HEALTH INSURANCE PREMIUM PREDICTION USING ML

A PROJECT REPORT

Submitted by

K.Rahul (RA1911008020027)

A.Mohammed Thanish(RA1911008020045)

V.Abhishek Narayan (RA1911008020049)

Under the Guidance of

Dr.K.Danesh

(Asst.Professor, Department of Information Technology)

in partial fulfilment of the requirements for the degree

of

BACHELOR OF TECHNOLOGY

in

INFORMATION TECHNOLOGY

DEPARTMENT OF INFORMATION TECHNOLOGY

FACULTY OF ENGINEERING AND TECHNOLOGY

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

RAMAPURAM-600 089

MAY 2023

SRM INSTITUTE OF SCIENCE & TECHNOLOGY
RAMAPURAM, CHENNAI-600 089

BONAFIDE CERTIFICATE

The Certified Bachelor of Technology project report titled **"*Health Insurance Premium Prediction Using ML*"** is the bonafide work by ***K.Rahul (RA1911008020027), A.Mohammed Thanish (RA1911008020045) and V.Abhishek Narayan (RA1911008020049)*** carried out the project work under my/our supervision. Certified further, that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion for this or any other candidate.

SIGNATURE	SIGNATURE
Dr. K.DANESH, AP	Dr. RAJESWARI MUKESH
Supervisor	**Head of the Department**
SRM Institute of Science & Technology	**SRM Institute of Science Technology**
Ramapuram, Chennai – 600089	**Ramapuram, Chennai - 600089**

SUBMISSION OF MAJOR PROJECT REPORT FOR VIVA-VOICE HELD ON _____

INTERNAL EXAMINER	EXTERNAL EXAMINER

Annexure II

Department of Information Technology

SRM Institute of Science & Technology

Own Work* Declaration Form

To be completed by the student for all assessments

Degree/ Course : **Bachelor of Technology/ Information Technology**

Student Name : **K.Rahul, A.Mohammed Thanish, V.Abhishek Narayan**

Registration Number : **RA1911008020027, RA1911008020045, RA1911008020049**

Title of Work : **Health Insurance Premium Prediction Using ML**

I / We hereby certify that this assessment compiles with the University's Rules and Regulations relating to Academic misconduct and plagiarism**, as listed in the University Website, Regulations, and the Education Committee guidelines.

I / We confirm that all the work contained in this assessment is my / our own except where indicated, and that I / We have met the following conditions:

- Clearly references / listed all sources as appropriate.
- Referenced and put in inverted commas all quoted text (from books, web, etc)
- Given the sources of all pictures, data etc. that are not my own.
- Not made any use of the report(s) or essay(s) of any other student(s) either past or present.
- Acknowledged in appropriate places any help that I have received from others. (e.g., fellow students, technicians, statisticians, external sources)
- Compiled with any other plagiarism criteria specified in the Course handbook / University website.

I understand that any false claim for this work will be penalised in accordance with the University policies and regulations.

DECLARATION:
I am aware of and understand the University's policy on Academic misconduct and plagiarism and I certify that this assessment is my / our own work, except where indicated by referring, and that I have followed the good academic practices noted above.
If you are working in a group, please write your registration numbers and sign with the date for every student in your group.

ABSTRACT

Everyone is aware of how crucial health insurance is to maintaining human life. It covers a wide range of insurance-related medical costs for a person who is insured and has a number of medical difficulties, including disease, accidents, and surgery. It is known as the amount of business and individual people's inverse the amount to policies based insurance premium. Insurance premiums became the insurance company's source of income once we started making money. In an emergency, we will utilise that health insurance. In this paper, we propose using supervised machine learning, the best type for stochastic gradient descent optimisation algorithm and multi layer perceptron, to remove unwanted data (null values, load the dataset), predict whether or not people will purchase health insurance premiums, and load the dataset. The major goal of this project is to operate in real-time, thus we'll predict the accuracy score and heat map (confusion matrix) for it. The primary goal of this research is to forecast whether consumers will purchase health insurance premium plans using machine learning techniques and various types of algorithms.

Keywords- insurance, premium, dataset, gradient, forecast.

TABLE OF CONTENTS

S.NO	CONTENTS	PAGE NUMBER
1.	INTRODUCTION	1
2.	LITERATURE SURVEY 2.1 Overview 2.2 Literature Review 2.3 Disadvantages of Existing Work 2.4 Proposed Work 2.5 Summary	2 2 2 4 4 5
3.	HEALTH INSURANCE PREMIUM PREDICTION USING ML ARCHITECTURE	6
4.	HEALTH INSURANCE PREMIUM PREDICTION USING ML IMPLEMENTATION 4.1 Overview 4.2. Requirements 4.2.1 Software Requirement 4.2.2 Hardware Requirement 4.3 Module I 4.3.1 Dataset Collection 4.3.2 Pre-processing 4.3.3 Linear Regression Algorithm 4.4 Module II 4.5 Summary	7 7 7 7 7 7 8 8 8 9 15
5.	PERFORMANCE ANALYSIS 5.1 Overview 5.2 Performance Analysis 5.3 Summary	16 16 16 17
6.	CONCLUSION	18

7.	FUTURE ENHANCEMENT	19
8.	REFERENCES	20
	APPENDIX I.SAMPLE CODES II.SAMPLE OUTPUT	

LIST OF TABLES

S.NO	NAME OF THE TABLE	PAGE NUMBER
2.2	LITERATURE REVIEW	3
5.2.1	TABLE OF DIFFERENCE	17

LIST OF FIGURES

FIGURE NO	NAME OF THE FIGURES	PAGE NUMBER
3.1	ARCHITECTURE	6
4.3	MODULE DIAGRAM	9
4.4.1	UML DIAGRAM-USE CASE DIAGRAM	10
4.4.2	SEQUENCE DIAGRAM	11
4.4.3	CLASS DIAGRAM	12
4.4.4	ACTIVITY DIAGRAM	13
4.4.5	DFD	14
5.2.2	TENSORFLOW	20

LIST OF SYMBOLS AND ABBREVIATIONS

S.No	ACRONYM	Full form
1.	ML	Machine Learning
2.	DL	Deep Learning
3.	HR	Human Resources
4.	AI	Artificial Intelligence
5.	SARSA	State-Action-Reward-State-Action
6.	DDPG	Deep Deterministic Policy Gradient

CHAPTER 1

INTRODUCTION

We inhabit a planet that is rife with perils and unknowns. All individuals, homes, businesses, buildings, and property are susceptible to different kinds of danger. and the danger level can change. The potential hazards include loss of life, disease, and property. assets. Health and happiness are the two most significant components of people's lives. However, the financial industry has developed a wide range of measures to protect you because risks can't always be avoided. By using people and organizations. Therefore, having insurance is necessary.the importance of insurance in people's lives and a policy that lowers or eliminates the costs of losses caused by various risks. It is critical for organizations to comprehend what customers demand. It's crucial for insurance companies to be able to precisely measure or quantify the amount covered by the policy and the required insurance payments when it comes to the value of insurance in people's lives.

CHAPTER 2

LITERATURE REVIEW

2.1 Overview

Currently, everything is carried out manually . The insurance agent goes to the clients and then he collects the client's personal details and then based on that , he gives policy suggestions . Based on their age and sex, the insurance agents can provide policy suggestions. Data Mining is an emerging technology in today's world. It is nothing but collecting data , analyzing data , finding patterns hidden in the data and then making predictions. Using these techniques , the purchase patterns of the customers are analyzed and policy suggestions are given to the customers. All this involves a lot of paperwork and consumes a lot of human effort and time. This method increases the overhead on the policy owners as well as the customers. The insurance agent needs to travel to the place where the customer lives , needs to collect all the details and he needs to maintain all the records. Similarly , the customer needs to sit with the insurance agent , discuss with them about the various policies , pros and cons of various policies , analyze them and choose the best policy which suits their needs.

2.2 Literature Review

Today, a considerable amount of data is being produced in numerous industries. We have a lot of work ahead of us since we must gather data, analyze it, look for hidden patterns, and then predict the future. Different machine learning algorithms are being used to complete all of these tasks. The ability of machines to learn on their own without being explicitly taught is known as machine learning. supervised learning and unsupervised learning are the two categories of machine learning techniques that are available. It is referred to as supervised learning whenever we train machine learning models using labels.

By giving the machine learning models some tagged data, they can be taught. in the event that some unknown data

In the below presented article, we provided an automated model to identify premiums in the insurance business. The results for the multi-layer perceptron and stochastic gradient descent optimization algorithms, which are described in the section, are the highest precision accuracy for the problem of insurance premium utilizing machine learning data. The insurance business

can use this model as a consequence and get precise results rapidly. In order to minimize financial loss in the insurance sector and decrease the need for human labor, any insurance company can adopt this automated framework.

s.no	Year	Title	author	Algorithm	Disadvantages
1	2017	An Ensemble Random Forest Algorithm for Insurance Big Data Analysis	WEIWEI LIN1,2, ZIMING WU1, LONGXIN	Random Forest	A forest is less interpretable than a single decision tree
2	2019	An Adaptive Estimation of Distribution Algorithm for Multipolicy	Wen Shi, Student Member, IEEE, Wei-Neng Chen	EDA	Anticipating the Unforeseen
3	2020	Research on the UBI Car Insurance Rate Determination Model Based on the CNN-HVSVM Algorithm	CHUN YAN1, XINDONG WANG 1, XINHONG LIU2,	CNN	A small change in the dataset can make the tree structure unstable which can cause variance
4	2022	Blockchain and AI-Empowered Healthcare Insurance Fraud Detection: An Analysis, Architecture, and Future Prospects	KHYATI KAPADIYA1, USHA PATEL, RAJESH GUPTA, MOHAMMAD	SVM	Choosing a "good" kernel function is not easy
5	2021	Clustering Application for	Tallal Omar,	K-means	Require high memory – need to store all of the training data

Fig 2.2 Literature Review

2.3 Disadvantages of Existing Work

- Sequence mining is the system which is used here for enhancement.
- Most outlier identification algorithms fall into one of two categories: The second class of algorithms considers the data as the sequence in creating the model, while the first class of algorithms concentrates on the identification of abnormalities in individual data points. These approaches are not advised in a real-time setting.
- Health insurance premium fraud detection methodology can be further improved with the design of computerized physician order entry system for each disease using machine learning techniques.
- Upon finding a set of sequences for every disease, provider level fraud detection would be more effective.

2.4 Proposed Work

The suggested system's primary goal is to develop a smart system that uses multiple Machine learning and Deep learning approaches to anticipate the cost of the insurance policy. The cost of the insurance policy is always fluctuating based on the needs of the consumer. The cost of the insurance policy is predicted using a variety of supervised machine learning methods. An example of a regression problem is predicting the cost of insurance policies. In order to forecast the cost of the insurance policy, a regression approach known as the Linear Regression approach is employed in this research. This algorithm forecasts the statistical association between the dependent and independent variables, and then it makes recommendations. The project's several modules include dataset collecting.

