 |
|  | 6.1 System Design | 35 |
|  | 6.2 UML Diagram | 36 |
|  | 6.2.1 Use Case Diagram | 37 |
|  | 6.2.2 Activity Diagram | 37 |
|  | 6.2.3 Class Diagram | 38 |
|  | 6.2.4 Sequence Diagram | 38 |
|  | 6.2.5 Data Flow Diagram | 39 |
| **7** | **MODULE DESRIPTION** | 41 |
|  | 7.1 Module | 41 |
|  | 7.1.1 Data Collection | 41 |
|  | 7.1.2 Data Processing | 42 |
|  | 7.1.3 Creation of Tree Using C4.5 Algorithm | 43 |
|  | 7.2 Performance Evaluation | 43 |
|  | 7.2..1 Model Selection | 43 |
| **8** | **PROJECT MODEL & DEPLOYMENT** | 45 |
| **9** | **CONCLUSION AND FUTURE SCOPE** | 49 |
|  | 9.1 Conclusion | 49 |
|  | 9.2 Future Scope | 50 |
|  | **SAMPLE CODE** | 51 |
|  | **REFERENCES** | 53 |

# CHAPTER 1

## INTRODUCTION

## 1.1 INSURANCE CLAIM FRAUD PREDICTION

The insurance market is a highly profitable market that moves large sums of money over the years. In Brazil alone, about 10.8 billion USD was paid in insurance policies in 2017 (Brazilian National Confederation of Insurance Companies, 2017). Similarly, frauds can bring huge losses to the companies: In the same year of 2017, the total value of all occurred claims was around 10.0 billion USD, while the value of proven frauds totaled 221.2 million USD (Brazilian National Confederation of Insurance Companies, 2017).

Bearing in mind the economic relevance of this market and the challenge of fraud detection by professional analysts, the search for data mining and machine learning techniques had been showing its predicting potential in financial applications, as seen in works like Hsu et al. (2016), notably when involving complex problems and non-linear patterns (Huang et al., 2004, Soman et al., 2009).

Fraud in insurance is an unethical activity performed systematically to get some financial gain. These fraudulent claims present overpriced and large problem for insurance company leading to billions of dollars of needless expenses per year. Also due to some flaws in traditional process most of the companies are in search of some new technique to find fraud claims. So here we propose ML based automated framework employed with XGBoost algo to classify claims. We also compare the performance of XGBoost algo with other algorithms to obtain most accurate results.

## 1.2 PROBLEM STATEMENT

The goal of this project is to build a model that can detect auto insurance fraud. The challenge behind fraud detection in machine learning is that frauds are far less common as compared to legit insurance claims.

Insurance fraud detection is a challenging problem, given the variety of fraud patterns and relatively small ratio of known frauds in typical samples. While building detection models, the savings from loss prevention needs to be balanced with the cost of false alerts. Machine learning techniques allow for improving predictive accuracy, enabling loss control units to achieve higher coverage with low false positive rates.

Insurance frauds cover the range of improper activities which an individual may commit in order to achieve a favourable outcome from the insurance company. This could range from staging the incident, misrepresenting the situation including the relevant actors and the cause of incident and finally the extent of damage caused.

**1.3 MOTIVATION AND PURPOSE**

**MOTIVATION:** Early detection of fraud by employees or customers can significantly reduce industry costs and positively impact insurance premiums. By learning collected customer data, you can analyze and predict whether newly filed claims are inappropriate manipulation behavior.

**PURPOSE :** The purpose of this project is to analyze and detect the frauds in insurance claims with the help of data mining techniques like predictive models, rule learning, time series learning etc..,thereby helping insurance companies to savings from loss prevention needs to be balanced with the cost of false alerts.

**1.4 ANALYSIS & SCOPE**

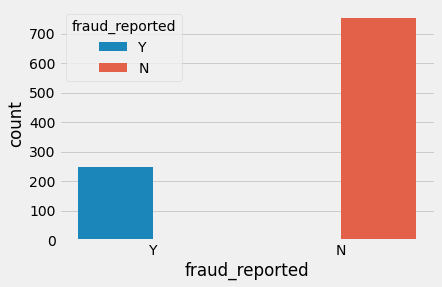
**DATA ANALYSIS:**

In this project, we have a dataset which has the details of the insurance policy along with the customer details. It also has the details of the accident on the basis of which the claims have been made.

The given dataset contains 1000 rows and 40 columns. The column names like policy number, policy bind date, policy annual premium, incident severity, incident location, auto model, etc.all currently available information and any price changes that are not based on newly revealed information thus are inherently unpredictable.

**EXPLORATORY DATA ANALYSIS:**

**Dependent variable:** Exploratory data analysis was conducted starting with the dependent variable, Fraud\_reported. There were 247 frauds and 753 non-frauds. 24.7% of the data were frauds while 75.3% were non-fraudulent claims.



**FIG 1 : Dependent variable**

**Correlations data analysis:** Heatmap was plotted for variables with at least 0.3 Pearson’s correlation coefficient, including the DV. Month as customer and age had a correlation of 0.92. Probably because drivers buy auto insurance when they own a car and this time measure only increase with age. Apart from that, there don’t seem to be many correlations in the data. There don’t seem to be multicollinearity problems except maybe that all the claims are all correlated, and somehow total claims have accounted for them. However, the other claims provide some granularity that will not otherwise be captured by total claims. Thus, these variables were kept.

**Visualizing varibales:** The value of fraud reported differs across hobbies of the customer seems like chess players and crossfitters have higher tendencies to fraud.

**SCOPE:**

The insurance fraud detection market is expected to register a CAGR of over 17.4% during the forecast period (2021-2026). Fraudulent claims in the insurance industry have steadily grown to be the single largest expense to property and casualty insurers, taking up to 10% of an insurer’s revenue. Accurate prediction gives a chance to reduce financial loss for the company. A major cause of  the increased costs are the payment errors made by the insurance companies while processing claims.

## MACHINE LEARNING

XGBoost Is A Fast And Powerful Machine Learning Algorithm, Which Has Become A Popular Choice For Online Machine Learning Competitions Due To Its High Efficacy. Its Complete Moniker Is Extreme Gradient Boosting. XGBoost Is An Enhanced Implementation Of Gradient Boosting In Terms Of Both Speed And Performance. XGBoost Is A High-Performance Decision-Tree-Based Ensemble Learning Method Usually Used For Supervised Learning. In Ensemble Learning, Results From Different Individual Models Are Combine In Order To Make An Optimize Prediction. Instead Of Depending On A Single Model’s Outcome, The Ensemble Learning Method Makes A Decision By Aggregating Results From Different Models To Put Forth A Model With Better Accuracy.

The random forest (RF) classifier is a supervised learning (SL) algorithm which you can use for regression and classification problems. It is among the most popular machine learning algorithms due to its high flexibility and ease of implementation.  it uses randomness to enhance its accuracy and combat overfitting, which can be a huge issue for such a sophisticated algorithm. These algorithms make decision trees based on a random selection of data samples and get predictions from every tree. After that, they select the best viable solution through votes.

