**AI/ML and Data Science**

**Insurance Claim Data Set**

Summer Internship Report Submitted in partial fulfillment of the requirement for undergraduate degree of In **Bachelor of Technology**

**COMPUTER SCIENCE AND ENGINEERING**

By

**N.Sai Pramod**

**222010311021**

Under the Guidance of

**Mr. M. Yogi Reddy**

Assistant Professor



Department Of Computer Science and Engineering

GITAM School of Technology

GITAM (Deemed to be University)

Hyderabad-502329

September 2023

**DECLARATION**

I submit this industrial training work entitled **“Insurance Claim Data Set”** to GITAM (Deemed to be University), Hyderabad in partial fulfillment of the requirements for the award of the degree of “**Bachelor of Technology**” in “**Computer Science and Engineering**”. I declare that it was carried out independently by me under the guidance of **Mr.M.Yogi Reddy** Asst. Professor, GITAM (Deemed to be University), Hyderabad, India.

The results embodied in this report have not been submitted to any other University or Institute for the award of any degree or diploma.

Place: HYDERABAD Name: N.Sai Pramod

Date: 08-09-2023 Student Roll No:222010311021

GITAM (DEEMED TO BE UNIVERSITY) 

Hyderabad-502329, India

Dated: 08 -09-2023

**CERTIFICATE**

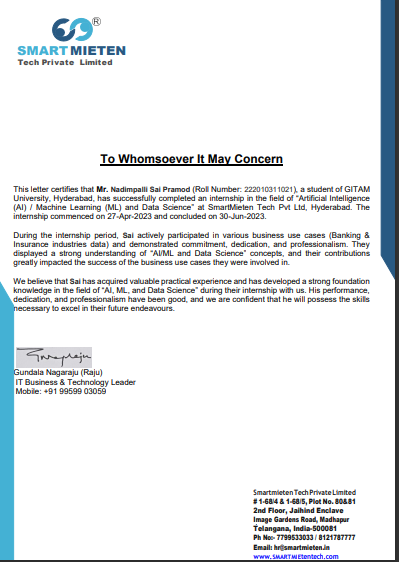
This is to certify that the Industrial Training Report entitled **“Insurance Claim Data Set”** is being submitted by N.Sai Pramod(222010311021) in partial fulfilment of the requirement for the award of **Bachelor of Technology in Computer Science and Engineering** at GITAM (Deemed to Be University), Hyderabad during the academic year 2023-2024.

It is faithful record work carried out by him at the **Computer Science and Engineering Department**, GITAM University Hyderabad Campus under my guidance and supervision.

### Mr.M.Yogi Reddy Dr. K.S Sudeep

Assistant Professor Professor and HOD

Department of CSE Department of CSE



**ACKNOWLEDGEMENT**

Apart from my effort, the success of this internship largely depends on the encouragement and guidance of many others. I take this opportunity to express my gratitude to the people who have helped me in the successful competition of this internship.

I would like to thank respected **Prof. D. Sambasiva Rao**, Pro Vice Chancellor, GITAM Hyderabad and **Prof. N. Seetharamaiah**, Principal, GITAM Hyderabad.

I would like to thank respected **Dr. K S Sudeep,** Head of the Computer Science and Engineering for giving me such a wonderful opportunity to expand my knowledge for my own branch and giving me guidelines to present an internship report. It helped me a lot to realize of what we study for.

I would like to thank the respected faculties **Mr.M.Yogi Reddy** who helped me to make this internship a successful accomplishment.

I would also like to thank my friends who helped me to make my work more organized and well-stacked till the end.

N. Sai Pramod

222010311021

**ABSTRACT**

"In response to the rising tide of fraudulent insurance claims, this project focuses on the development of a robust machine learning-based predictive model. The objective is to effectively identify and flag potential fraudulent insurance claims, mitigating financial losses and bolstering trust within the insurance industry. The envisioned model promises to streamline claim processing, enhance risk management, and ultimately deliver substantial cost savings for insurance companies. By harnessing advanced algorithms, this project seeks to provide a proactive solution to combat insurance fraud, thereby improving the overall efficiency and integrity of the insurance ecosystem."

Table Of Contents

1.0 Introduction ….............................................................................. 8

1.1 Different Types Of ML …............................................................... 8

1.2 Benefits of ML …........................................................................... 9

1.3 About Industry …...........................................................................10

1.3.1 AI/ML Role in Insurance claim Prediction................................. 11

2.0 Insurance Prediction …................................................................ 11

2.1 Main Insurance for AI Quote Analysis …...................................... 12

2.2 Internship Project – Link ….......................................................... 13

3.0 AI/ML Results and Results …....................................................... 13

3.1 Your Problem Of Statement ….......................................................13

3.2 Data Science Project Life Cycle …............................................... 13

3.2.1 Data Exploratory Analysis …..................................................... 14

3.2.2 Data Preprocessing …............................................................... 14

3.2.2.1 Checking for Duplicates and Low Variation Data ….............14

3.2.2.2 Label Encoding …...................................................................15

3.2.2.3 Imputer Techniques …............................................................15

3.2.2.4 Identifying Target Variable …................................................ 15

3.2.3 Splitting Into Train & Test Dataset …...................................... 16

3.2.4 Feature Scaling …......................................................................16

4.0[: MODEL BUILDING AND EVALUATION](https://docs.google.com/document/d/1ruBVdyx_9yiMPSQ8EImC_7vizmUo-7jN957qiLGz8jI/edit?pli=1#heading=h.ihv636).........................................17