Advantages:

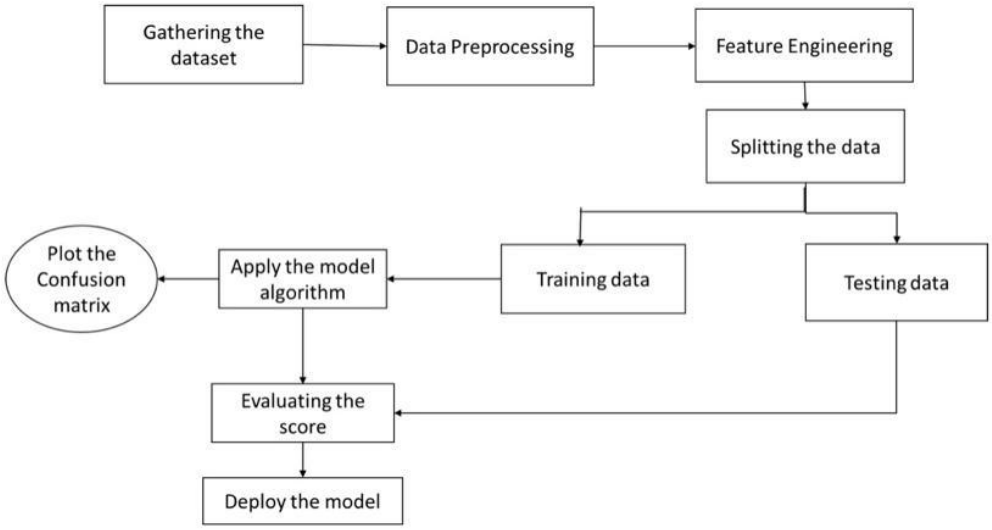
- Accuracy will be compared and analyzed.
- Time consumption is less for classifier algorithms.

2.5 Summary

In the dataset module , the insurance datasets are collected from various sources. Once the datasets are collected , in the next phase , scaling techniques and preprocessing techniques are applied in order to scale the values and remove the unwanted null values. After pre-processing , the next module is the feature extraction module. In this module , the essential features are extracted.

CHAPTER 3

HEALTH INSURANCE PREMIUM PREDICTION USING ML ARCHITECTURE



- Gathering the dataset and formatting it accordingly based on the data stored in the system.
- Data preprocessing is preprogramming the system for making it suitable for the developer and the user level usage.
- Splitting the data is the data allocation based on its feature, quality and mode of originality. It is of two kinds: Training data and Testing data.
- Evaluating the score is the evaluation of the performance of the system during the user implementation phase.
- Deploying the model is the deployment of the system based on user interface and user requirements.

CHAPTER 4

HEALTH INSURANCE PREMIUM PREDICTION USING ML IMPLEMENTATION

4.1 Overview

This chapter briefly demonstrates about the health insurance, premium production, using ML implementation, requirements, software requirements and hardware requirements along with the modules subdivided into data collection, pre-processing, and linear regression algorithm.

4.2 Requirements

4.2.1 Software Requirement

i. Windows OS
ii. Python GUI or Anaconda Navigator
iii. **Operating System:** Windows 7 Ultimate 32 bit / Windows XP

4.2.2 Hardware Requirement

i. 1 GB RAM
ii. 80 GB Hard Disk
iii. Intel Processor
iv. LAN

4.3 Module I

This project consists of four main modules:

● Dataset Collection

● Pre-Processing

● Feature Extraction

● Linear Regression Algorithm

● Evaluation

Dataset Collection:

In this module , the insurance datasets are collected from various datasources . The datasets are stored in the form of a csv file. Data is stored in a tabular format in a csv file . The first step in any machine learning project is that we need to load the datasets. The datasets are loaded by using the pandas object. The quality of the dataset is very important in order to train the machine learning model so that it will predict the insurance prices properly.

Preprocessing:

The next module is the preprocessing module . This module is used for Removing any null values present in a dataset and the values are encoded Using the LabelEncoder. Some algorithms need the data to be scaled in .The range of 0 to 1. The data is scaled by using MinMaxScaler. This Process is required in order to clean the data or remove the unwanted Noise value from the data.

Feature Extraction:

This module is useful for extracting the essential features needed for Training the Machine Learning model. The independent and dependent variables are separated and they are stored in x and y values respectively. Then, using the train_test_split module of the Scikit-learn library , the data is split into training data and testing data. All the extracted features are given as input to train the machine learning model.

Linear Regression Algorithm:

Whenever the predicted output is a continuous variable , then it is called a regression problem. Linear regression algorithm makes predictions based on the statistical relationship between the independent and the dependent variables. Various input features like name , age ,sex and

smoking status are given as input to the machine learning algorithm and the algorithm predicts the insurance prices based on the input features.

Evaluation:

After prediction , the final module is the evaluation module . The accuracy score is calculated in order to evaluate the performance of the model. It is a performance score to evaluate how well the model has predicted the insurance prices based on the features given to the machine learning Model.