## CHAPTER 2 LITERATURE SURVEY

**2.1 TITLE :** DETECTING INSURANCE CLAIMS FRAUD USING MACHINE LEARNING TECHNIQUES

**AUTHOR:** Riya Roy, Thomas George K

**YEAR & PUBLICATION:** 19 October 2017,IEEE

**ALGORITHM USED:** Decision tree(DT), Random forest(RF), Naïve Bayes(NB).

**ABSTRACT:** The insurance industries consist of more than thousand companies in worldwide. And collect more than one trillions of dollars premiums in each year. When a person or entity make false insurance claims in order to obtain compensation or benefits to which they are not entitled is known as an insurance fraud. The total cost of an insurance fraud is estimated to be more than forty billions of dollars. So detection of an insurance fraud is a challenging problem for the insurance industry. The traditional approach for fraud detection is based on developing heuristics around fraud indicator. The auto\vehicle insurance fraud is the most prominent type of insurance fraud, which can be done by fake accident claim. In this paper, focusing on detecting the auto\vehicle fraud by using, machine learning technique. Also, the performance will be compared by calculation of confusion matrix. This can help to calculate accuracy, precision, and recall.

**DISADVANTAGES:**

compare with the algorithms, decision tree and random forest algorithms have better performance than naïve bayes.

## 2.2 TITLE: ROBUST FUZZY RULE BASED TECHNIQUE TO DETECT

## FRAUDS IN VEHICLE INSURANCE

**AUTHOR:** K. Supraja, S.J. Saritha

**YEAR & PUBLICATION:** 21 June 2018, IEEE

**ALGORITHM USED:** Bayesian algorithm, Fuzzy Logic, S-curve

**ABSTRACT:**  insurance companies have its significance in the society. Customers are interested in claiming the insurance on their property, vehicles. All the vehicle users approach different insurance companies which provide better security to their vehicles. In the same manner the Fraudulent cases also increases. There are different mining techniques in detecting fraud and analyse the data. In this literature survey we present some techniques to fraud analysis, classification and prediction which we consider important to handle fraud detection. Among those The Naïve Bayesian model is more powerful fraud detection in automobile insurance. Bayesian visualization is selected to analyse and interpret the classifier predictions. However this visualization technique is not suitable for abundant data with little frauds. To avoid this limitation, we are using Fuzzy Logic by framing fuzzy rules to improve the Fraud Detection. By using this technique time complexity will be decreased and implementation is easy and interprets the results accurately.

**DISADVANTAGES:**

Bayesian algorithm gives the less accuracy and complexity is very high.

**2.3 TITLE**: INSURANCE FRAUD DETECTION USING MACHINE LEARNING

**AUTHOR:** Soham Shah, Shrutee Phadke, Princia Koli, Shweta Sharma

**YEAR & PUBLICATION:** Apr 2021, IRJET

**ALGORITHM USED:** XGBoost Algorithm,decision tree

**ABSTRACT**: Nowadays fraudulent insurance claims is the problem faced by many of the insurance companies which leads to huge financial loss yearly. These frauds have adverse consequences on society as the losses are settled down by increasing the premium cost of policy holders. Also the traditional claim investigation process being time consuming and tedious that generally leads to inaccurate results has been identified as main culprit. Thus, in this paper we develop an automated fraud detection application framework based on machine learning and XGBoost algorithm. The aim is to identify fraud claims accurately within shorter period of time. Throughout the process data analysis is used to validate, clean and extract the relevant data. Hence, by using this framework insurance company can maintain its respectability in outside world and can also share trustworthy relationship with customers.

**DISADVANTAGES:**

Decision tree algo only gives 70% of accuracy and it takes time.

**2.4 TITLE:** ENHANCING CLAIMS HANDLING PROCESSES WITH INSURANCE BASED LANGUAGE MODELS

**AUTHOR:** suraj,delip

**YEAR & PUBLICATION:** feb 19,IEEE

**ALGORITHM USED:** IBLMs , BERT

**ABSTRACT**: Insurance companies manage a large number of claims on a daily basis as new claims are reported and existing claims are serviced. A key component for servicing a claim is the ability for Claims personnel to enter in raw text, aka claims notes. Claims notes contain invaluable information often beyond that of structured data, capturing this information in a machine learning setting offers remarkable benefits to many downstream tasks in a Claims department. The ability to leverage claims notes enables an insurance company not only to make data-driven and insightful decisions while handling claims, but to create value through working more efficiently and serve their customers more effectively. To best leverage the information contained claims notes, we develop insurance-based language models (IBLMs) by further pre-training existing general domain language models (ULMFiT and BERT) on a large number of claim notes with enhanced vocabulary.

**DISADVANTAGES:**

* It needs high storage space and costly.
* Prone to underfitting, Sensitive to outliers

# 2.5 TITLE: DETECTING FRAUDULENT MOTOR INSURANCE CLAIMS USING SUPPORT VECTOR MACHINES WITH ADAPTIVE SYNTHETIC SAMPLING METHOD

**AUTHOR:** ahmed ali

**YEAR & PUBLICATION:** 19 November 2020 ,IEEE

**ALGORITHM USED:** SVM and ADASYN

**ABSTRACT:** Classification algorithms suffer from imbalanced training sets. In the area of detecting fraudulent claims in the insurance industry, fraud cases are rare as compared to the genuine ones. Therefore, algorithms of detecting fraud have fewer training samples of positive cases, leading to lower performance metrics compared to when there are equal cases. In this paper, we propose a machine learning method of detecting fraudulent claims. The proposed method uses the adaptive synthetic sampling method (ADASYN) to remove imbalances in the dataset. We then used Support Vector Machines (SVM) to classify the claim cases. The outcome of the algorithm is compared to the imbalanced datasets and other existing methods.

**DISADVANTAGES:** Different machine learning algorithms and feature selection methods performed differently for various stocks, which would not be the case if the fraud had followed a random pattern.

# 2.6 TITLE: AUTOMOBILE INSURANCE FRAUD DETECTION USING SL

# CLASSIFIER

**AUTHOR:** arinin deni ,enrico loah

**YEAR & PUBLICATION:** 21 November 2020,IEEE

**ALGORITHM USED:** MLP, random forest , SMOTE

**ABSTRACT**: Automobile fraudulent claim leads to several consequences for the company and policyholder. The current detection system is costly and inefficient. This research aims to design a prediction model in detecting automobile insurance fraud using a machine learning approach. The study used realworld data on an automobile insurance company in Indonesia. The dataset has a high imbalanced distribution between the data of policyholders who commit fraud and legitimate data. This research handles the imbalanced dataset problem by using the Synthetic Minority Oversampling Technique (SMOTE) and undersampling methods. The proposed supervised classifiers are Multilayer Perceptron (MLP), Decision Tree C4.5, and Random Forest(RF). The performance of models is evaluated through the confusion matrix, ROC Curve, and parameters such as sensitivity. This research found that Random Forest outperformed the results comparing to other classifiers with 98.5% accuracy.

**DISADVANTAGE:**

It higher time to train the model.