4[.1. Building a Random Forest Classifier Model...............................](https://docs.google.com/document/d/1ruBVdyx_9yiMPSQ8EImC_7vizmUo-7jN957qiLGz8jI/edit?pli=1#heading=h.32hioqz) **17**

4[.3. Comparision with other Algorithms](https://docs.google.com/document/d/1ruBVdyx_9yiMPSQ8EImC_7vizmUo-7jN957qiLGz8jI/edit?pli=1#heading=h.3fwokq0)........................ ....................25

5.0 Analysis..........................................................................................34

6.0 Conclusions and Future Work …...................................................35

7.0 References …...............................................................................35

1. **Introduction:**

With the increasing power of computer technology, companies and institutions can nowadays store large amounts of data at reduced cost. The amount of available data is increasing exponentially and cheap disk storage makes it easy to store data that previously was thrown away. There is a huge amount of information locked up in databases that is potentially important but has not yet been explored. The growing size and complexity of the databases makes it hard to analyse the data manually, so it is important to have automated systems to support the process. Hence there is the need of computational tools able to treat these large amounts of data and extract valuable information.

In this context, Data Mining provides automated systems capable of processing large amounts of data that are already present in databases. Data Mining is used to automatically extract important patterns and trends from databases seeking regularities or patterns that can reveal the structure of the data and answer business problems. Data Mining includes learning techniques that fall into the field of Machine learning. The growth of databases in recent years brings data mining at the forefront of new business technologies.

A key challenge for the insurance industry is to charge each customer an appropriate price for the risk they represent. Risk varies widely from customer to customer and a deep understanding of different risk factors helps predict the likelihood and cost of insurance claims. The goal of this program is to see how well various statistical methods perform in predicting auto Insurance claims based on the characteristics of the driver, vehicle and driver / vehicle coverage details.

A number of factors will determine BI claims prediction among them a driver's age, past accident history, and domicile, etc. However, this contest focused on the relationship between claims and vehicle characteristics well as other characteristics associated with the auto insurance policies.

**1.1 Different Types Of Machine Learning:**

Machine learning can be broadly categorized into three main types: supervised learning, unsupervised learning, and reinforcement learning.

* **Supervised Learning**: In supervised learning, the algorithm is trained on labeled data, where the input data is paired with corresponding target labels or outcomes. The algorithm learns from this labeled data to make predictions or classify new, unseen data. Common algorithms used in supervised learning include decision trees, support vector machines (SVM), and neural networks.
* **Unsupervised Learning**: Unsupervised learning involves training algorithms on unlabeled data, where there are no predetermined target labels or outcomes. The algorithms learn patterns, relationships, and structures within the data. Unsupervised learning can be used for tasks such as clustering similar data points, dimensionality reduction, and anomaly detection. Clustering algorithms like k-means and hierarchical clustering, as well as dimensionality reduction techniques like principal component analysis (PCA), are examples of unsupervised learning.
* **Reinforcement Learning**: Reinforcement learning is an area of machine learning where an agent learns to make decisions in an interactive environment to maximize a reward signal. Through trial and error, the agent learns to optimize its actions to achieve the maximum cumulative reward.

These three types of machine learning provide a framework for training algorithms to learn from data and make predictions or decisions based on the patterns and relationships they discover.

**1.2 Benefits of Using Machine Learning in Insurance claim data set:**

The utilization of machine learning in safe driver prediction offers several significant benefits. Here are some key advantages:

* Improved Accuracy: Machine learning algorithms can process large volumes of data and identify complex patterns that may not be apparent through traditional statistical analysis. By considering a wide range of driver attributes and historical data, machine learning models can make more accurate predictions of claim initiation probabilities. This enhanced accuracy enables insurance companies to better assess risk and make informed decisions regarding premiums, coverage, and risk mitigation strategies.
* Personalized Risk Assessment: Machine learning models can take into account individual driver characteristics and driving behavior to provide personalized risk assessments. This enables insurance companies to offer tailored insurance policies and pricing based on an individual's specific risk profile. Policyholders with lower predicted claim probabilities can be rewarded with lower premiums, while high-risk drivers can be charged higher rates, fostering fairness and encouraging safer driving habits.
* Fraud Detection: Machine learning algorithms can be trained to identify patterns indicative of fraudulent insurance claims. By analyzing historical data and identifying anomalous behaviors or suspicious patterns, these models can help insurance companies flag potentially fraudulent claims for further investigation. This helps reduce fraudulent activities, leading to cost savings for insurers and fairer premiums for genuine policyholders.