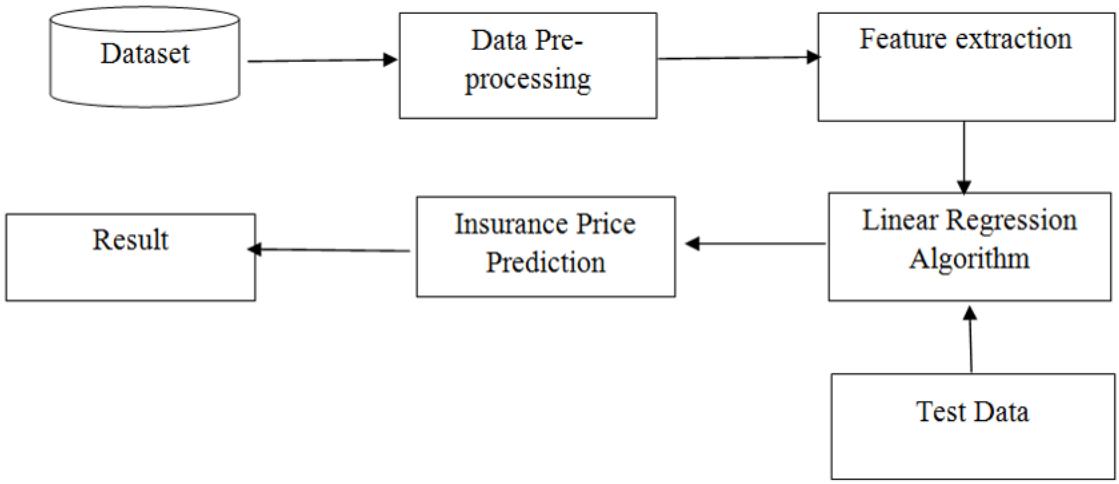


Fig 4.3 Module Diagram

4.4 Module II

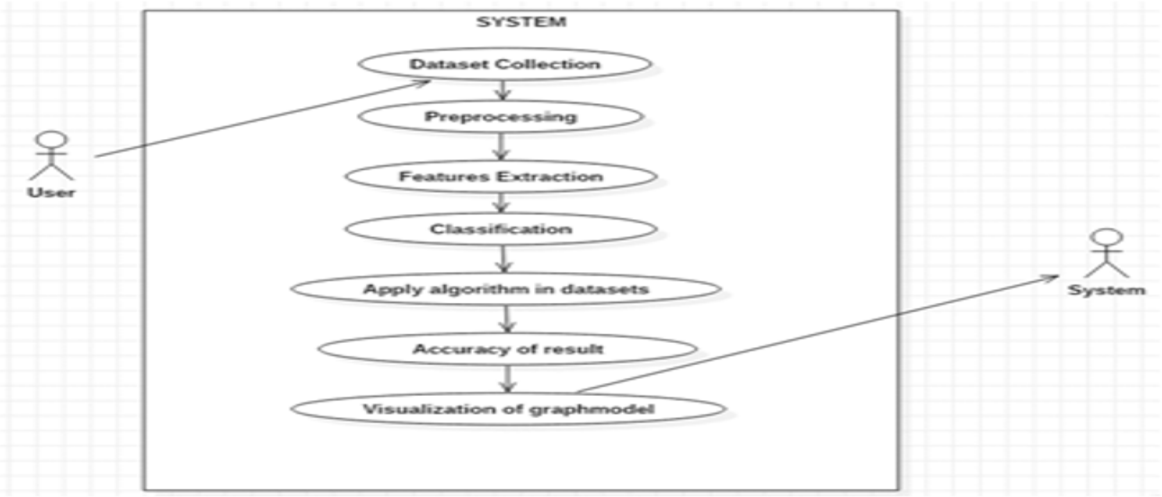


Fig 4.4.1 UML Diagrams-Use Case Diagrams

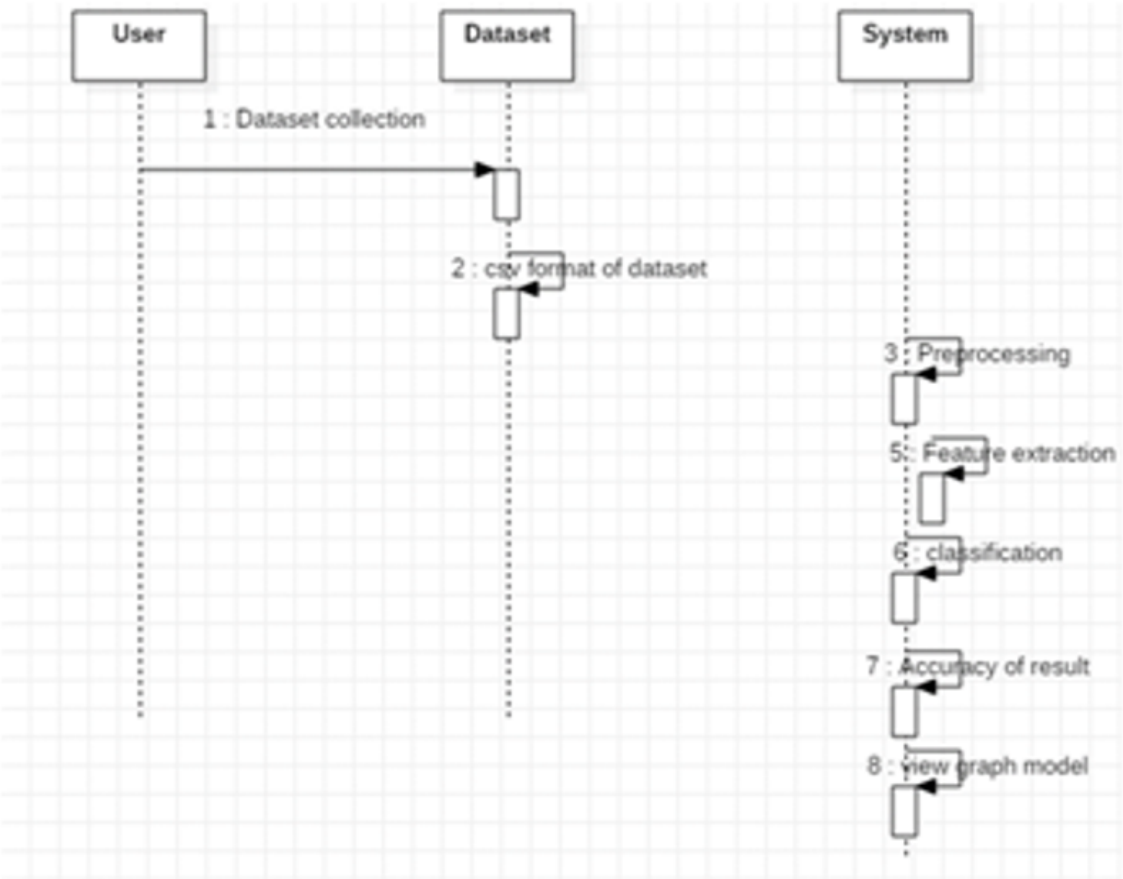


Fig 4.4.2 Sequence Diagram

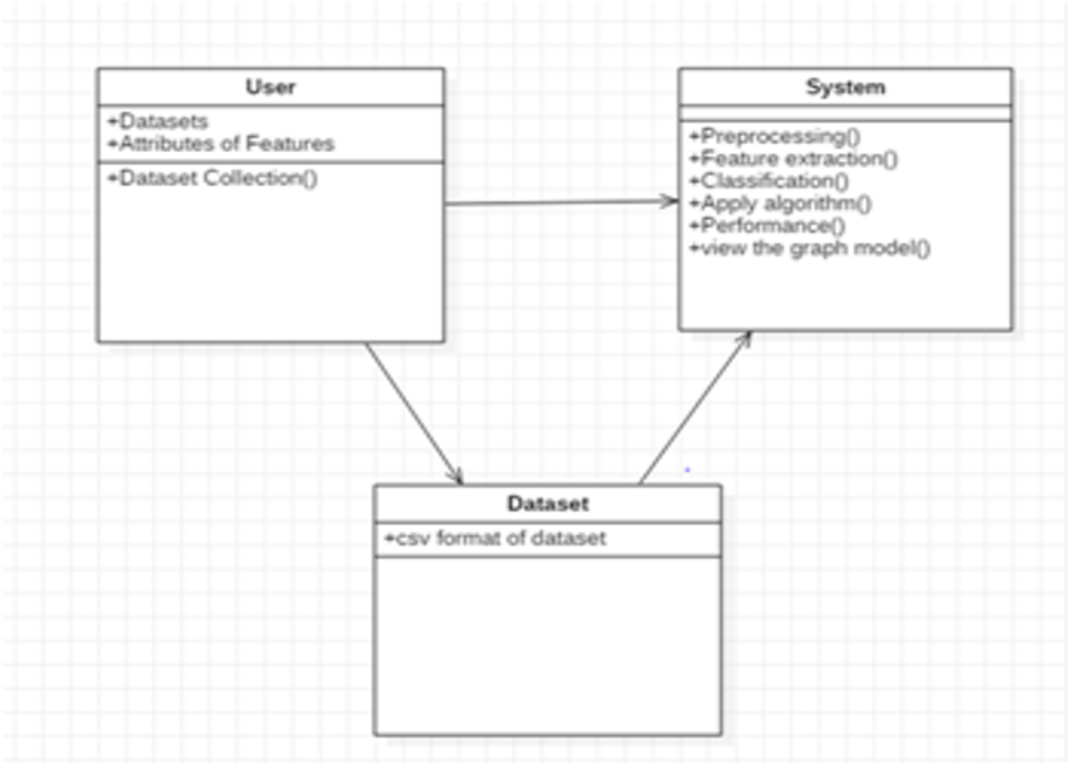


Fig 4.4.3 Class Diagram

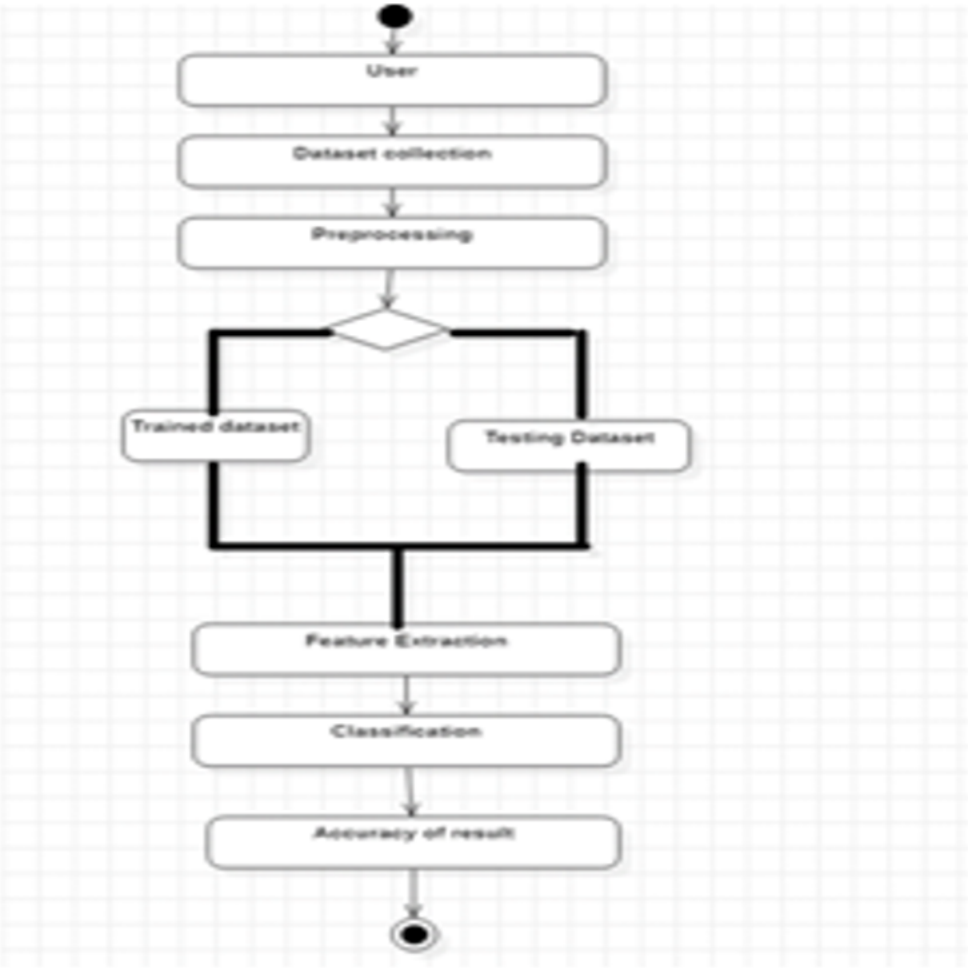
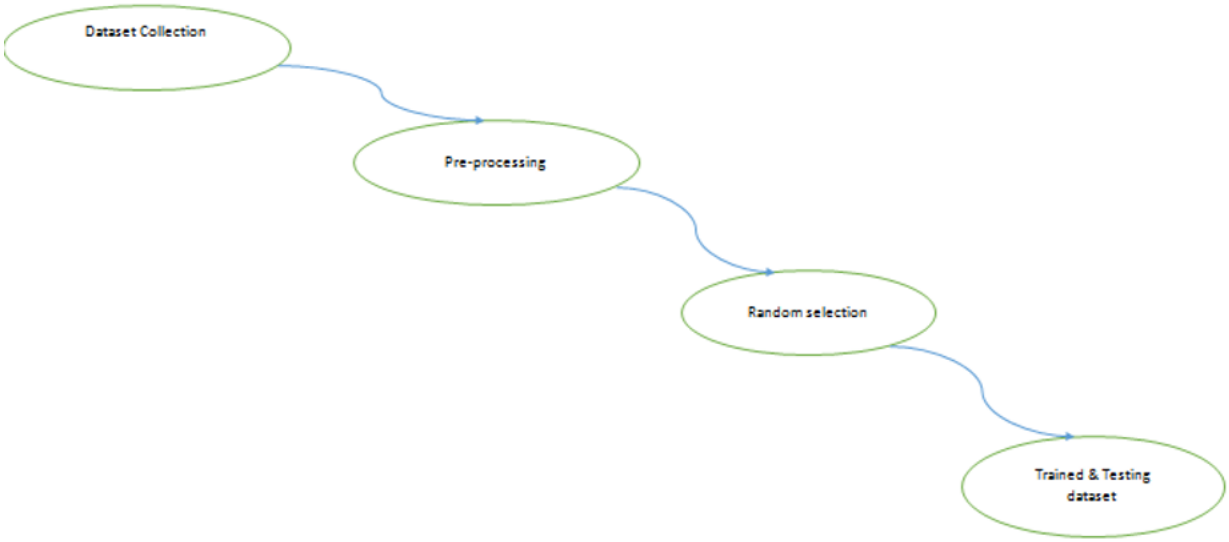
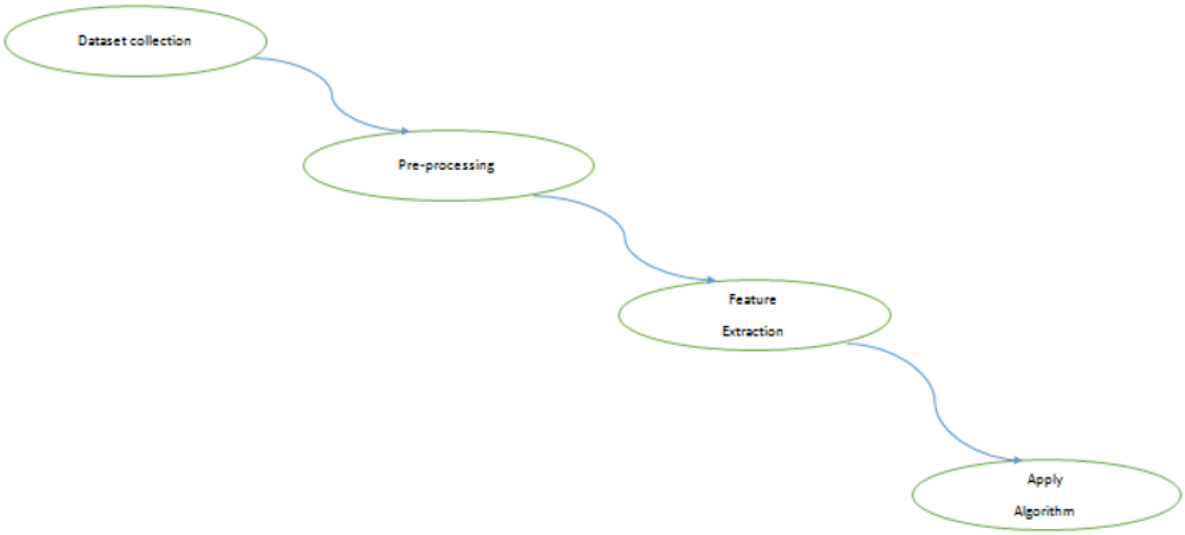


Fig 4.4.4 Activity Diagram

Level 0:



Level 1:



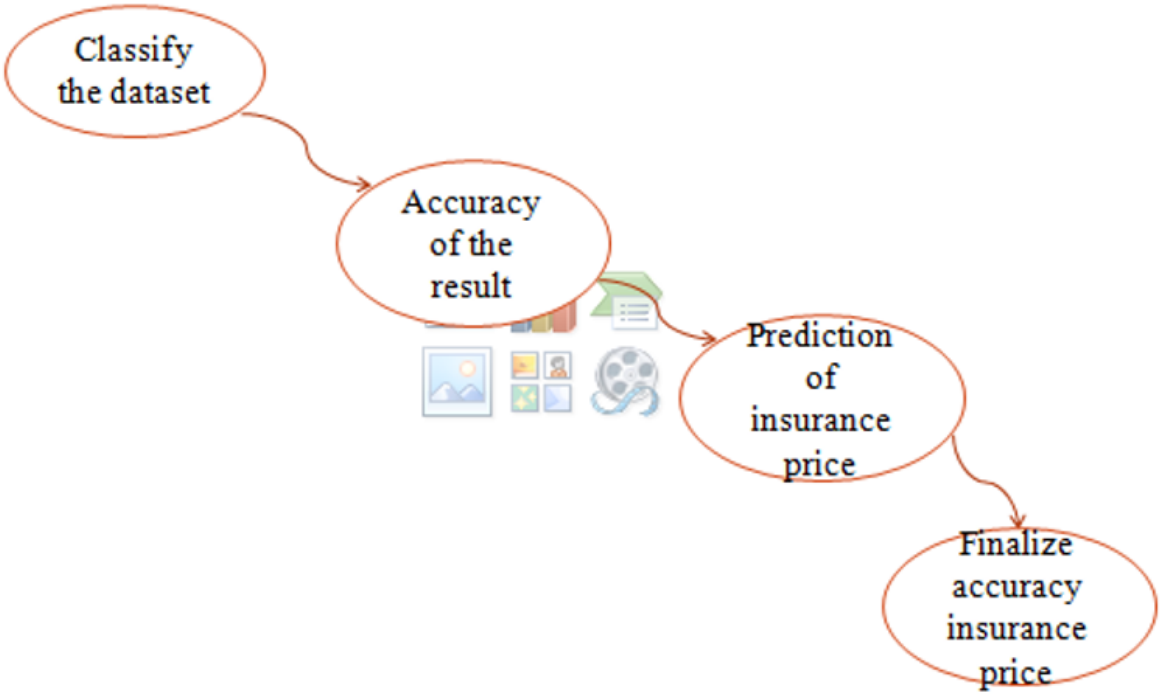


Fig 4.4.5 DFD

4.5 Summary

AI is a valuable material for customer demand management and personalization tasks. Improved speech recognition in BPOs and call forwarding as a result of AI techniques allows a more seamless experience for customers.