# 2.7 TITLE: ANALYSIS ACCURACY OF XGBOOST MODEL FOR MULTICLASS CLASSIFICATION - A CASE STUDY OF APPLICANT LEVEL RISK PREDICTION FOR LIFE INSURANCE

**AUTHOR:** Yekti Widyaningsih,hendri

**YEAR & PUBLICATION:** 10 February 2020,IEEE

**ALGORITHM USED:** XGBoost,bayesian

**ABSTRACT**: Risk level assessment for insurance applicants is an important part of life insurance, so it needs to be classified. Determination of the level of risk claims on life insurance is based on the applicant's historical data. Submission to become a member of a life insurance requires a short time. But the application of a machine learning model can help classify prospective insurance applicants based on the level of risk quickly. One machine learning model is Extreme Gradient tree boosting (XGBoost) which is a decision tree based model. This model is used to predict risk in life insurance. The missing values in the data used are overcome by several strategies in the data processing process to increase the accuracy value of the XGBoost model. The results of this study show that the accuracy of the XGBoost model is 0.60730 with kappa units which indicates that the XGBoost model is very good and can be applied to the problem of predicting the level of risk claims for life insurance applicants.

**DISADVANTAGES:**

Hardware dependence, Unexplained behavior of the network design , prediction accuracy of 78%

**2.**8 **TITLE:** A DATA MINING BASED TARGET REGRESSION – ORIENTED APPROACH TO MODELING OF HEALTH INSURANCE CLAIMS

**AUTHOR:** Harshvardhan GM

**YEAR & PUBLICATION:**06 May 2021 , IEEE

**ALGORITHM USED:** Random Forest Regression

**ABSTRACT:** Machine learning or Data mining algorithms can be used for predicting future management and thus treated as powerful tools. In recent days, data mining has become very important for gaining vital information in healthcare industries. The Health care insurance cost plays a vital role in developing medical facilities. To provide better medical facilities, it is very essential to forecast the cost of medical insurance which is one of the possibilities to enhance medical facilities. The paper deals with predicting the cost of the health insurance which has to be paid by the patient. Here various data mining regression algorithms such as decision tree, random forest, polynomial regression and linear regression are implemented to achieve the best prediction analysis. A comparison has been done between the actual and predicted expenses of the prediction premium and eventually, a graph has been plotted on this basis which will enlighten us to choose the best-suited regression algorithm for the insurance policy prediction. After the execution of these regression algorithms for prediction, correctness has been measured by the Coefficient of determination (r2\_score), Root Mean Squared Error (RMSE) and Mean Squared Error (MSE) of each algorithm to check for the best-suited algorithm.

**DISADVANTAGES:**

Parameters have been chosen for high time data consumption .

# CHAPTER 3

## SYSTEM SPECIFICATION

* 1. **H/W SYSTEM CONFIGURATION: -**
     + Processor – Intel Corei5,GPU
     + RAM - 4 GB (Min) Or Higher
     + Hard Disk - 20 GB
     + Keyboard – Standard Keyboard
     + Monitor – 20 Inch Color Monitor

## S/W SYSTEM CONFIGURATION: -

* + - Operating System : Windows 8 and above
    - Front End : PYTHON
    - Back End : MySQL
    - APP : web app
    - PLATFORM : VS CODE

## SOFTWARE DESCRIPTION:

Python is an interpreter, high-level, general-purpose programming language. Python is dynamically typed, and garbage collected. It supports multiple programming paradigms, including procedural, object-oriented, and functional programming. Python is often described as a "batteries included" language due to its comprehensive standard library.

Python was conceived in the late 1980s as a successor to the ABC language. Python 2.0, released in 2000, introduced features like list of the

comprehensions and a garbage collection system capable of collecting reference cycles. Python 3.0, released in 2008, was a major revision of the language that is not completely backward compatible, and much Python 2 code does not run unmodified on Python 3.

Python interpreters are available for many operating systems. A global community of programmers develops and maintains Python, an open-source reference implementation. A non-profit organization, the Python Software Foundation, manages and directs resources for Python and Python development.

## .LIBRARIES

Numpy

Pandas

Matplotlib

Sklearn

Flask etc….

## 3.5.DEVELOPMENT ENVIRONMENTS

Most Python implementations (including Python) include a read–eval– print loop (REPL), permitting them to function as a command line interpreter for which the user enters statements sequentially and receives results immediately.

Other shells, including IDLE and I Python, add further abilities such as auto- completion, session state retention and syntax highlighting.

As well as standard desktop integrated development environments, there are Web browser-based IDEs.

**CHAPTER-4**

**IMPLEMENTATIONS**

**4.1 REFERENCE IMPLEMENTATION**

CPython is the reference implementation of Python. It is written in C, meeting the C89 standard with several select C99 features. It compiles Python programs into an intermediate bytecode which is then executed by its virtual machine. CPython is distributed with a large standard library written in a mixture of C and native Python. It is available for many platforms, including Windows and most modern Unix-like systems. Platform portability was one of its earliest priorities.

**4.2 OTHER IMPLEMENTATIONS**

PyPy is a fast, compliant interpreter of Python 2.7 and 3.5. Its just-intime compiler brings a significant speed improvement over CPython but several libraries written in C cannot be used with it. Stackless Python is a significant fork of CPython that implements microthreads; it does not use the C memory stack, thus allowing massively concurrent programs. PyPy also has a stackless version. MicroPython and CircuitPython are Python 3 variants optimized for microcontrollers. This includes Lego Mindstorms EV3. RustPython is a Python 3 interpreter written in Rust.

**4.2.1 UNSUPPORTED IMPLEMENTATIONS**

Other just-in-time Python compilers have been developed, but are now unsupported: Google began a project named Unladen Swallow in 2009, with the aim of speeding up the Python interpreter five-fold by using the LLVM, and of improving its multithreading ability to scale to thousands of cores, while ordinary implementations suffer from the global interpreter lock. Psyco is a just-in-time specialising compiler that integrates with CPython and transforms bytecode to machine code at runtime. The emitted code is specialized for certain data types and is faster than standard Python code. In 2005, Nokia released a Python interpreter for the Series 60 mobile phones named PyS60. It includes many of the modules from the CPython 43 implementations and some additional modules to integrate with the Symbian operating system. The project has been kept up-to-date to run on all variants of the S60 platform, and several third-party modules are available.

**4.2.2 CROSS-COMPILERS TO OTHER LANGUAGES**

There are several compilers to high-level object languages, with either unrestricted Python, a restricted subset of Python, or a language similar to Python as the source language:

* Python enables the use of the Java class library froma Python program.
* IronPython follows a similar approach in order to run Pythonprograms on the.NET Common Language Runtime.
* The RPython language can be compiled to C, and is usedto build the PyPyinterpreter of Python.
* Pyjs compiles Python to JavaScript.
* Cython compiles Python to C and C++
* Numba uses LLVM to compile Python to machine code.
* Python compiles Python to C++.
* Somewhat dated Pyrex (latest release in 2010)

**CHAPTER 5**

**SYSTEM ANALYSIS**

**5.1.EXISTING SYSTEM**

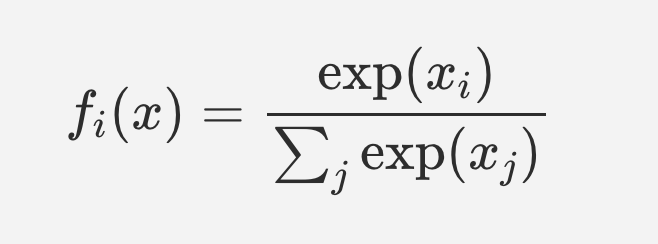
An insurance claim fraud predicition is a challenging task since the factors Involved in pricing dynamically change overtime and make the price fluctuate. In the last decade, researcher have incorporated machine learning algorithms and Data mining strategies to better model observed prices but accurate results are not predicted since there is no day to day price comparison.