**1.3 About Industry (Insurance Claim Data Set):**

About the Insurance Industry:

The insurance industry, a cornerstone of modern financial security, operates as a guardian against life's unpredictable twists and turns. It is an intricate ecosystem that encompasses a wide array of policies, covering everything from health and property to automobiles and more. However, as the global landscape evolves, the industry finds itself at a crossroads, grappling with a formidable challenge—the relentless surge in the number of insurance claims being lodged each day.

This surge, while emblematic of the industry's mission to provide protection, is also marked by a darker facet—fraudulent insurance claims. Fraudulent claims represent a growing concern that reverberates throughout the industry, posing a significant threat not only in terms of financial losses but also in eroding the bedrock of trust between insurers and policyholders.

The insurance industry stands as a complex web of transactions, financial strategies, and risk assessment. It plays a vital role in shielding individuals and businesses from the financial repercussions of unforeseen events. Yet, the rising tide of fraudulent claims casts a shadow over the industry's noble mission, necessitating innovative solutions to safeguard its integrity.

In response to this challenge, the industry is increasingly turning to advanced technology, specifically machine learning algorithms, to develop predictive models capable of accurately identifying and flagging potential fraudulent insurance claims. This innovative approach represents a watershed moment for the industry, with the potential to revolutionize the way it operates.

The implications of successfully deploying such models are profound. Beyond the immediate financial benefits of reducing fraudulent claims, these models promise to streamline the often cumbersome claim processing workflows within insurance companies. They enable insurers to promptly separate legitimate claims from fraudulent ones, facilitating the efficient settlement of valid claims and ultimately enhancing the overall customer experience.

Furthermore, these models have the potential to reshape the industry's risk management strategies. By pinpointing fraudulent claims with precision, insurers can fine-tune their risk assessments, allocate resources more strategically, and secure their long-term sustainability.

In essence, the insurance industry finds itself at an inflection point, where the fusion of data science and insurance expertise holds the key to overcoming the challenge of insurance fraud. This endeavor signifies the industry's commitment to fortify itself against fraudulent practices, restore trust, and unlock the full potential of a future where technology empowers it to serve its policyholders with unparalleled efficiency and integrity.

**1.3.1 Role Of AI/ML in Insurance Claim Data Set :**

The Role of AI/ML in Insurance Claim Data:

Artificial Intelligence (AI) and Machine Learning (ML) are poised to revolutionize the insurance industry, particularly in the domain of insurance claim data. Here's an exploration of their pivotal roles:

1. Fraud Detection: AI/ML algorithms excel at identifying patterns and anomalies. In insurance, they can analyze vast datasets to uncover unusual behavior, making them highly effective in detecting fraudulent claims. These algorithms can learn from historical fraud cases, continuously adapting to evolving fraudulent tactics, thereby significantly reducing financial losses for insurance companies.

2. Risk Assessment: AI/ML models can assess risk more accurately by considering a multitude of factors. They can analyze customer data, claims history, and external variables (e.g., weather, economic trends) to fine-tune risk profiles. This leads to more precise pricing of policies, ultimately benefiting both insurers and policyholders.

3. Claim Processing Automation: AI-powered chatbots and natural language processing (NLP) algorithms can handle routine claim inquiries and assist policyholders in filing claims. This not only expedites the process but also reduces administrative costs for insurers.

4. Customer Experience Enhancement: AI-driven personalization enables insurers to offer tailored policies and services. Through analyzing customer data, AI can recommend relevant insurance products, thereby improving customer satisfaction and retention.

5. Predictive Underwriting: ML models can predict customer behavior and identify potential risks. By analyzing data on an individual's lifestyle, health, and habits, insurers can assess risks more accurately and offer customized coverage.

In conclusion, AI and ML are poised to transform the insurance industry by enhancing fraud detection, risk assessment, customer experience, and operational efficiency. These technologies empower insurers to better serve their policyholders while reducing financial losses and strengthening trust within the industry.

**2.0 Insurance Claim Data Set Prediction:**

The term "insurance claim dataset prediction" typically refers to using machine learning or statistical modeling techniques to predict the likelihood of an insurance claim being filed or to estimate the potential cost of an insurance claim. Insurance companies collect vast amounts of data related to policyholders, their characteristics, and historical claim information. This data can be used to make predictions and inform business decisions. Here's how it works:

**2.1 Factors Influencing:**

In insurance claim datasets, various factors can influence the likelihood and cost of insurance claims. These factors are crucial for insurance companies to understand and analyze to make accurate predictions and informed decisions. Here are some of the key factors that can influence insurance claims:

1. \*\*Policyholder Characteristics:\*\*

- Age: Younger or older policyholders may have different risk profiles.

- Gender: Some insurance types may be gender-specific in terms of risk.

- Marital Status: Married individuals may have different risk profiles than singles.

- Occupation: Certain occupations may have higher risk levels.

- Location: Geographic factors, such as urban vs. rural areas, can impact risk.