For example, deep learning is used to analyze the customer's emotions. If the customer is not satisfied then the AI changes its plan of action.

CHAPTER 5
PERFORMANCE ANALYSIS

5.1 Overview

This chapter briefly describes the applicational performance of the system by a comparative study between the previously available version or software with that of the currently proposed methodology.

5.2 Performance Analysis

In the table below, we summarize the difference between machine learning and deep learning.

	Machine learning	Deep learning
Training dataset	Minimum	Maximum
Choose features	Available	Not Available
Number of algorithms	Several	less
Training time	Short period	Long period

Fig 5.2.1 Table of Difference

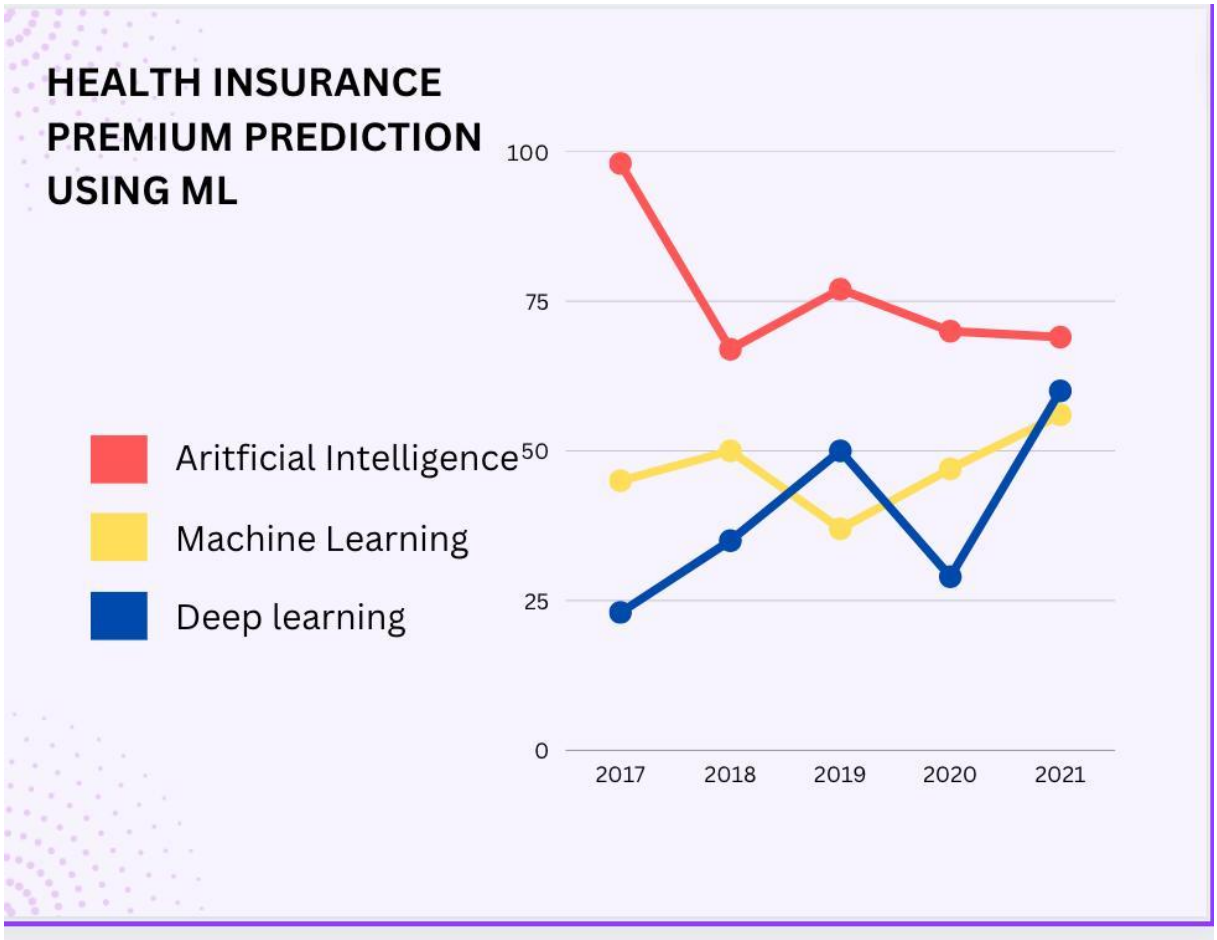


Fig 5.2.2 TensorFlow

5.3 Summary

The performance depends on the comparative study between the earlier existence and currently proposed methodology. The comparison displays the proposed system is effectively processed. It performs as

CHAPTER 6

CONCLUSION

Thus , through this project , insurance prices are predicted using various Machine learning techniques. Linear Regression algorithm is employed in Order to predict the prices of various insurance policies and medical Insurance policies. Various parameters are considered in order to predict the price of insurance policies like age , gender ,sex and smoking status. Then , finally after making predictions , the accuracy score is calculated and the overall performance of the system is evaluated based on the accuracy score.

CHAPTER 7

FUTURE ENHANCEMENT

In future , the project can be further enhanced by using some other Algorithms like logistic regression instead of linear regression algorithm. Similarly , one or more algorithms can be combined in order to see how well the model predicts and the result is shown in the form of bar graphs by comparing the algorithm's performance.

- In future , the project can be further enhanced by using some other Algorithms like logistic regression instead of linear regression algorithm. Similarly , one or more algorithms can be combined in order to see how well the model predicts and the result is shown in the form of bar graphs by comparing the algorithm's performance.
- The project's security goals will be improved in subsequent development, and the project will be implemented on the blockchain. It is practical to utilise this module to extract the crucial elements required for
- the machine learning model's training. Separated and kept in x and y values, respectively, are the independent and dependent variables.
- The Data is then divided into training data and testing data using the Scikit-learn library's train_test_split function. The machine learning model is trained using all the retrieved features as input.

REFERENCES

[1]Kowshalya, G., & Nandhini, M. (2018, April). Predicting fraudulent claims in automobile insurance.
In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT) (pp. 1338-1343). IEEE.

[2] Kayri, M., Kayri, I., & Gencoglu, M. T. (2017, June). The performance comparison of multiple linear
regression, random forest and artificial neural network by using photovoltaic and atmospheric data. In
2017 14th International Conference on Engineering of Modern Electric Systems (EMES) (pp. 1-4). IEEE.

[3] Denuit, Michel & Hainaut, Donatien & Trufin, Julien. (2019). Effective Statistical Learning Methods
for Actuaries I: GLMs and Extensions. 10.1007/978-3-030-25820-7.

[4] Breiman, Leo. 2001. ―Random Forests.‖ Machine Learning 45 (1). Springer: 5–32.

[5] Chen, T., & Guestrin, C. (2016). XGBoost: a scalable tree boosting system 22nd ACM SIGKDD Int.In Conf. on Knowledge Discovery and Data Mining

APPENDIX

report new

by Danesh K

Submission date: 12-May-2023 04:05AM (UTC-0400)
Submission ID: 2091173443
File name: Content_pgs.docx (1.06M)
Word count: 1548
Character count: 8395

CHAPTER 1

INTRODUCTION

We inhabit a planet that is rife with perils and unknowns. All individuals, homes, businesses, buildings, and property are susceptible to different kinds of danger. and the danger level can change. The potential hazards include loss of life, disease, and property. assets. Health and happiness are the two most significant components of people's lives. However, the financial industry has developed a wide range of measures to protect you because risks can't always be avoided. By using people and organizations. Therefore, having insurance is necessary.the importance of insurance in people's lives and a policy that lowers or eliminates the costs of losses caused by various risks. It is critical for organizations to comprehend what customers demand. It's crucial for insurance companies to be able to precisely measure or quantify the amount covered by the policy and the required insurance payments when it comes to the value of insurance in people's lives.

2.4 Proposed Work

The suggested system's primary goal is to develop a smart system that uses multiple Machine learning and Deep learning approaches to anticipate the cost of the insurance policy. The cost of the insurance policy is always fluctuating based on the needs of the consumer. The cost of the insurance policy is predicted using a variety of supervised machine learning methods. An example of a regression problem is predicting the cost of insurance policies. In order to forecast the cost of the insurance policy, a regression approach known as the Linear Regression approach is employed in this research. This algorithm forecasts the statistical association between the dependent and independent variables, and then it makes recommendations. The project's several modules include dataset collecting.

Advantages:

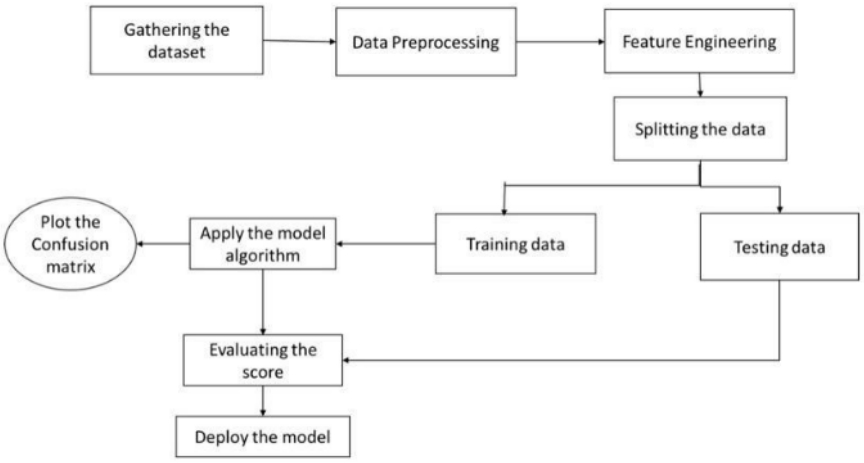
- Accuracy will be compared and analyzed.
- Time consumption is less for classifier algorithms.

2.5 Summary

In the dataset module , the insurance datasets are collected from various sources. Once the datasets are collected , in the next phase , scaling techniques and preprocessing techniques are applied in order to scale the values and remove the unwanted null values. After pre-processing , the next module is the feature extraction module. In this module , the essential features are extracted.

CHAPTER 3

HEALTH INSURANCE PREMIUM PREDICTION USING ML ARCHITECTURE



- Gathering the dataset and formatting it accordingly based on the data stored in the system.
- Data preprocessing is preprogramming the system for making it suitable for the developer and the user level usage.
- Splitting the data is the data allocation based on its feature, quality and mode of originality. It is of two kinds: Training data and Testing data.
- Evaluating the score is the evaluation of the performance of the system during the user implementation phase.
- Deploying the model is the deployment of the system based on user interface and user requirements.

CHAPTER 4

HEALTH INSURANCE PREMIUM PREDICTION USING ML IMPLEMENTATION

4.1 Overview

This chapter briefly demonstrates about the health insurance, premium production, using ML implementation, requirements, software requirements and hardware requirements along with the modules subdivided into data collection, pre-processing, and linear regression algorithm.