**ALGORITHMS USED:**

**1) LOGISTIC ALGORITHM**: Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.Mathematically, a logistic regression model predicts P(Y=1) as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection etc.

**2)NAÏVE BAYES ALGORITHM:** Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is mainly used in text classification that includes a high-dimensional training dataset. Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.

3) **SOFTMAX REGRESSION ALGORITHM:**  The [Softmax function](https://en.wikipedia.org/wiki/Softmax_function) converts raw values (as an outcome of functions) into probabilities. In the function of softmax function was invented in 1959 by the social scientist R. Duncan Luce in context of choice models*.* Here is what the softmax function looks like:



**FIG 2:** **SOFTMAX REGRESSION DIAGRAM**

**4) SUPPORT VECTOR MACHINE:** Support Vector Machine or SVM is one of the most popular Supervised Learning (SL) algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

1. **LINEAR REGRESSION ALGORITHM:** Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables,Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

**5.2 .PROPOSED SYSTEM**

Our system predicts whether the insurance claim is real or fake which means fraudulent. It considers the policy deductible , policy annual premium and the collision type to predict the premiums. XG Boost algorithm and Random forest algorithm have been used for the prediction, the output of which is obtained by performing voting among the class label obtained as result by each of these trees.

**RANDOM FOREST ALGORITHM:**

A Random Forest Algorithm is a supervised machine learning algorithm which is extremely popular and is used for Classification and Regression problems in Machine Learning. We know that a forest comprises numerous trees, and the more trees more it will be robust. Similarly, the greater the number of trees in a Random Forest Algorithm, the higher its accuracy and problem-solving ability.  Random Forest is a classifier that contains several decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.

1. It is incompetent in terms of accuracy among all other current algorithms.

2. Its runs proficiently on large data bases.

3. It can leverage thousands of input variables.

**XGBOOST ALGORITHM:**

The XGBoost (eXtreme Gradient Boosting) is a popular and efficient open-source implementation of the gradient boosted trees algorithm. Gradient boosting is a supervised learning algorithm that attempts to accurately predict a target variable by combining an ensemble of estimates from a set of simpler and weaker models. The XGBoost algorithm performs well in machine learning competitions because of its robust handling of a variety of data types, relationships, distributions, and the variety of hyperparameters that you can fine-tune. You can use XGBoost for regression, classification (binary and multiclass), and ranking problems.

## ****BOOSTING****

## Boosting is an ensemble learning technique to build a strong classifier from several weak classifiers in series. Boosting algorithms play a crucial role in dealing with bias-variance trade-off. Unlike bagging algorithms, which only controls for high variance in a model, boosting controls both the aspects (bias & variance) and is considered to be more effective.

Below are the few types of boosting algorithms:

1. AdaBoost (Adaptive Boosting)
2. Gradient Boosting
3. XGBoost
4. CatBoost
5. Light GBM

## ****XGBoost****

## XGBoost stands for eXtreme Gradient Boosting. It became popular in the recent days and is dominating applied machine learning and Kaggle competition for structured data because of its scalability.XGBoost is an extension to gradient boosted decision trees (GBM) and specially designed to improve speed and performance.

## 

## ****ADABOOST:****

## AdaBoost is short for Adaptive Boosting. AdaBoost was the first successful boosting algorithm developed for binary classification. Also, it is the best starting point for understanding boosting algorithms. It is adaptive in the sense that subsequent classifiers built are tweaked in favour of those instances misclassified by previous classifiers. It is sensitive to noisy data and outliers.

## AdaBoost uses multiple iterations to generate a single composite strong learner. It creates a strong learner by iteratively adding weak learners. During each phase of training, a new weak learner is added to the ensemble, and a weighting vector is adjusted to focus on examples that were misclassified in previous rounds. The result is a classifier that has higher accuracy than the weak learner classifiers.

## ****GRADIENT BOOSTING:****

## Gradient boosting is one of the most powerful techniques for building predictive models, and it is called a Generalization of AdaBoost. The main objective of Gradient Boost is to minimize the loss function by adding weak learners using a gradient descent optimization algorithm. The generalization allowed arbitrary differentiable loss functions to be used, expanding the technique beyond binary classification problems to support regression, multi-class classification and more.

**TYPES OF MACHINE LEARNING:**

To better understand Random Forest algorithm and how it works, it's helpful to review the three main types of machine learning –

## 1. Supervised Machine Learning:

As its name suggests, Supervised machine learning is based on supervision. It means in the supervised learning technique, we train the machines using the "labelled" dataset, and based on the training, the machine predicts the output. Here, the labelled data specifies that some of the inputs are already mapped to the output. More preciously, we can say; first, we train the machine with the input and corresponding output, and then we ask the machine to predict the output using the test dataset.

**The main goal of the supervised learning technique is to map the input variable(x) with the output variable(y).**Some real-world applications of supervised learning are**Risk Assessment, Fraud Detection, Spam filtering,**etc.

### Categories of Supervised Machine Learning

Supervised machine learning can be classified into two types of problems, which are given below:

* **Classification**
* **Regression**

### a) Classification

Classification algorithms are used to solve the classification problems in which the output variable is categorical, such as **"Yes" or No, Male or Female, Red or Blue, etc.** The classification algorithms predict the categories present in the dataset. Some real-world examples of classification algorithms are **Spam Detection, Email filtering, etc.**

Some popular classification algorithms are given below:

* **Random Forest Algorithm**
* **Decision Tree Algorithm**
* **Logistic Regression Algorithm**
* **Support Vector Machine Algorithm**

### b) Regression

Regression algorithms are used to solve regression problems in which there is a linear relationship between input and output variables. These are used to predict continuous output variables, such as market trends, weather prediction, etc.

Some popular Regression algorithms are given below:

* Simple Linear Regression Algorithm
* Multivariate Regression Algorithm
* Decision Tree Algorithm
* Lasso Regression

## 2. Unsupervised Machine Learning

Unsupervised learning is different from the Supervised learning technique; as its name suggests, there is no need for supervision. It means, in unsupervised machine learning, the machine is trained using the unlabeled dataset.

In unsupervised learning, the models are trained with the data that is neither classified nor labelled.

**The main aim of the unsupervised learning algorithm is to group or categories the unsorted dataset according to the similarities, patterns, and differences.** Machines are instructed to find the hidden patterns from the input dataset.

### Categories of Unsupervised Machine Learning

Unsupervised Learning can be further classified into two types, which are given below:

* **Clustering**
* **Association**

### 1) Clustering

The clustering technique is used when we want to find the inherent groups from the data. It is a way to group the objects into a cluster such that the objects with the most similarities remain in one group and have fewer or no similarities with the objects of other groups. An example of the clustering algorithm is grouping the customers by their purchasing behaviour.