2. \*\*Policy Details:\*\*

- Coverage Type: Different types of coverage (e.g., comprehensive, liability) affect the nature of claims.

- Policy Limits: The maximum amount the policy will pay out affects claim amounts.

- Deductibles: Higher deductibles may discourage small claims.

3. \*\*Claim History:\*\*

- Previous Claims: A history of previous claims can indicate future claims.

- Claim Frequency: How often a policyholder files claims.

- Claim Severity: The average cost of claims filed.

4. Vehicle or Property Information:

- Vehicle Make and Model: In auto insurance, the type of vehicle can impact risk.

Property Characteristics: In property insurance, details about the property's age and condition matter.

5. Driving Record (for Auto Insurance):

- Accident History: Past accidents can predict future accidents.

- Traffic Violations: Traffic violations may indicate risky driving behavior.

6. Health Information (for Health Insurance):

- Preexisting Conditions: Health conditions can impact medical claims.

- Lifestyle Factors: Habits such as smoking or exercise can influence health risks.

7. Environmental Factors:

- Weather: Climate and weather conditions can affect claims (e.g., storms, floods).

- Natural Disasters: Areas prone to earthquakes, hurricanes, etc., face higher risks.

.

**2.2 Internship Project – Data Link**

<https://www.kaggle.com/datasets/shashwatwork/insurance-claim-dataset>

**2.0AI/ML Modelling and Results:**

**3.1 Problem Statement:**

With the increasing number of insurance claims being filed daily, insurance companies face significant challenges in identifying fraudulent claims. Fraudulent claims lead to financial losses and can undermine the trust between insurers and policyholders. To address this issue, we aim to develop a predictive model using machine learning algorithms that can accurately detect and flag potential fraudulent insurance claims.

By successfully developing an accurate and efficient fraud detection model, insurance companies can significantly reduce financial losses due to fraudulent claims, streamline their claim processing workflows, and enhance their overall risk management strategy.

**3.2 Data Science Lifecycle:**

Data Science is an interdisciplinary domain that merges programming abilities, subject matter expertise, and statistical and mathematical knowledge to derive valuable insights and information from data.

**3.2.1 Data Exploration Analysis:**

This process encompasses condensing the key attributes of the data, recognizing patterns, and identifying any anomalies or gaps. These actions enable data analysts to acquire a deeper understanding of the data, identify possible connections or associations, and make knowledgeable choices regarding the next stages of analysis.

* Target variable: "target" represents insurance claim data set (1) or non-initiation (0).
* Gender: "Gender" column indicates driver gender .
* Numericalfeatures:HCPCS\_CD\_1','HCPCS\_CD\_2','HCPCS\_CD\_3','HCPCS\_CD\_4','HCPCS\_CD\_5','HCPCS\_CD\_6','HCPCS\_CD\_7','HCPCS\_CD\_8','HCPCS\_CD\_9','HCPCS\_CD\_10','HCPCS\_CD\_11','HCPCS\_CD\_12','HCPCS\_CD\_13','HCPCS\_CD\_14','HCPCS\_CD\_15','HCPCS\_CD\_16','HCPCS\_CD\_17','HCPCS\_CD\_18','HCPCS\_CD\_19','HCPCS\_CD\_20','HCPCS\_CD\_21','HCPCS\_CD\_22','HCPCS\_CD\_23','HCPCS\_CD\_24','HCPCS\_CD\_25','HCPCS\_CD\_26','HCPCS\_CD\_27','HCPCS\_CD\_28','HCPCS\_CD\_29','HCPCS\_CD\_30','HCPCS\_CD\_31','HCPCS\_CD\_32','HCPCS\_CD\_33','HCPCS\_CD\_34','HCPCS\_CD\_35','HCPCS\_CD\_36','HCPCS\_CD\_37','HCPCS\_CD\_38','HCPCS\_CD\_39','HCPCS\_CD\_40','HCPCS\_CD\_41','HCPCS\_CD\_42','HCPCS\_CD\_43','HCPCS\_CD\_44','HCPCS\_CD\_45','OP\_PHYSN\_NPI','OT\_PHYSN\_NPI','ICD9\_DGNS\_CD\_8','ICD9\_DGNS\_CD\_9','ICD9\_DGNS\_CD\_10','ICD9\_PRCDR\_CD\_1','ICD9\_PRCDR\_CD\_2','ICD9\_PRCDR\_CD\_3','ICD9\_PRCDR\_CD\_4','ICD9\_PRCDR\_CD\_5','ICD9\_PRCDR\_CD\_6

**3.2.2 Data Preprocessing:**

Variables that do not have an impact on the target variable have been excluded to minimize noise and computation time. While it is crucial to identify outliers and anomalous data points, this particular dataset does not contain any such outliers or anomalies, therefore making this step unnecessary.