4.2 Requirements

4.2.1 Software Requirement

i. Windows OS
ii. Python GUI or Anaconda Navigator
iii. **Operating System:** Windows 7 Ultimate 32 bit / Windows XP

4.2.2 Hardware Requirement

i. 1 GB RAM
ii. 80 GB Hard Disk
iii. Intel Processor
iv. LAN

4.3 Module I

This project consists of four main modules:

- Dataset Collection
- Pre-Processing
- Feature Extraction
- Linear Regression Algorithm

● Evaluation

Dataset Collection:

In this module , the insurance datasets are collected from various datasources . The datasets are stored in the form of a csv file. Data is stored in a tabular format in a csv file . The first step in any machine learning project is that we need to load the datasets. The datasets are loaded by using the pandas object. The quality of the dataset is very important in order to train the machine learning model so that it will predict the insurance prices properly.

Preprocessing:

The next module is the preprocessing module . This module is used for Removing any null values present in a dataset and the values are encoded Using the LabelEncoder. Some algorithms need the data to be scaled in .The range of 0 to 1. The data is scaled by using MinMaxScaler. This Process is required in order to clean the data or remove the unwanted Noise value from the data.

Feature Extraction:

This module is useful for extracting the essential features needed for Training the Machine Learning model. The independent and dependent variables are separated and they are stored in x and y values respectively. Then, using the train_test_split module of the Scikit-learn library , the data is split into training data and testing data. All the extracted features are given as input to train the machine learning model.

Linear Regression Algorithm:

Whenever the predicted output is a continuous variable , then it is called a regression problem. Linear regression algorithm makes predictions based on the statistical relationship between the independent and the dependent variables. Various input features like name , age ,sex and smoking

status are given as input to the machine learning algorithm and the algorithm predicts the insurance prices based on the input features.

Evaluation:

After prediction , the final module is the evaluation module . The accuracy score is calculated in order to evaluate the performance of the model. It is a performance score to evaluate how well the model has predicted the insurance prices based on the features given to the machine learning Model.