Some of the popular clustering algorithms are given below:

* **K-Means Clustering algorithm**
* **Mean-shift algorithm**
* **DBSCAN Algorithm**
* **Principal Component Analysis**
* **Independent Component Analysis**

### 2) Association

Association rule learning is an unsupervised learning technique, which finds interesting relations among variables within a large dataset. The main aim of this learning algorithm is to find the dependency of one data item on another data item and map those variables accordingly so that it can generate maximum profit. Some popular algorithms of Association rule learning are **Apriori Algorithm, Eclat, FP-growth algorithm.**

## 3. Semi-Supervised Learning

**Semi-Supervised learning is a type of Machine Learning algorithm that lies between Supervised and Unsupervised machine learning.** It represents the intermediate ground between Supervised (With Labelled training data) and Unsupervised learning (with no labelled training data) algorithms and uses the combination of labelled and unlabeled datasets during the training period.As labels are costly, but for corporate purposes, they may have few labels. It is completely different from supervised and unsupervised learning as they are based on the presence & absence of labels.

**To overcome the drawbacks of supervised learning and unsupervised learning algorithms, the concept of Semi-supervised learning is introduced.**

## 4. Reinforcement Learning

**Reinforcement learning works on a feedback-based process, in which an AI agent (A software component) automatically explore its surrounding by hitting & trail, taking action, learning from experiences, and improving its performance.** Agent gets rewarded for each good action and get punished for each bad action.The reinforcement learning process is similar to a human being; for example, a child learns various things by experiences in his day-to-day life. An example of reinforcement learning is to play a game, where the Game is the environment, moves of an agent at each step define states

A reinforcement learning problem can be formalized using**Markov Decision Process(MDP).**

### Categories of Reinforcement Learning:

Reinforcement learning is categorized mainly into two types of methods/algorithms.

* **Positive Reinforcement Learning**
* **Negative Reinforcement Learning**

**GOALS:**

Cost management will be attained and the insurance company will be reduced with the burden of fraud insurance claims. AI technology not only automates the fraud detection process but also identifies fraud patterns allowing early flagging and prompt response to any potential incidents.

## CHAPTER 6

## ARCHITECTURAL DESIGN

## 6.1.SYSTEM DESIGN

A system architecture is the conceptual model that defines the structure, behaviour, and more views of a system. An architecture description is a formal description and representation of a system, organized in a way that supports reasoning about the structures and behaviours of the system.

# devLifeCycle_2

**FIG 3 : SYSTEM ARCHITECTURE**

## UML DIAGRAM

UML stands for Unified Modelling Language. UML is a standardized general-purpose modelling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group.

The goal is for UML to become a common language for creating models of object-oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with uml.

The UML is a very important part of developing objects-oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

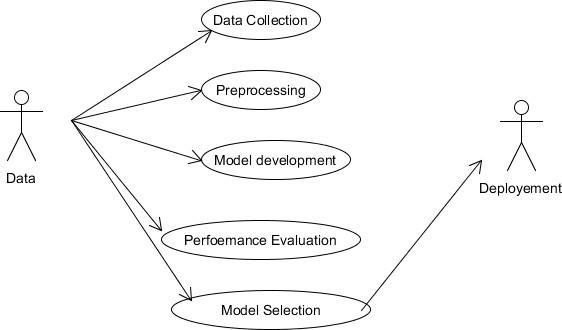
## GOALS:

The Primary goals in the design of the UML are as follows:

* + 1. Provide users a ready-to-use, expressive visual modeling Language so that they can develop and exchange meaningful models.
    2. Provide extendibility and specialization mechanisms to extend the core concepts.
    3. Be independent of particular programming languages & development - process.
    4. Provide a formal basis for understanding the modeling language.
    5. Encourage the growth of OO tools market.

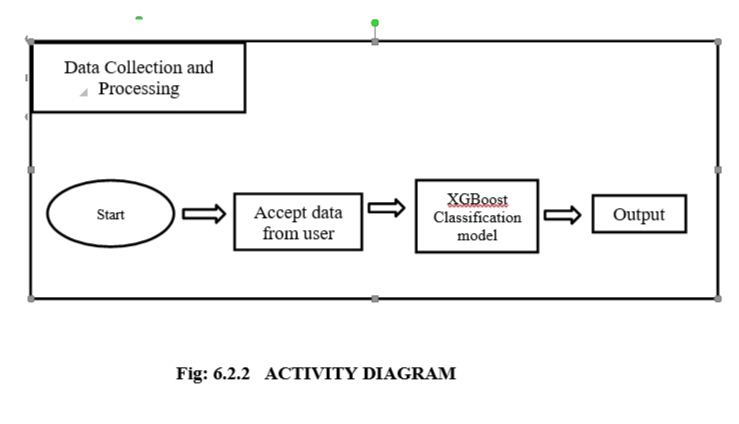
* + 1. **USE CASE DIAGRAM:**

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis.

Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases.

**FIG 4 :USE CASE DIAGRAM**

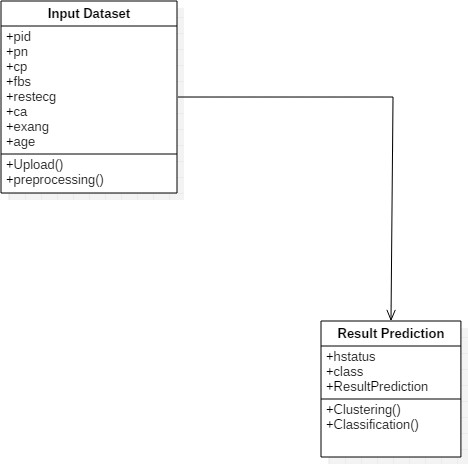
## 6.2.2.ACTIVITY DIAGRAM



**FIG 5: ACTIVITY DIAGRAM**

**6.2.3.CLASS DIAGRAM:**

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.



**FIG.6: CLASS DIAGRAM**

**6.2.4 SEQUENCE DIAGRAM:**

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

:Server

:Client

Enter the detail

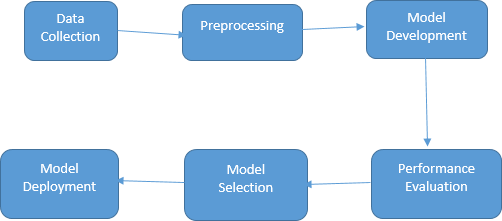
**XGBOOST**

RESULT

**Fig: 7 SEQUENCE DIAGRAM**

**6.2.5.DATA FLOW DIAGRAM:**

* + - 1. The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system.
      2. The data flow diagram (DFD) is one of the most important modeling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system.
      3. DFD shows how the information moves through the system and how it is modified by a series of transformations. It is a graphical technique that depicts information flow and the transformations that are applied as data move from input to output.
      4. DFD is also known as bubble chart. A DFD may be used to represent a system at any level of abstraction.The DFD belongs to structured-analysis modeling tools. Data Flow diagrams are very popular because they help us to visualize the major steps and data involved in software-system processes.