Fig:3.2.2









**3.2.2.1 Checking for Duplicates and Low Varying Data:**

Here are some reasons why it is important for datasets to be free of duplicate values:

* Inflated Performance Metrics: Duplicate values can artificially boost performance metrics, such as accuracy or precision, making the evaluation of models or algorithms unreliable.
* Redundant Information: Duplicate values provide redundant information, which can skew data analysis and lead to biased results.
* Data Integrity: Duplicate values can compromise the integrity of the dataset, as they may indicate data entry errors or inconsistencies.
* Efficient Storage and Processing: Removing duplicate values allows for more efficient storage and processing of data, reducing storage space requirements and improving computational efficiency.

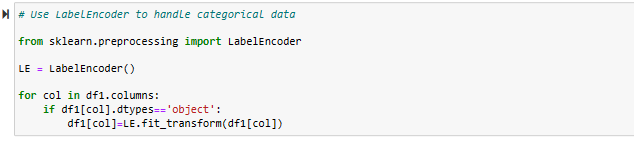


* Therefore, it is recommended to check for duplicates in the dataset based on the above reasons. Since no duplicate values were found upon checking, we can proceed further.

**3.2.2.2 Label Encoding:**

A label encoder is a preprocessing technique used to convert categorical variables into numerical values. Here's a brief description of a label encoder in four lines:

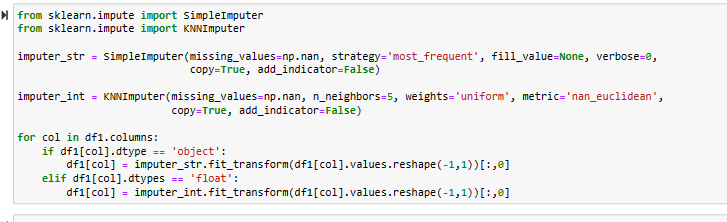
* It assigns a unique numerical value to each unique category in the variable.
* The encoded labels are integers ranging from 0 to (n\_classes - 1), where n\_classes is the number of unique categories.
* Label Encoder is commonly used for encoding target variables in machine learning tasks.



**3.2.2.3 Imputer Techniques:**

KNN Imputer is a technique used to fill in missing values in a dataset using the K-Nearest Neighbors algorithm. In this stage, we are finding the missing values for Driver\_Age using KNN imputing techniques.

**Fig:3.2.2.3**



**3.2.2.4 Identifying Target Variable:**

The target variable in the provided context is not explicitly mentioned. However, based on the given information, the main objective is to build a predictive model that can estimate the probability of a driver initiating an auto insurance claim in the next year. The goal is to develop a model that can help insurance companies assess the risk associated with individual drivers and price their insurance policies accordingly. Here, from the given dataset the target variable is ‘Target’.

**Fig:3.2.2.4**

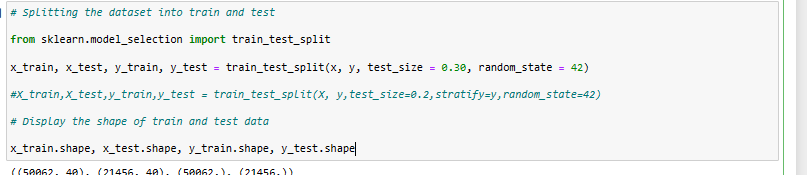


**3.2.3 Splitting the Dataset into Train and Test:**

The main reason for splitting a dataset into training and test sets is to evaluate the performance of a machine learning model on unseen data. Some, of the reason to split the dataset are because of: Model Evaluation and Preventing Overfitting.

In this case, the chosen test size is 0.3 with random state = 42.

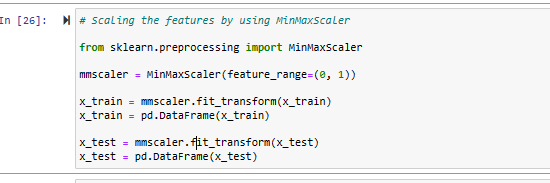
Fig:3.2.3



**3.2.4 Feature Scaling:**

Feature scaling is a common practice used to ensure that all features in a dataset have comparable magnitudes and units. This is beneficial for various data analysis and machine learning tasks. Before, performing Feature Scaling it is important to remove columns which are not affecting target variable and also target variable.

Fig:3.2.4



# 4.0 MODEL BUILDING AND EVALUATION

##### 4.1 BUILDING A RANDOM FOREST CLASSIFIER MODEL:

**Random Forest Classifier:**

Random Forest Classifier is a machine learning algorithm that belongs to the ensemble learning family. It is widely used for both classification and regression tasks. The "forest" in the name is a collection of decision trees, where each tree is trained on a different subset of the data with bootstrapping (random sampling with replacement) and can vote on the final classification result (in the case of classification tasks) or average the predictions (in the case of regression tasks). Random Forests are known for their versatility and robustness, and they are considered one of the most powerful and popular machine learning algorithms.

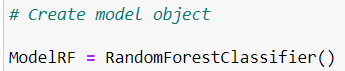
Here we need to build a Random Forest classification model, evaluate its performance using various metrics, and visualize its performance with an ROC curve. The metrics and visualizations help assess how well the model is doing in classifying data points, especially in binary classification problems. We achieve this by doing the following steps:

1. Import Necessary Libraries:



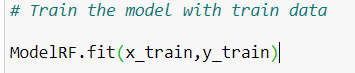
**Importing libraries**

1. Create a Random Forest Model: An instance of the Random Forest classifier is created. You can also see commented code with parameters that you can configure for the Random Forest classifier.