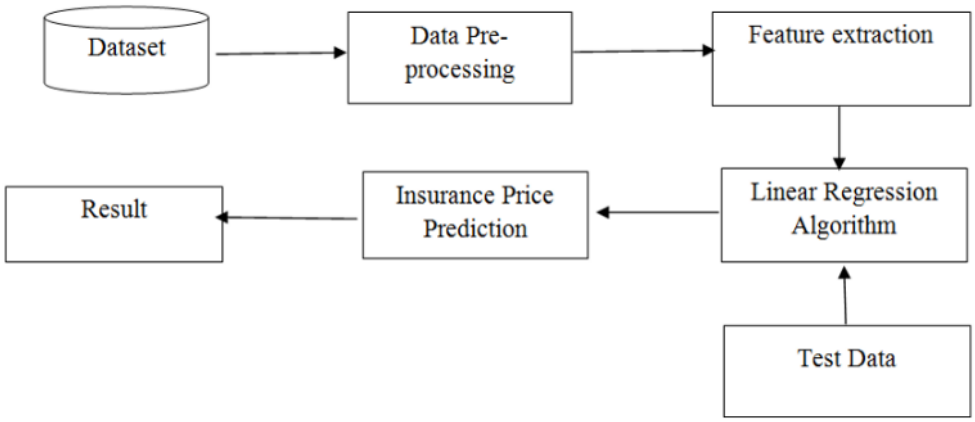


Fig 4.3 Module Diagram

4.4 Module II

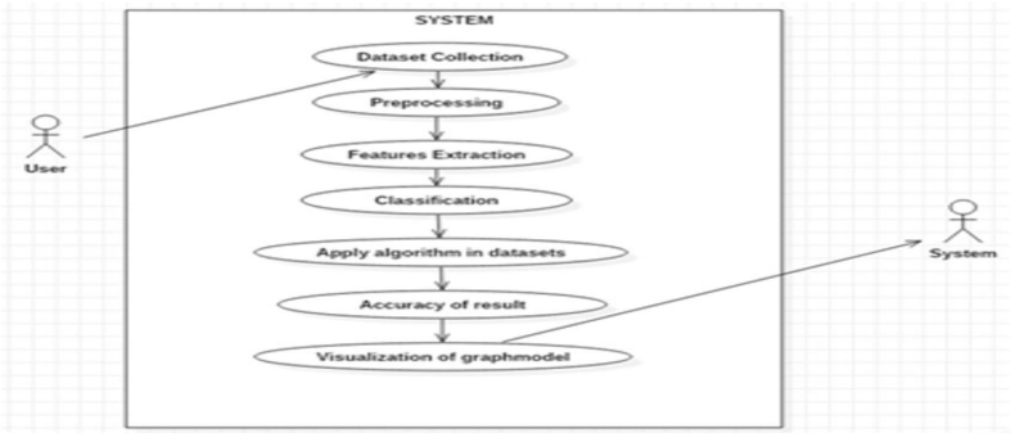


Fig 4.4.1 UML Diagrams-Use Case Diagrams

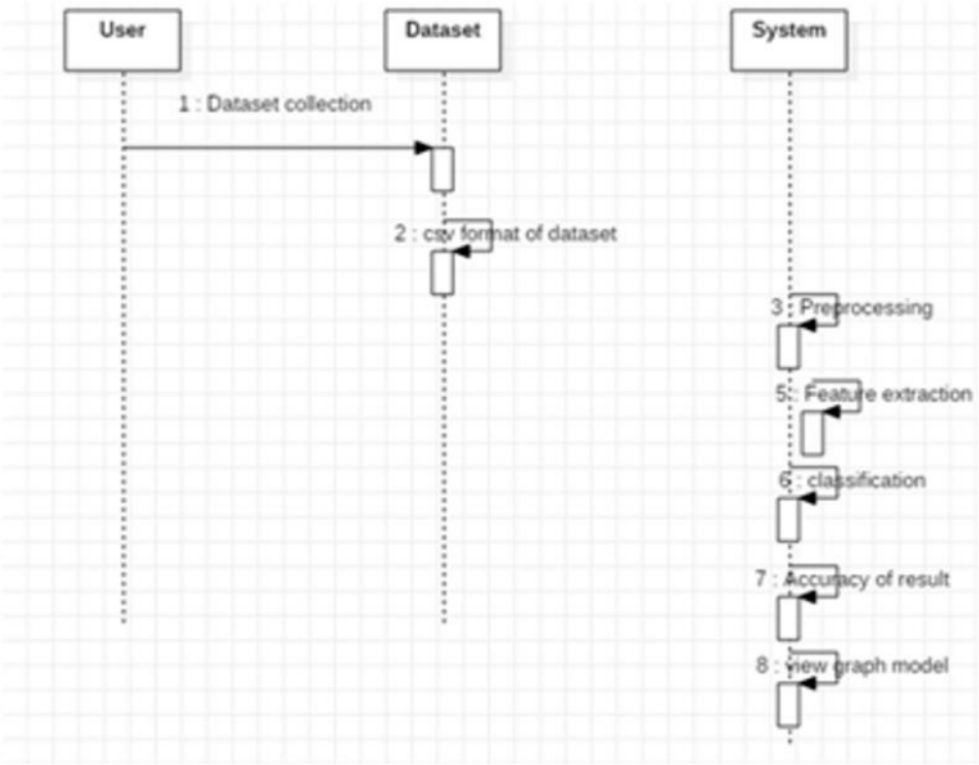


Fig 4.4.2 Sequence Diagram

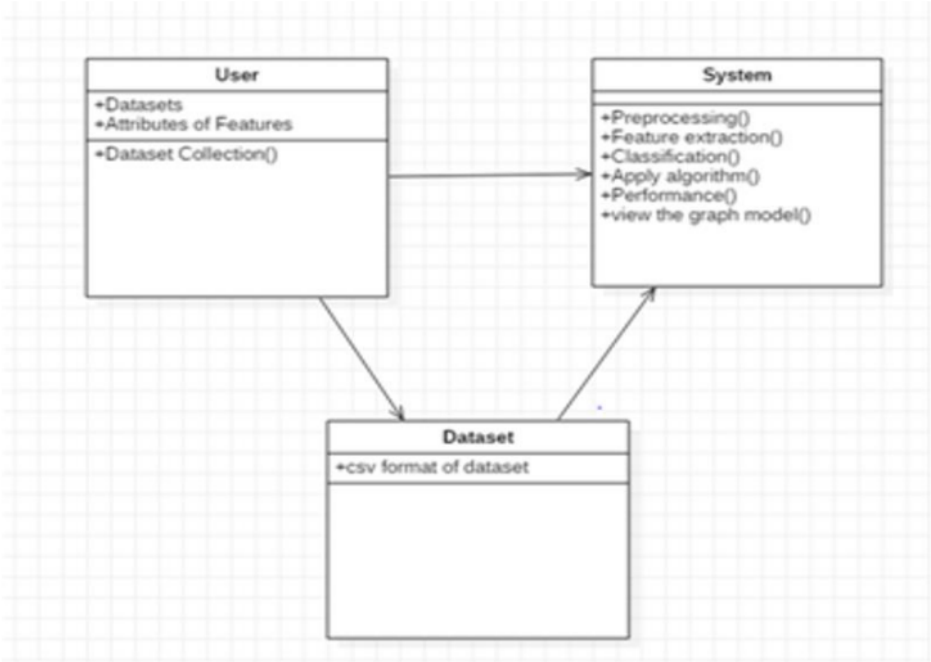


Fig 4.4.3 Class Diagram

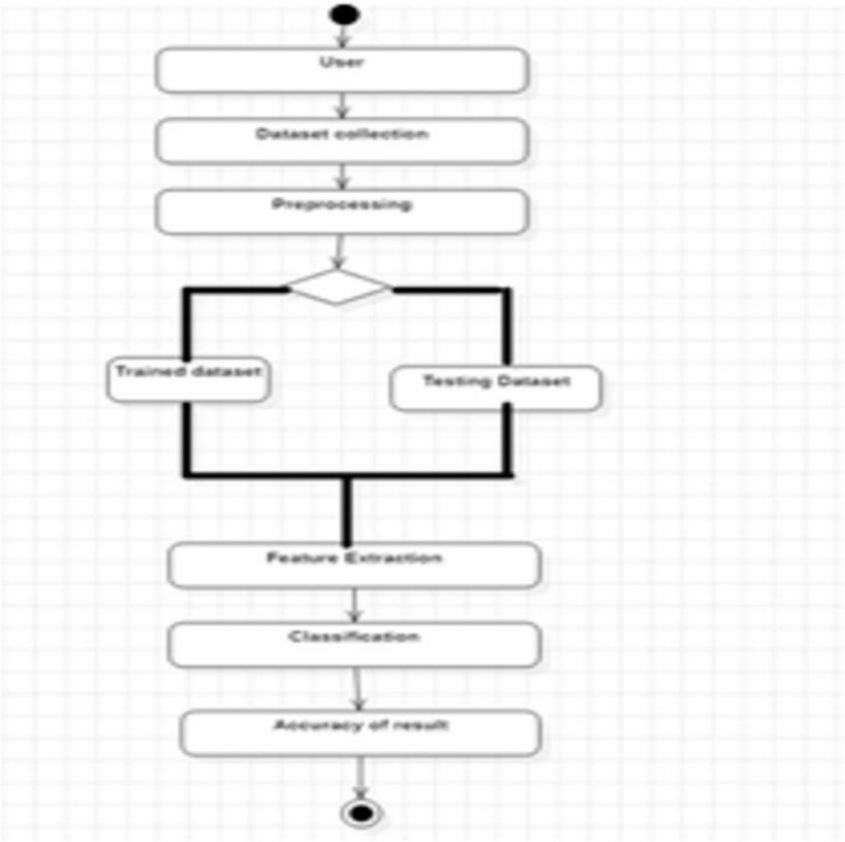
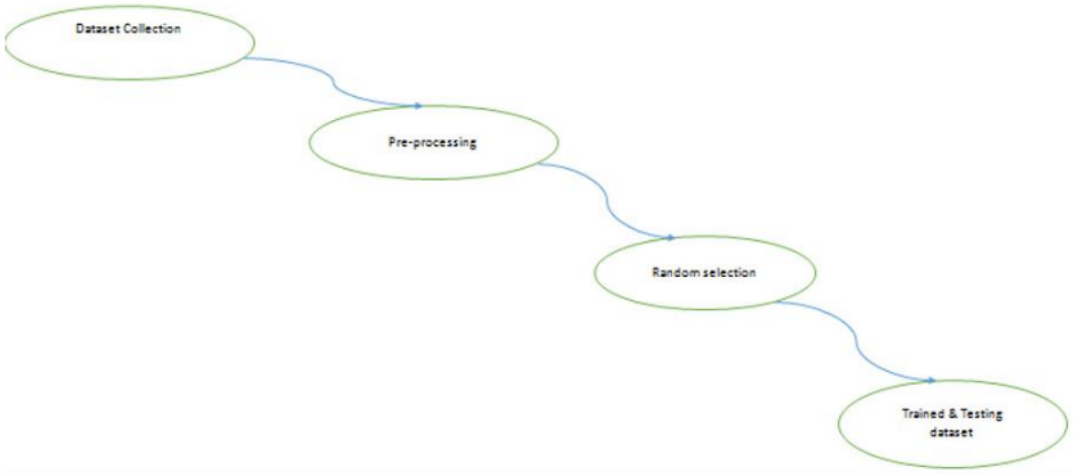
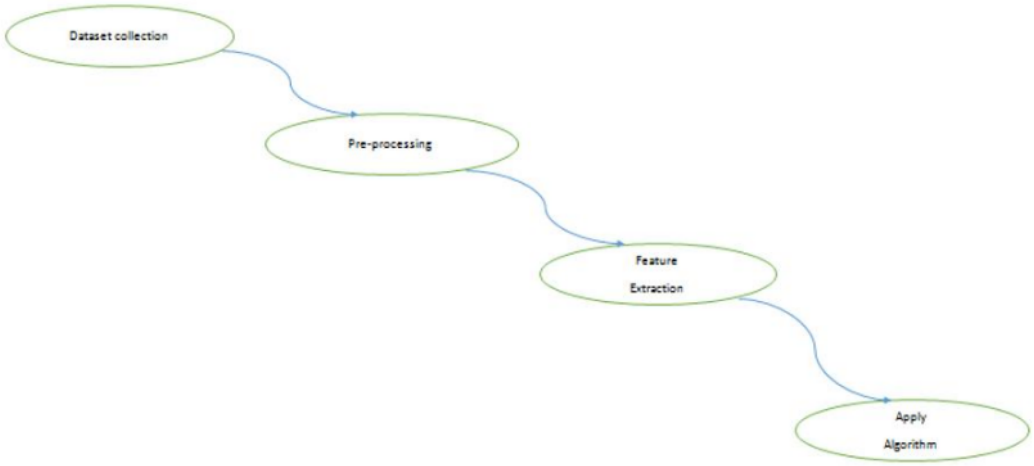


Fig 4.4.4 Activity Diagram

Level 0:



Level 1:



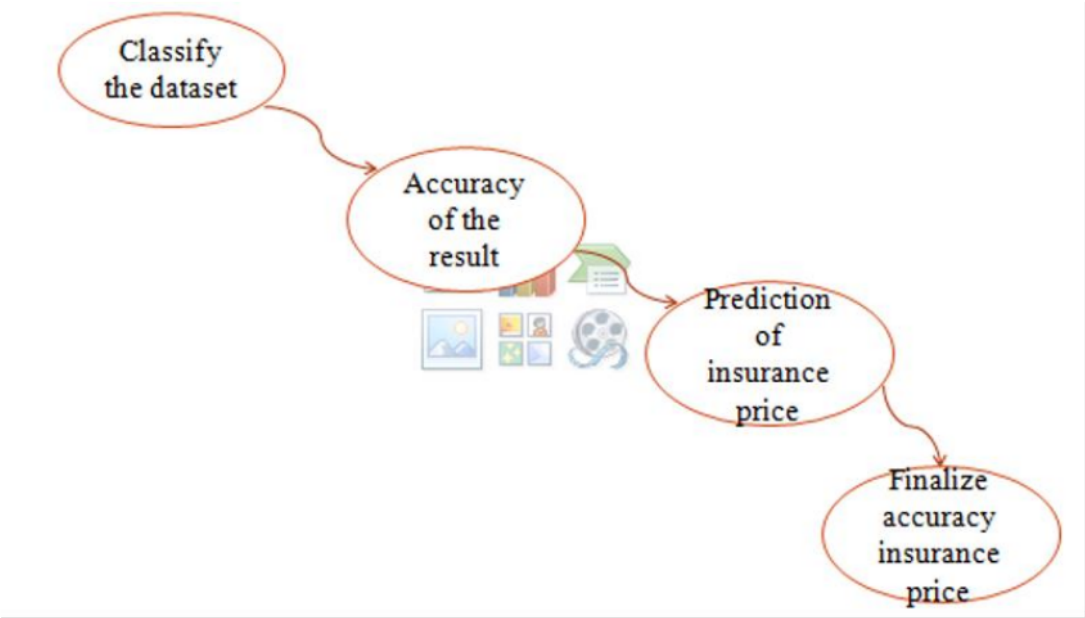


Fig 4.4.5 DFD

4.5 Summary

AI is a valuable material for customer demand management and personalization tasks. Improved speech recognition in BPOs and call forwarding as a result of AI techniques allows a more seamless experience for customers.

For example, deep learning is used to analyze the customer's emotions. If the customer is not satisfied then the AI changes its plan of action.

CHAPTER 5
PERFORMANCE ANALYSIS

5.1 Overview

This chapter briefly describes the applicational performance of the system by a comparative study between the previously available version or software with that of the currently proposed methodology.

5.2 Performance Analysis

In the table below, we summarize the difference between machine learning and deep learning.

	Machine learning	Deep learning
Training dataset	Minimum	Maximum
Choose features	Available	Not Available
Number of algorithms	Several	less
Training time	Short period	Long period

Fig 5.2.1 Table of Difference

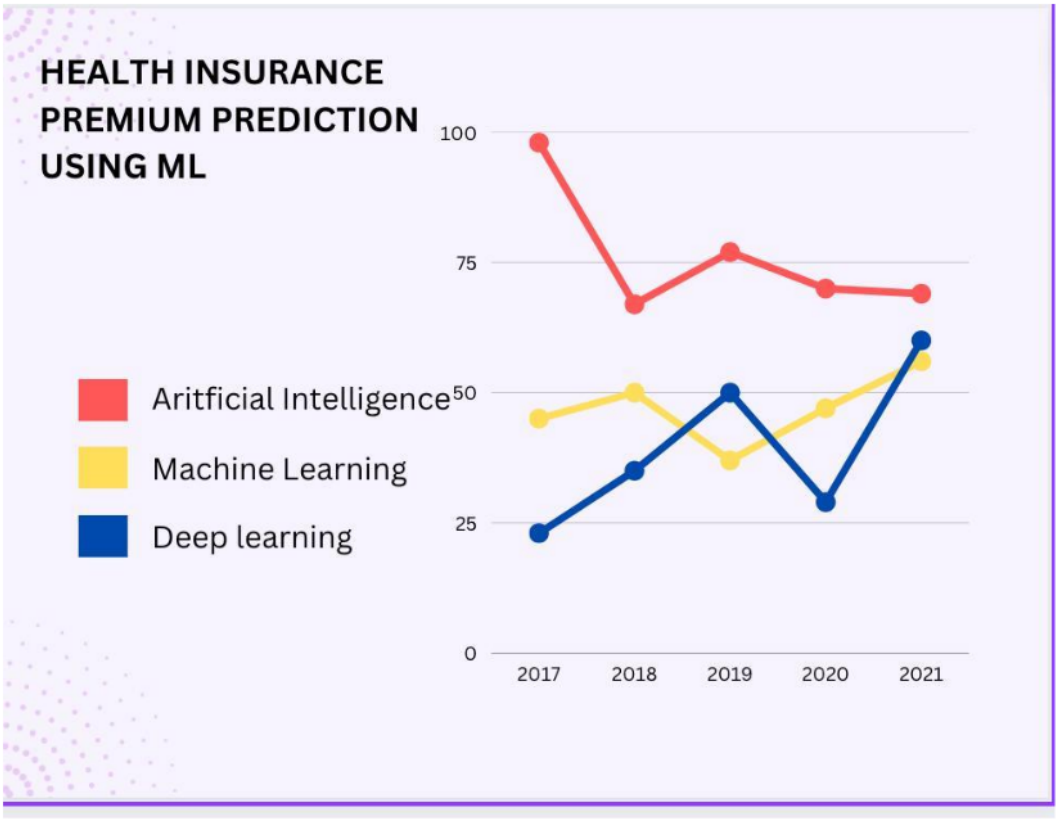


Fig 5.2.2 TensorFlow

5.3 Summary

The performance depends on the comparative study between the earlier existence and currently proposed methodology. The comparison displays the proposed system is effectively processed. It performs as the successor of the existing core project.

CHAPTER 6

CONCLUSION

Thus , through this project , insurance prices are predicted using various Machine learning techniques. Linear Regression algorithm is employed in Order to predict the prices of various insurance policies and medical Insurance policies. Various parameters are considered in order to predict the price of insurance policies like age , gender ,sex and smoking status. Then , finally after making predictions , the accuracy score is calculated and the overall performance of the system is evaluated based on the accuracy score.

CHAPTER 7

FUTURE ENHANCEMENT

In future , the project can be further enhanced by using some other Algorithms like logistic regression instead of linear regression algorithm. Similarly , one or more algorithms can be combined in order to see how well the model predicts and the result is shown in the form of bar graphs by comparing the algorithm's performance.

- In future , the project can be further enhanced by using some other Algorithms like logistic regression instead of linear regression algorithm. Similarly , one or more algorithms can be combined in order to see how well the model predicts and the result is shown in the form of bar graphs by comparing the algorithm's performance.
- The project's security goals will be improved in subsequent development, and the project will be implemented on the blockchain. It is practical to utilise this module to extract the crucial elements required for
- the machine learning model's training. Separated and kept in x and y values, respectively, are the independent and dependent variables.
- The Data is then divided into training data and testing data using the Scikit-learn library's train_test_split function. The machine learning model is trained using all the retrieved features as input.

REFERENCES

[1]Kowshalya, G., & Nandhini, M. (2018, April). Predicting fraudulent claims in automobile insurance.
In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT) (pp. 1338-1343). IEEE.

[2] Kayri, M., Kayri, I., & Gencoglu, M. T. (2017, June). The performance comparison of multiple linear
regression, random forest and artificial neural network by using photovoltaic and atmospheric data. In
2017 14th International Conference on Engineering of Modern Electric Systems (EMES) (pp. 1-4). IEEE.

[3] Denuit, Michel & Hainaut, Donatien & Trufin, Julien. (2019). Effective Statistical Learning Methods
for Actuaries I: GLMs and Extensions. 10.1007/978-3-030-25820-7.

[4] Breiman, Leo. 2001. ―Random Forests.‖ Machine Learning 45 (1). Springer: 5–32.