**FIG 8: DATA FLOW DIAGRAM**

**CHAPTER 7**

**MODULE DESCRIPTION**

* 1. **MODULES**

**7.1.1 DATA COLLECTION**

* + - * In this process to collect a data from github and Kaggle. This Kaggle data collection has an only numerical value and, in this data, using multiple purpose.
      * Github & Kaggle supports a variety of dataset publication formats, but we strongly encourage dataset publishers to share their data in an accessible, non- proprietary format if possible. Not only are open, accessible data formats better supported on the platform, but they are also easier to work with for more people regardless of their tools.
      * The data that is used to train and test the module is obtained by extracting it from various sources. The extracted data is stored in a prescribed format on Microsoft excel in xls format.

Extracted data in Excel(xls)

Real time data

Automation Anywhere Software

**Fig: 9 DATA COLLECTION**

* + 1. **DATA PREPROCESSING**
       - Data pre-processing is a process of preparing the raw data and making it suitable for a deep learning model. It is the first and crucial step while creating a deep learning model.
       - When creating a machine learning project, it is not always a case that we come across the clean and formatted data.
       - And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So, for this, we use data pre-processing task.

**FIG 10: DATA PREPROCESSING**

Processed data in Excel

Manual pre- processing of data

Extracted data in excel

**7.1.3.CREATION OF TREES USING C4.5 ALGORITHM**

The trees that are built in our system are done so using C4.5 algorithm that is available on the internet. This algorithm uses gini index in order to find efficient split conditions for formation of daughter trees.

**7.2.PERFORMANCE EVALUVATION**

In this process is used to implement the project Accuracy.

**7.2.1.MODEL SELECTION**

* + - * Model selection is the process of selecting one final machine learning model from among a collection of candidate machine learning models for a training dataset.
      * Model selection is a process that can be applied both across different types of models and across models of the same type configured with different model hyper parameters.

**XGBoost ALGORITHM:**

The XGBoost (eXtreme Gradient Boosting) is a popular and efficient open-source implementation of the gradient boosted trees algorithm. Gradient boosting is a supervised learning algorithm that attempts to accurately predict a target variable by combining an ensemble of estimates from a set of simpler and weaker models. The XGBoost algorithm performs well in machine learning competitions because of its robust handling of a variety of data types, relationships, distributions, and the variety of hyper parameters that you can fine-tune. You can use XGBoost for regression, classification (binary and multiclass), and ranking problems.

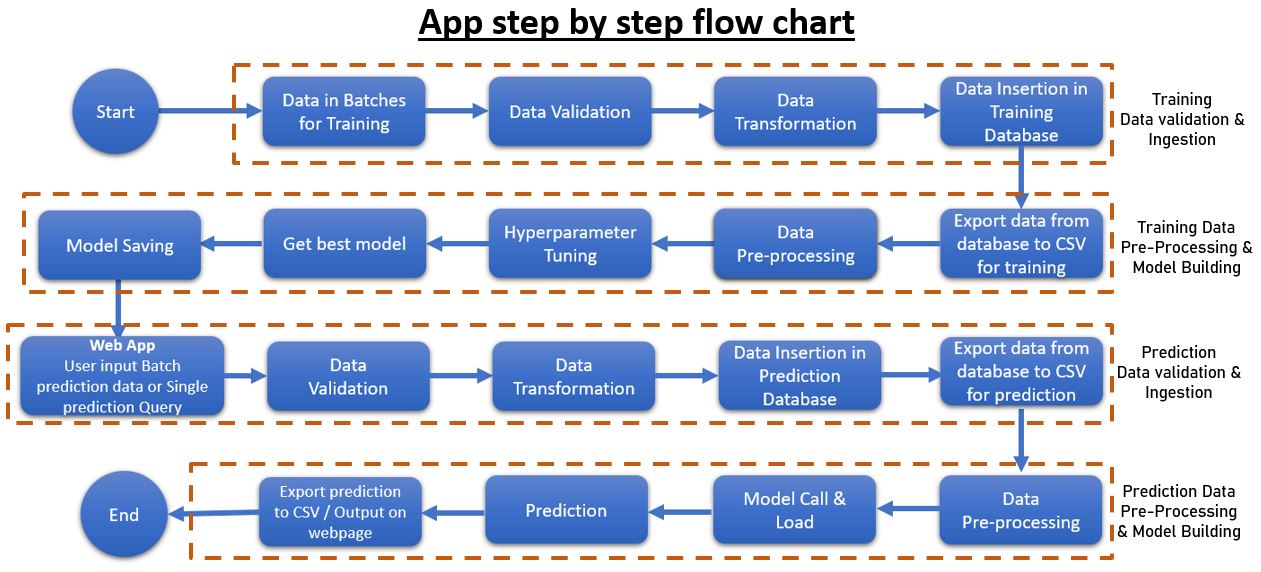
XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements Machine Learning algorithms under the Gradient Boosting framework. It provides a parallel tree boosting to solve many data science problems in a fast and accurate way.

**STEPS:**

1. Create a single leaf tree.
2. For the first tree, compute the average of target variable as prediction and calculate the residuals using the desired loss function. For subsequent trees the residuals come from prediction made by previous tree.
3. Using similarity score we select the appropriate node. Higher the similarity score more the homogeneity.
4. Using similarity score we calculate Information gain. Information gain gives the difference between old similarity and new similarity and thus tells how much homogeneity is achieved by splitting the node at a given point. https://www.csias.in/wp-content/uploads/2021/01/XGBoost-2.png
5. Create the tree of desired length using the above method. Pruning and regularization would be done by playing with the regularization hyper parameter.
6. Predict the  residual values using the Decision Tree you constructed.
7. Go back to step 1 and repeat the process for all the trees.