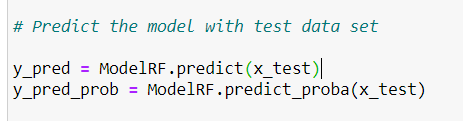
**Creating model**

1. Train the Model: The Random Forest model is trained using the training data (x\_train and y\_train).



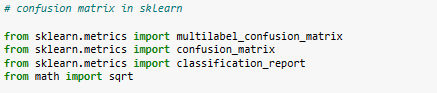
**Model Training**

1. Make Predictions: The model is used to make predictions on the test data (x\_test).



**Predicting the model**

1. Calculate Confusion Matrix: A confusion matrix is calculated to evaluate the model's performance. It shows the counts of true positives, true negatives, false positives, and false negatives.



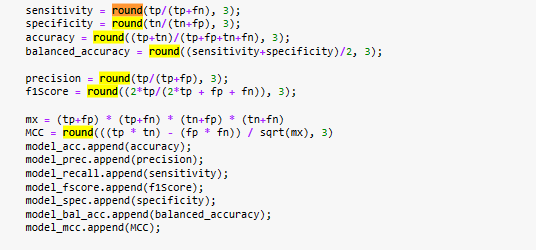
**Confusion Matrix**

1. Display Metrics and Scores:

**. Model Name**

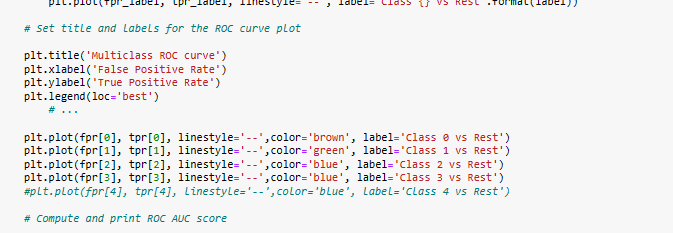
* **True\_Positive**
* **False\_Negative**
* **False\_Positive**
* **True\_Negative**
* **Accuracy**
* **Precision**
* **Recall**
* **F1 Score**
* **Specificity**
* **MCC**
* **ROC\_AUC\_Score**
* **ROC\_AUC\_Score**

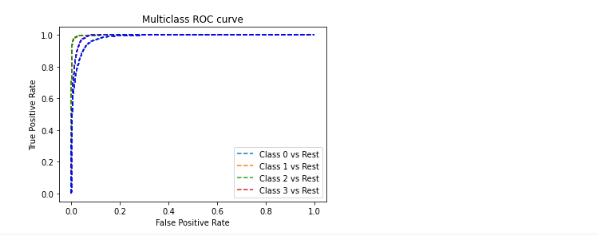
Fig:4.1



**Displaying Metrics**

1. ROC Curve Plotting: The code calculates and plots the Receiver Operating Characteristic (ROC) curve for the model's performance.

Fig:4.1 



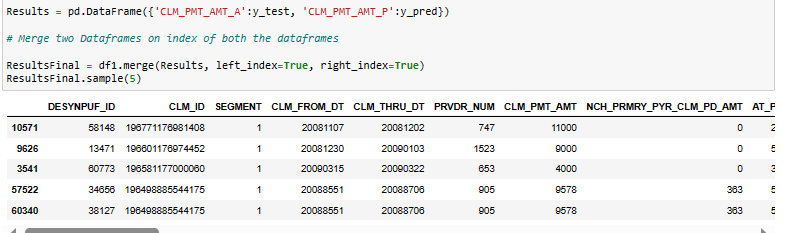
**ROC Curve**

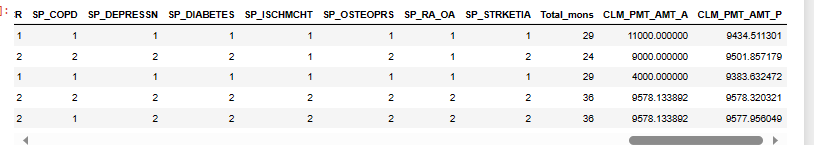
##### 4.2 RESULTS:

##### 

We store the results in two different sets of Actual results and Predicted results. Then we merge them and compare with each other.