[5] Chen, T., & Guestrin, C. (2016). XGBoost: a scalable tree boosting system 22nd ACM SIGKDD Int.In Conf. on Knowledge Discovery and Data Mining

APPENDIX

report new

ORIGINALITY REPORT

15%	10%	7%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	library.samdu.uz Internet Source	5%
2	"Intelligent Systems Design and Applications", Springer Science and Business Media LLC, 2020 Publication	3%
3	www.researchgate.net Internet Source	1%
4	jti.aisyahuniversity.ac.id Internet Source	1%
5	Bishal Baniya, Damien Giurco, Scott Kelly. "Green growth in Nepal and Bangladesh: Empirical analysis and future prospects", Energy Policy, 2021 Publication	1%
6	www.coursehero.com Internet Source	1%
7	Submitted to University of North Texas Student Paper	1%

8	Submitted to CSU, San Jose State University Student Paper	1 %
9	pastel.archives-ouvertes.fr Internet Source	1 %

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		

HEALTH INSURANCE PREMIUM PREDICTION USING ML

Dr.K.Danesh, AP
Department of IT
SRMIST Ramapuram
Chennai, India
daneshk@srmist.edu.in

Rahul K
Department of IT
SRMIST Ramapuram
Chennai, India
rk7450@srmist.edu.in

Mohammed Thanish
Department of IT
SRMIST Ramapuram
Chennai, India
mt5840@srmist.edu.in

Abhishek Narayan V
Department of IT
SRMIST Ramapuram
Chennai, India
an5954@srmist.edu.in

***Abstract*—**Everyone is aware of how crucial health insurance is to maintaining human life. It covers a wide range of insurance-related medical costs for a person who is insured and has a number of medical difficulties, including disease, accidents, and surgery. It is known as the amount of business and individual people's inverse the amount to policies based insurance premium. Insurance premiums became the insurance company's source of income once we started making money. In an emergency, we will utilise that health insurance. In this paper, we propose using supervised machine learning, the best type for stochastic gradient descent optimisation algorithm and multilayer perceptron, to remove unwanted data (null values, load the dataset), predict whether or not people will purchase health insurance premiums, and load the dataset. The major goal of this project is to operate in real-time, thus we'll predict the accuracy score and heat map (confusion matrix) for it. The primary goal of this research is to forecast whether consumers will purchase health insurance premium plans using machine learning techniques and various types of algorithms.

Keywords- insurance, premium, dataset, gradient, forecast.

INTRODUCTION

We inhabit a planet that is rife with perils and unknowns. All individuals, homes, businesses, buildings, and property are susceptible to different kinds of danger. and the danger level can change. The potential hazards include loss of life, disease, and property. assets. Health and happiness are the two most significant components of people's life. However, the financial industry has developed a wide range of measures to protect you because risks can't always be avoided. By using people and organizations. Therefore, having insurance is necessary the importance of insurance in people's lives and a policy that lowers or eliminates the costs of losses caused by various risks. It is critical for organizations to comprehend what customers demand. It's crucial for insurance companies to be able to precisely measure or quantify the amount covered by the policy and the required insurance payments when it comes to the value of insurance in people's lives.

I. APPROACHES

The suggested system's primary goal is to develop a smart system that uses multiple Machine learning and Deep learning approaches to anticipate the cost of the insurance policy. The cost of the insurance policy is always fluctuating based on the needs of the consumer. The cost of the insurance policy is predicted using a variety of supervised machine learning methods. An example of a regression problem is predicting the cost of insurance policies. In order to forecast the cost of the insurance policy, a regression approach known as the Linear Regression approach is employed in this research. This algorithm forecasts the statistical association between the dependent and independent variables, and then it makes recommendations. The project's several modules include dataset collecting.

Advantages:

- Accuracy will be compared and analyzed.
- Time consumption is less for classifier algorithms.

II. HEALTH INSURANCE PREMIUM PREDICTION USING ML ARCHITECTURE

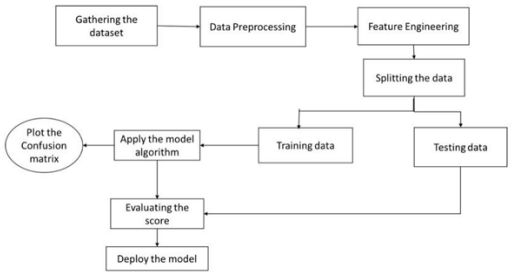


Fig 1: Architecture diagram

- Gathering the dataset and formatting it accordingly based on the data stored in the system.
- Data pre-processing is pre-programming the system for making it suitable for the developer and the user level usage.
- Splitting the data is the data allocation based on its feature, quality and mode of originality. It is of two kinds: Training data and Testing data.
- Evaluating the score is the evaluation of the performance of the system during the user implementation phase.

- Deploying the model is the deployment of the system based on user interface and user requirements.

III. PROPOSED WORK

Dataset Collection:

In this module , the insurance datasets are collected from various datasources . The datasets are stored in the form of a csv file. Data is stored in a tabular format in a csv file . The first step in any machine learning project is that we need to load the datasets. The datasets are loaded by using the pandas object. The quality of the dataset is very important in order to train the machine learning model so that it will predict the insurance prices properly.

Preprocessing:

The next module is the preprocessing module . This module is used for Removing any null value present in a dataset and the values are encoded Using the LabelEncoder. Some algorithms need the data to be scaled in .The range of 0 to 1. The data is scaled by using MinMaxScaler. This Process is required in order to clean the data or remove the unwanted Noise value from the data.

Feature Extraction:

This module is useful for extracting the essential features needed for Training the Machine Learning model. The independent and dependent variables are separated and they are stored in x and y values respectively. Then, using the train_test_split module of the Scikit-learn library , the data is split into training data and testing data. All the extracted features are given as input to train the machine learning model.

Linear Regression Algorithm:

Whenever the predicted output is a continuous variable , then it is called a regression problem. Linear regression algorithm makes predictions based on the statistical relationship between the independent and the dependent variables. Various input features like name , age ,sex and smoking status are given as input to the machine learning algorithm and the algorithm predicts the insurance prices based on the input features.

Evaluation:

After prediction , the final module is the evaluation module . The accuracy score is calculated in order to evaluate the performance of the model. It is a performance score to evaluate how well the model has predicted the insurance prices based on the features given to the machine learning model.

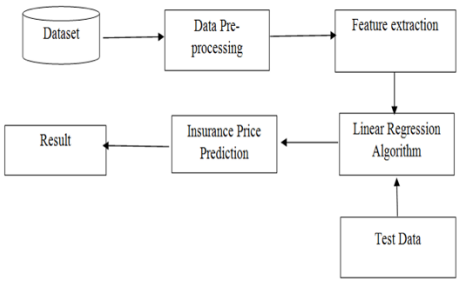


Fig 3.1 Proposed Module

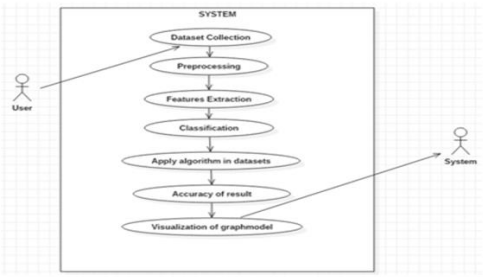


Fig 3.2 UML Diagram

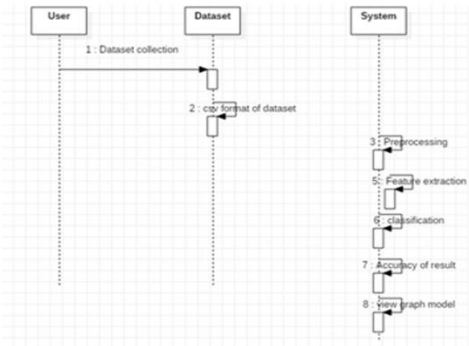


Fig 3.3 Sequence Diagram

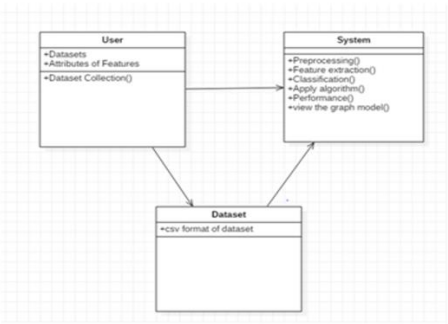


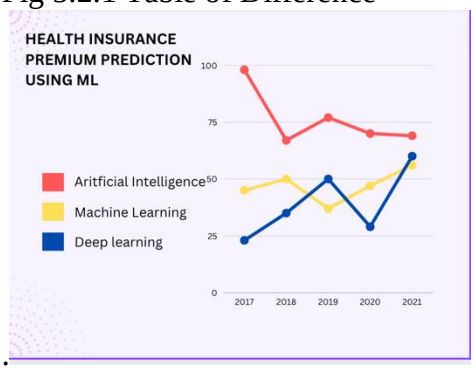
Fig 3.4 Class Diagram

IV. PERFORMANCE ANALYSIS

In the table below, we summarize the difference between machine learning and deep learning.

	Machine learning	Deep learning
Training dataset	Minimum	Maximum
Choose features	Available	Not available
Number of algorithms	Several	Less
Training time	Short period	Long period

Fig 5.2.1 Table of Difference



.

Fig 5.2.2 TensorFlow

The performance depends on the comparative study between the earlier existence and currently proposed methodology. The comparison displays the proposed system is effectively processed. It performs as the successor of the existing core project.

CONCLUSION

Thus , through this project , insurance prices are predicted using various Machine learning techniques. Linear Regression algorithm is employed in Order to predict the prices of various insurance policies and medical Insurance policies. Various parameters are considered in order to predict the price of insurance policies like age , gender ,sex and smoking status. Then , finally after making predictions , the accuracy score is calculated and the overall performance of the system is evaluated based on the accuracy score.

FUTURE ENHANCEMENTS

In future , the project can be further enhanced by using some other Algorithms like logistic regression instead of linear regression algorithm. Similarly , one or more algorithms can be combined in order to see how well the model predicts and the result is shown in the form of bar graphs by comparing the algorithm's performance.

- In future , the project can be further enhanced by using some other Algorithms like logistic regression instead of linear regression algorithm. Similarly , one or more algorithms can be combined in order to see how well the model predicts and the result is shown in the form of bar graphs by comparing the algorithm's performance.
- The project's security goals will be improved in subsequent development, and the project will be implemented on the blockchain. It is practical to utilise this module to extract the crucial elements required for
- the machine learning model's training. Separated and kept in x and y values, respectively, are the independent and dependent variables.

- The Data is then divided into training data and testing data using the Scikit-learn library's train_test_split function. The machine learning model is trained using all the retrieved features as input.