# CHAPTER 8

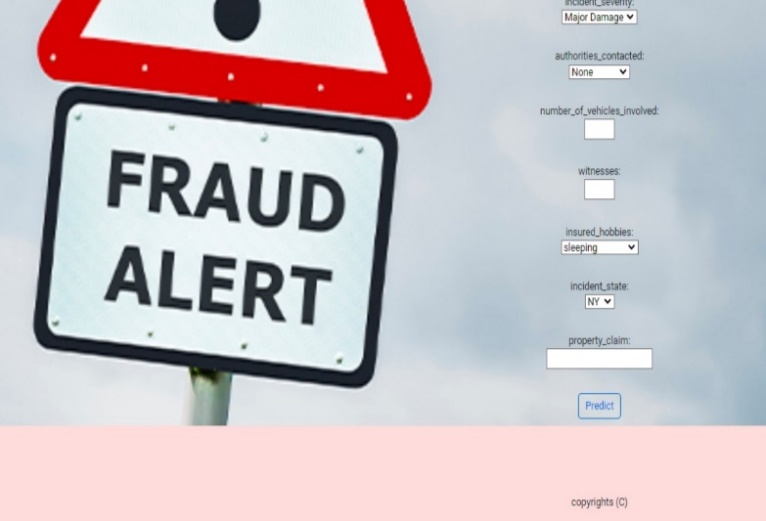
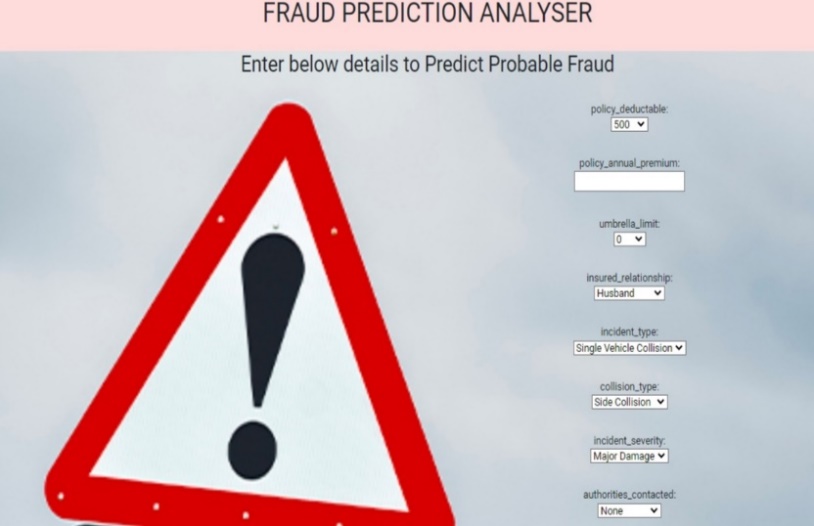
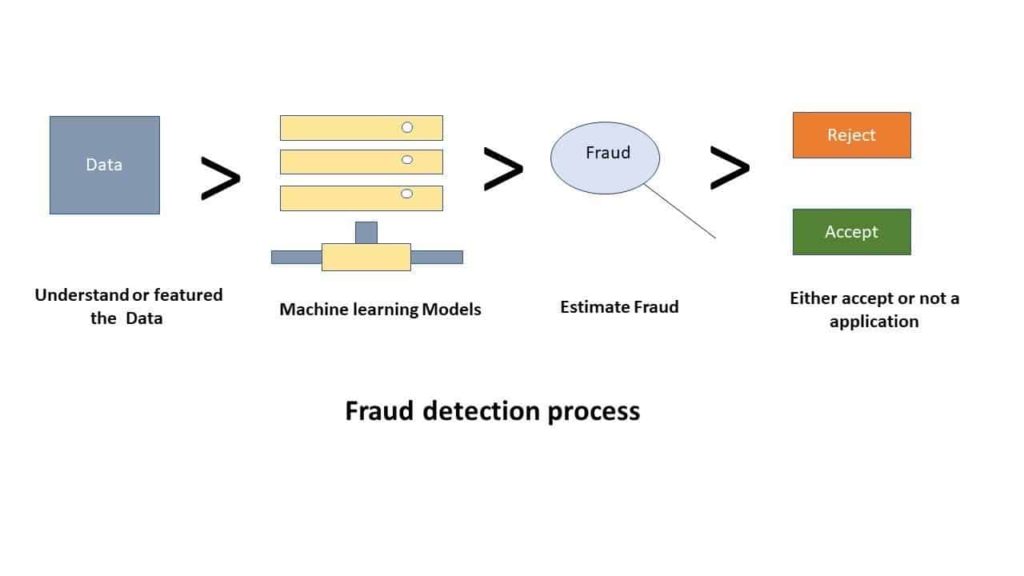
**PROJECT MODEL & DEPLOYMENT**



## FIG 11: APP FLOWCHART

## 

**FIG 12 : HISTOGRAM**

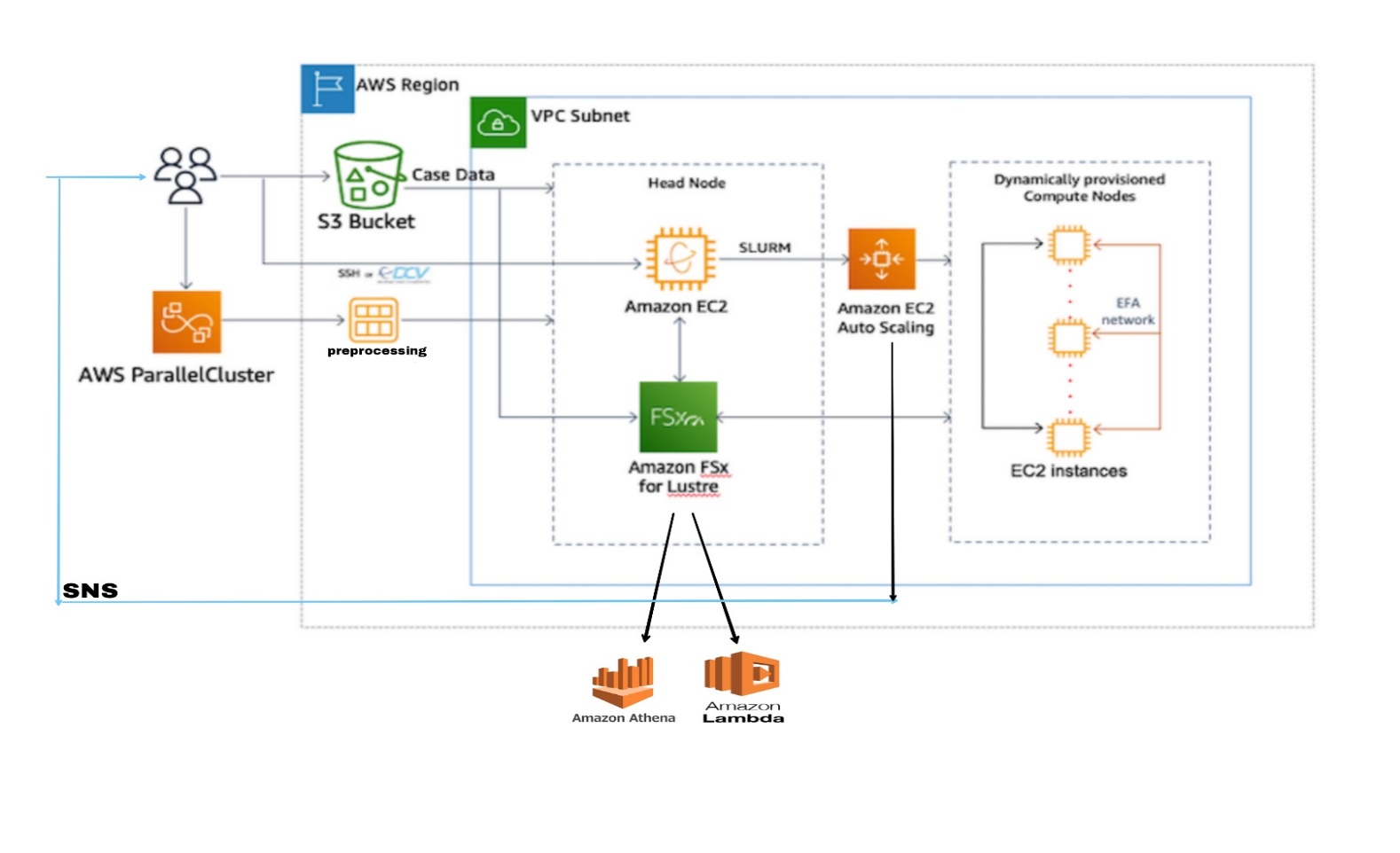
**FIG 13: WEB APP MODEL**

**FIG 14**

**DEPLOYMENT:**

Amazon Web Services (AWS) delivers reliable, scalable, and cost-effective computing resources on which to host our web applications. You can use the following AWS components alone or combined to host your application.