Fig:4.1





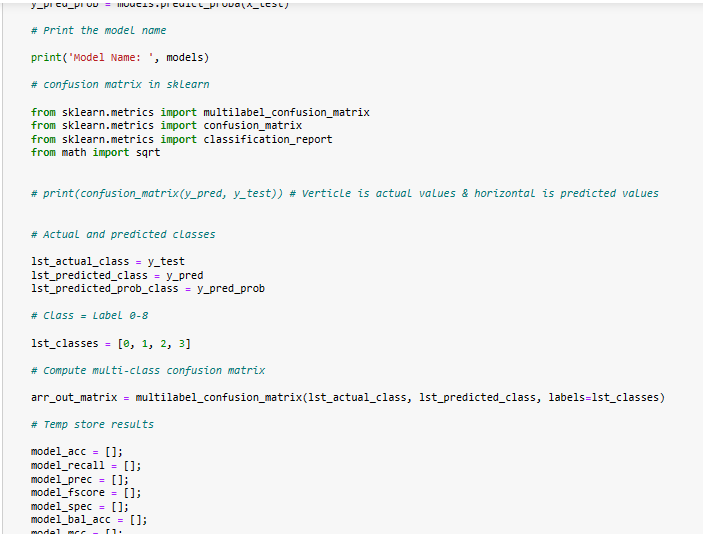
**Results**

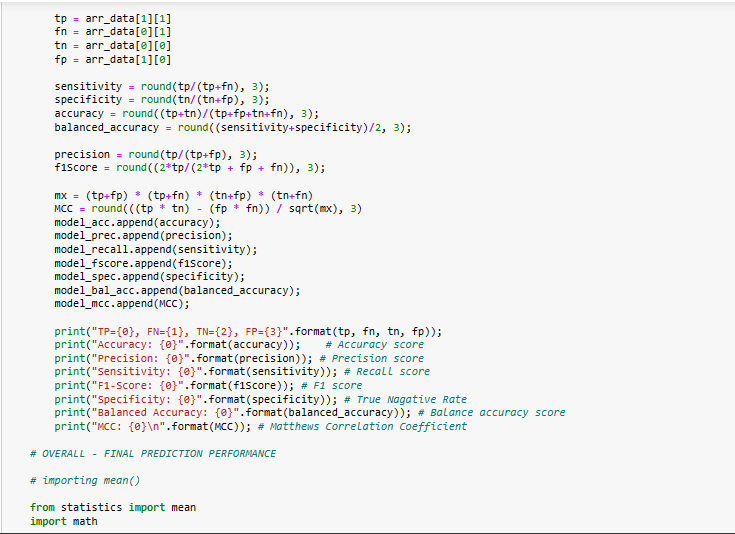
##### 4.3 COMPARISION WITH OTHER ALGORITHMS:

We now compare the results of this model with different other models.