REFERENCES

[1]Kowshalya, G., & Nandhini, M. (2018, April). Predicting fraudulent claims in automobile insurance.
In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT) (pp. 1338-1343). IEEE.
[2] Kayri, M., Kayri, I., & Gencoglu, M. T. (2017, June). The performance comparison of multiple linear
regression, random forest and artificial neural network by using photovoltaic and atmospheric data. In
2017 14th International Conference on Engineering of Modern Electric Systems (EMES) (pp. 1-4). IEEE.
[3] Denuit, Michel & Hainaut, Donatien & Trufin, Julien. (2019). Effective Statistical Learning Methods
for Actuaries I: GLMs and Extensions. 10.1007/978-3-030-25820-7.
[4] Breiman, Leo. 2001. ―Random Forests.‖ Machine Learning 45 (1). Springer: 5–32.
[5] Chen, T., & Guestrin, C. (2016). XGBoost: a scalable tree boosting system 22nd ACM SIGKDD Int.In Conf. on Knowledge Discovery and Data Mining

journal

by Danesh K

Submission date: 12-May-2023 03:40AM (UTC-0400)
Submission ID: 2091160987
File name: journal_paper_HEALTH_INSURANCE_1_-1.pdf (440.94K)
Word count: 1492
Character count: 8178

HEALTH INSURANCE PREMIUM PREDICTION USING ML

J.K.Danesh, AP
Department of IT
SRMIST Ramapuram
Chennai, India
daneshk@srmist.edu.in

Rahul K
Department of IT
SRMIST Ramapuram
Chennai, India
rk7450@srmist.edu.in

Mohammed Thanish
Department of IT
SRMIST Ramapuram
Chennai, India
mt5840@srmist.edu.in

Abhishek Narayan V
Department of IT
SRMIST Ramapuram
Chennai, India
an5954@srmist.edu.in

***Abstract*—**Everyone is aware of how crucial health insurance is to maintaining human life. It covers a wide range of insurance-related medical costs for a person who is insured and has a number of medical difficulties, including disease, accidents, and surgery. It is known as the amount of business and individual people's inverse the amount to policies based insurance premium. Insurance premiums became the insurance company's source of income once we started making money. In an emergency, we will utilise that health insurance. In this paper, we propose using supervised machine learning, the best type for stochastic gradient descent optimisation algorithm and multilayer perceptron, to remove unwanted data (null values, load the dataset), predict whether or not people will purchase health insurance premiums, and load the dataset. The major goal of this project is to operate in real-time, thus we'll predict the accuracy score and heat map (confusion matrix) for it. The primary goal of this research is to forecast whether consumers will purchase health insurance premium plans using machine learning techniques and various types of algorithms.

Keywords- insurance, premium, dataset, gradient, forecast.

INTRODUCTION

We inhabit a planet that is rife with perils and unknowns. All individuals, homes, businesses, buildings, and property are susceptible to different kinds of danger. and the danger level can change. The potential hazards include loss of life, disease, and property. assets. Health and happiness are the two most significant components of people's life. However, the financial industry has developed a wide range of measures to protect you because risks can't always be avoided. By using people and organizations. Therefore, having insurance is necessary the importance of insurance in people's lives and a policy that lowers or eliminates the costs of losses caused by various risks. It is critical for organizations to comprehend what customers demand. It's crucial for insurance companies to be able to precisely measure or quantify the amount covered by the policy and the required insurance payments when it comes to the value of insurance in people's lives.

I. APPROACHES

The suggested system's primary goal is to develop a smart system that uses multiple Machine learning and Deep learning approaches to anticipate the cost of the insurance policy. The cost of the insurance policy is always fluctuating based on the needs of the consumer. The cost of the insurance policy is predicted using a variety of supervised machine learning methods. An example of a regression problem is predicting the cost of insurance policies. In order to forecast the cost of the insurance policy, a regression approach known as the Linear Regression approach is employed in this research. This algorithm forecasts the statistical association between the dependent and independent variables, and then it makes recommendations. The project's several modules include dataset collecting.

Advantages:

- Accuracy will be compared and analyzed.
- Time consumption is less for classifier algorithms.

II. HEALTH INSURANCE PREMIUM PREDICTION USING ML ARCHITECTURE

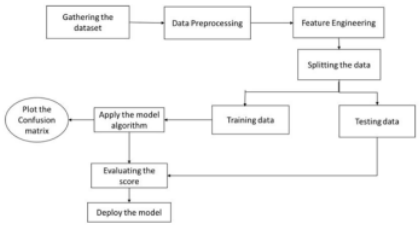


Fig 1: Architecture diagram

- Gathering the dataset and formatting it accordingly based on the data stored in the system.
- Data pre-processing is pre-programming the system for making it suitable for the developer and the user level usage.
- Splitting the data is the data allocation based on its feature, quality and mode of originality. It is of two kinds: Training data and Testing data.
- Evaluating the score is the evaluation of the performance of the system during the user implementation phase.

- Deploying the model is the deployment of the system based on user interface and user requirements.

III. PROPOSED WORK

Dataset Collection:

In this module , the insurance datasets are collected from various datasources . The datasets are stored in the form of a csv file. Data is stored in a tabular format in a csv file . The first step in any machine learning project is that we need to load the datasets. The datasets are loaded by using the pandas object. The quality of the dataset is very important in order to train the machine learning model so that it will predict the insurance prices properly.

Preprocessing:

The next module is the preprocessing module . This module is used for Removing any null value present in a dataset and the values are encoded Using the LabelEncoder. Some algorithms need the data to be scaled in .The range of 0 to 1. The data is scaled by using MinMaxScaler. This Process is required in order to clean the data or remove the unwanted Noise value from the data.

Feature Extraction:

This module is useful for extracting the essential features needed for Training the Machine Learning model. The independent and dependent variables are separated and they are stored in x and y values respectively. Then, using the train_test_split module of the Scikit-learn library , the data is split into training data and testing data. All the extracted features are given as input to train the machine learning model.

Linear Regression Algorithm:

Whenever the predicted output is a continuous variable , then it is called a regression problem. Linear regression algorithm makes predictions based on the statistical relationship between the independent and the dependent variables. Various input features like name , age ,sex and smoking status are given as input to the machine learning algorithm and the algorithm predicts the insurance prices based on the input features.

Evaluation:

After prediction , the final module is the evaluation module . The accuracy score is calculated in order to evaluate the performance of the model. It is a performance score to evaluate how well the model has predicted the insurance prices based on the features given to the machine learning model.

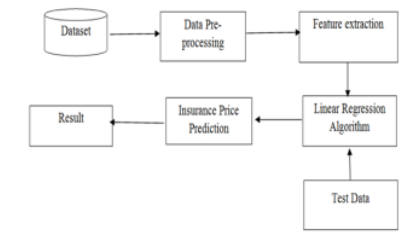


Fig 3.1 Proposed Module

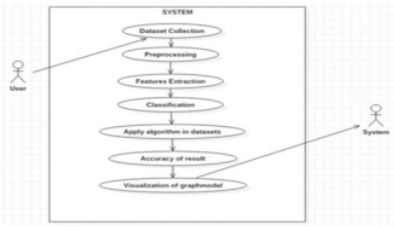


Fig 3.2 UML Diagram

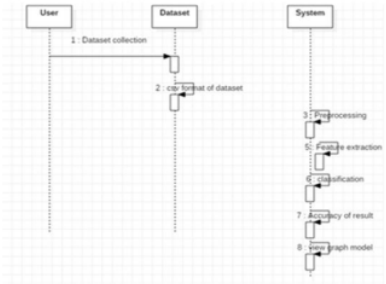


Fig 3.3 Sequence Diagram

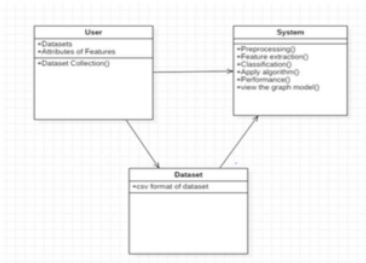


Fig 3.4 Class Diagram

IV. PERFORMANCE ANALYSIS

In the table below, we summarize the difference between machine learning and deep learning.

	Machine learning	Deep learning
Training dataset	Minimum	Maximum
Choose features	Available	Not available
Number of algorithms	Several	Less
Training time	Short period	Long period

Fig 5.2.1 Table of Difference

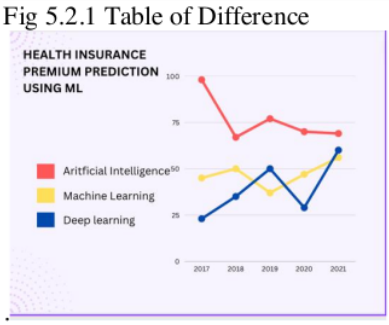


Fig 5.2.2 TensorFlow

The performance depends on the comparative study between the earlier existence and currently proposed methodology. The comparison displays the proposed system is effectively processed. It performs as the successor of the existing core project.

CONCLUSION

Thus , through this project , insurance prices are predicted using various Machine learning techniques. Linear Regression algorithm is employed in Order to predict the prices of various insurance policies and medical Insurance policies. Various parameters are considered in order to predict the price of insurance policies like age , gender ,sex and smoking status. Then , finally after making predictions , the accuracy score is calculated and the overall performance of the system is evaluated based on the accuracy score.

FUTURE ENHANCEMENTS

In future , the project can be further enhanced by using some other Algorithms like logistic regression instead of linear regression algorithm. Similarly , one or more algorithms can be combined in order to see how well the model predicts and the result is shown in the form of bar graphs by comparing the algorithm's performance.

- In future , the project can be further enhanced by using some other Algorithms like logistic regression instead of linear regression algorithm. Similarly , one or more algorithms can be combined in order to see how well the model predicts and the result is shown in the form of bar graphs by comparing the algorithm's performance.
- The project's security goals will be improved in subsequent development, and the project will be implemented on the blockchain. It is practical to utilise this module to extract the crucial elements required for
- the machine learning model's training. Separated and kept in x and y values, respectively, are the independent and dependent variables.

- The Data is then divided into training data and testing data using the Scikit-learn library's train_test_split function. The machine learning model is trained using all the retrieved features as input.

REFERENCES

[1]Kowshalya, G., & Nandhini, M. (2018, April). Predicting fraudulent claims in automobile insurance.
In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT) (pp. 1338-1343). IEEE.
[2] Kayri, M., Kayri, I., & Gencoglu, M. T. (2017, June). The performance comparison of multiple linear
regression, random forest and artificial neural network by using photovoltaic and atmospheric data. In
2017 14th International Conference on Engineering of Modern Electric Systems (EMES) (pp. 1-4). IEEE.
[3] Denuit, Michel & Hainaut, Donatien & Trufin, Julien. (2019). Effective Statistical Learning Methods
for Actuaries I: GLMs and Extensions. 10.1007/978-3-030-25820-7.
[4] Breiman, Leo. 2001. ―Random Forests.‖ Machine Learning 45 (1). Springer: 5–32.
[5] Chen, T., & Guestrin, C. (2016). XGBoost: a scalable tree boosting system 22nd ACM SIGKDD Int.In Conf. on Knowledge Discovery and Data Mining

journal

ORIGINALITY REPORT

13%	5%	10%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	"Intelligent Systems Design and Applications", Springer Science and Business Media LLC, 2020 Publication	3%
2	www.researchgate.net Internet Source	3%
3	S. Poongothai, S. Leoni Sharmila. "Chapter 11 A Comparison of Fuzzy and ACO-Based Fuzzy for Classification of Bio-Medical Database", Springer Science and Business Media LLC, 2022 Publication	3%
4	Submitted to Ngee Ann Polytechnic Student Paper	1%
5	Xinyi Tian, Jun Wang, Shuang Yuan, Jie Ji, Wei Ke, Chuyao Wang. "Investigation on the electrical performance of a curved PV roof integrated with CIGS cells for traditional Chinese houses", Energy, 2023 Publication	1%

jti.aisyahuniversity.ac.id

6	Internet Source	1 %
7	www.ncbi.nlm.nih.gov Internet Source	1 %
8	www.researchsquare.com Internet Source	1 %
9	www.scribd.com Internet Source	1 %

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		

NCCS '23 submission 25

1 message

NCCS '23 <nccs23@easychair.org> Fri, 12 May 2023 at 10:35 am
To: Mohamed Thanish <mt5840@srmist.edu.in>

Dear authors,

We received your submission to NCCS '23 (National conference on Cyber Security):

Authors : Dr.K Danesh, Rahul Kumar and Mohamed Thanish
Title : HEALTH INSURANCE PREMIUM PREDICTION USING ML
Number : 25

The submission was uploaded by Mohamed Thanish <mt5840@srmist.edu.in>. You can access it via the NCCS '23 EasyChair Web page

https://easychair.org/conferences/?conf=nccs23

Thank you for submitting to NCCS '23.

Best regards,
EasyChair for NCCS '23.