* Amazon Elastic Compute Cloud (Amazon EC2)
* Amazon Simple Storage Service (Amazon S3)
* Amazon Relational Database Service (Amazon RDS)
* Amazon CloudFront
* Amazon Simple Queue Service (Amazon SQS)
* Amazon DevPay

****

## FIG 15: AWS ARCHITECTURE FOR INSURANCE CLAIM FRAUD PREDICTION

**AWS SERVICES USED :**

1. **AWS PARALLELCLUSTER**

AWS ParallelCluster is an open source cluster management tool that makes it easy for you to deploy and manage High Performance Computing (HPC) clusters on AWS.

1. **AMAZON S3**

Amazon S3 to store and protect any amount of data for a range of use cases, such as data lakes, websites, mobile applications, backup and restore, archive, enterprise applications, IoT devices, and big data analytics.

1. **AMAZON EC2**

Amazon EC2 provides scalable computing capacity in the AWS Cloud. Using Amazon EC2 eliminates your need to invest in hardware up front, so you can develop and deploy applications faster.

1. **AMAZON EC2 AUTOSCALING**

It helps you maintain application availability and allows you to automatically add or remove EC2 instances according to conditions you define.

1. **AMAZON VPC**

Amazon VPC enables you to launch AWS resources into Amazon Virtual Private Cloud a virtual network and closely resembles a traditional network that you'd operate in your own data center, with the benefits of using the scalable infrastructure of AWS

1. **AMAZON FSx**

Amazon FSx makes it easy and cost effective to launch, run, and scale feature-rich, high-performance file systems in the cloud. In AMAZON FSx we use AWS LAMBDA & AWS ATHENA.

* AWS LAMBDA: It is built on the latest AWS compute, netwrking, and disk technologies to provide high performance and lower TCO.
* AWS ATHENA: It is an interactive query service that makes it easy to analyze data directly in Amazon S3 using standard SQL.

**CONCLUSION & FUTURE SCOPE**

**9.1. CONCLUSION**

Hence, we can conclude that machine learning is a most popular field among industry experts. After using this approach we can easily detect the potential frauds in the claims and also probability of fraud. By using this approach, we can early detect the frauds and this will be helpful in:

1. Reduction in fraud investigation expenses.
2. Lowers claim handling cost
3. Efficiently manages claims severity
4. Detection of early claims in the claim life cycle is paramount to managing overall claims costs.

An automated model to identify fraud claims in insurance industry. As explained in the results for XGBoost algorithm contain 94% which is the highest precision accuracy for fraud detection problem with machine learning data. Hence by the implementation of this model the insurance company can get accurate results in short duration of time. Thus, this automated framework can be used by any type of insurance company to reduce human labor and also to minimize the monetary loss in the insurance industry.

**9.2 FUTURE SCOPE:**

AI technology not only automates the fraud detection process but also identifies fraud patterns allowing early flagging and prompt response to any potential incidents.

As the number of clients increases, claims adjusters are put under higher pressure and should either sacrifice the accuracy or the speed of the claims process. On the contrary, the more data machine learning algorithms receive, the faster they provide accurate results.

Next, predictive analytics delivers way more accurate results than a human agent can do. As a result of processing big data, digital tools have more information to make decisions with never before seen accuracy.

By maximizing the use of technology and data analytics, insurers reduce the number of manual interventions in the claims management process. This reduces turnaround times and frees up insurance agents allowing them to focus on more valuable, high-impact tasks.

With more accurate fraud detection and reduced false positives made possible by AI technologies, insurers can decrease financial loss significantly. Also, by automating repetitive processes like fraud detection, you won’t need to increase your headcount as you scale up – which otherwise would have come with extra costs.

**SAMPLE CODE**

from flask import Flask, render\_template,request

import os

from data\_validation import dataValidation

from trainModel import trainingModel

from predDataVal import predDataVal

from predFromModel import predFromModel

from predSingleEntry import predFromRec

import  warnings

warnings.filterwarnings("ignore")

app = Flask(\_\_name\_\_)

@app.route('/')

def home():

    return render\_template('index.html')

def trainRouteClient():

    # get input train file path

    paths = "D:\Learning\Data Science\E2E Project\insuranceFraudDetection\Insurance-Claim-Fraud-Prediction-Web-App\Training batch files"+ "/"

    # Data Validation : Checking input data against input schema (as per agreement with client)

    if os.path.isdir(paths):

        # Input Batch file validate & export from DB for Training

        dataValidate = dataValidation(paths)

        dataValidate.train\_data\_val()

        # Model training

        modelTrain = trainingModel()

        modelTrain.trainingModel()

    else: print('Bad path dir')

# trainRouteClient()

@app.route('/predict', methods=['POST'])

def predictRouteClient():

    if request.method == 'POST':

        path = request.form['filepath']

            # path  = 'D:\Learning\Data Science\E2E Project\insuranceFraudDetection\Self\PredictionBatchFile'+ "/"

        if os.path.isdir(path):

            #Prediction Data Validation

            predictDataValid = predDataVal(path)

            predictDataValid.predDatavalInsertion()

            #Predicitons

            predicitor = predFromModel()

            res = predicitor.predModel()

            return render\_template('index.html',

             batchPredictionText='''Predcition Completed

            Output saved in predOutFile/Prediction.csv''')

            return res

        else: return 'Bad path dir'

@app.route('/singleRec', methods=['POST'])

def predRec():

    if request.method=='POST':

        predSinglerec = predFromRec()

        dictData = predSinglerec.getValues()

        res = predSinglerec.predFromRec(dictData)

        if (res == 'No'):

            return render\_template('index.html', prediction\_text='No, This Claim is not Fraud')

        else:

            return render\_template('index.html', prediction\_text= 'Yes, This calim is a Fraud')

            # return res

# predRec()

# predictRouteClient()

if \_\_name\_\_ == '\_\_main\_\_':

    app.run(debug=True)

**REFERENCE**

1. Detecting Insurance Claims Fraud Using Machine Learning (ML) Techniques, 19 OCTOBER 2017, IEEE

2. Robust Fuzzy Rule Based Technique To Detect Fraud In Vehicle Claim , 21 JUNE 2018,IEEE

3. Insurance Fraud Detection Using Machine Learning , APR 2021 ,IRJET.

4. Enhancing Claims Handling Processes With Insurance Based Language Models , FEB 19 2019,IEEE

5. Detecting Fraudulent Motor Insurance Claims Using Support Vector Machines With Adaptive Synthetic Sampling Method. 19 NOVEMBER 2020 , IEEE

6. Automobile Insurance Fraud Detection Using Supervised Classifiers, 21 NOVEMBER 2020, IEEE

7. Analysis Accuracy Of XGboost Model For Multiclass Classification - A Case Study Of Applicant Level Risk Prediction For Life Insurance, 10 FEBRUARY 2020,IEEE

8. A Data Mining Based Target Regression-Oriented Approach To Modelling Of Health Insurance Claims, 6 MAY 2021, IEEE

## 

# 