Fig:4.3

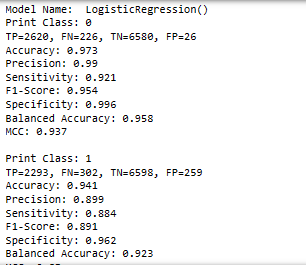


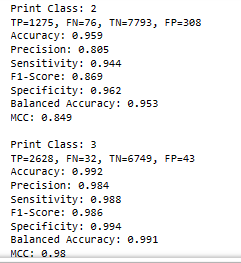




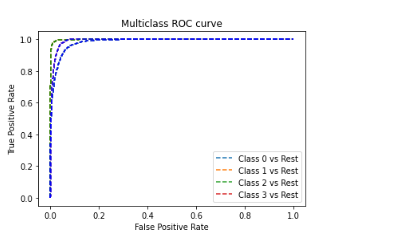


1. **Results of Logistic Regression:**

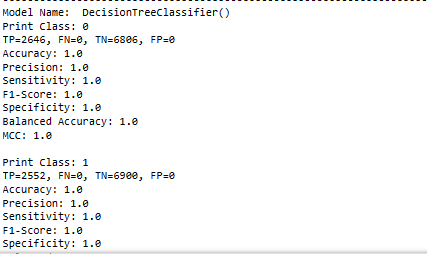
****

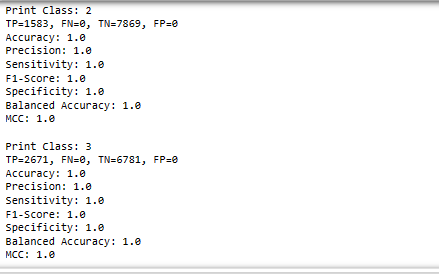
****

**Result:**

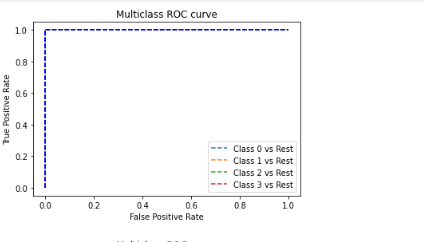
****

1. **Results of Classifier Decision Tree Classifier:**

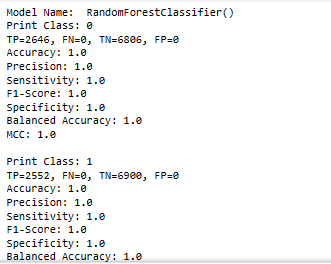
****

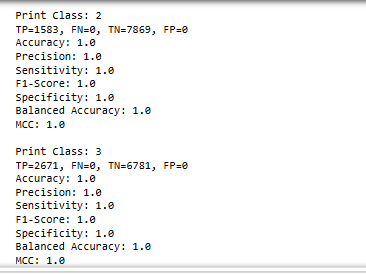
****

**Result:**

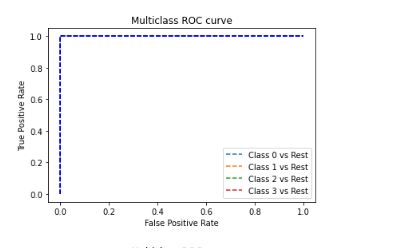
****

1. **Results of Random Forest Classifier :**

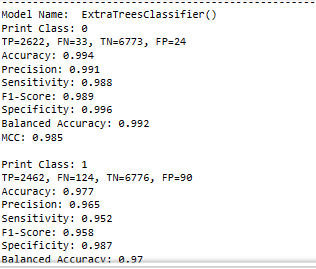
****

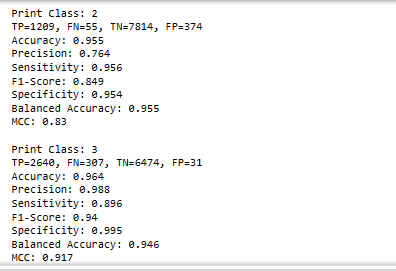
****

**Result:**

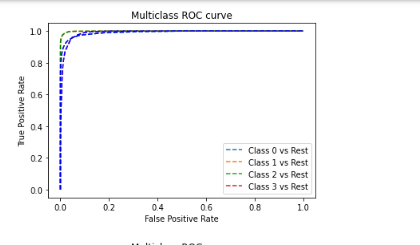
****

1. **Results of Extra Trees Classifier:**

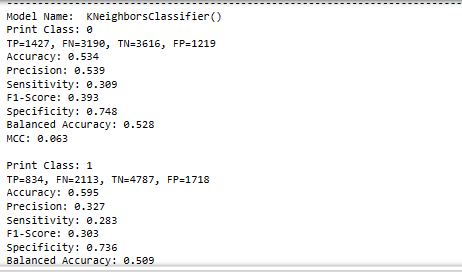
****

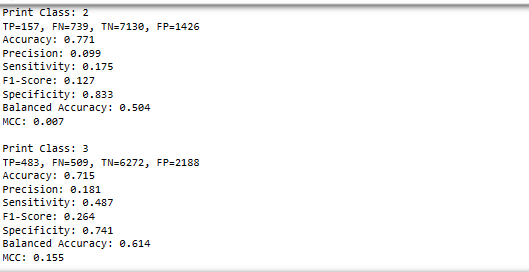
****

**Result :**

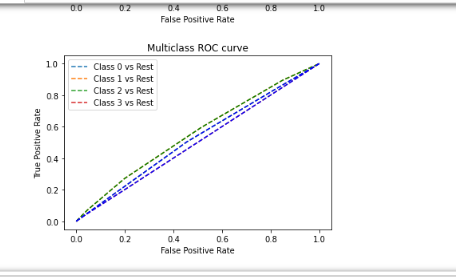
****

1. **Results of Kneighbours Classifier :**

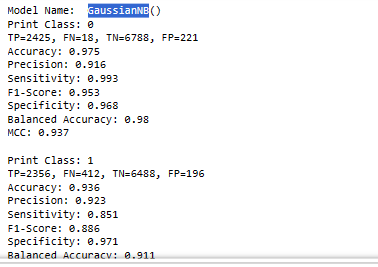
****

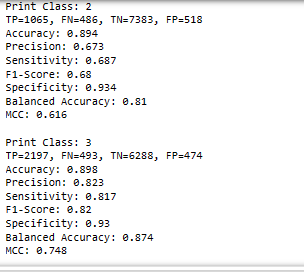
****

**Result:**

****

1. **Results of GuassianNB:**

****

****

**Result:**

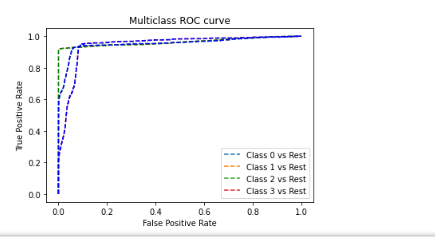


Fig: 5.0 ANALYSIS

## 

## 6.0 CONCLUSION

In this case, the BayesianRidge model is the better choice. The MAE is a measure of how close the predicted values are to the actual values, and the RMSE is a measure of how spread out the errors are. The lower the MAE and RMSE, the better the model is at predicting the actual values.

The BayesianRidge model has a lower MAE and RMSE than the DecisionTreeRegressor model, which means that it is closer to the actual values and the errors are less spread out. This suggests that the BayesianRidge model is a better choice for predicting the values in the dataset.

**7.0 REFERENCES**

<https://en.wikipedia.org/wiki/Machine_learning>

Dataset – <https://www.kaggle.com/datasets/shashwatwork/insurance-claim-dataset>

<https://www.javatpoint.com/machine-learning-random-forest-algorithm>

## 