A Novel Framework for Risk Prediction in the Health Insurance Sector using GIS and Machine Learning

Prasanta Baruah, Pankaj Pratap Singh, Sanjiv kumar Ojah
Department of Computer Science & Engineering, Central Institute of Technology Kokrajhar, Kokrajhar, India

Abstract—**Evaluation of risk is a key component to categorize the customers of the life insurance businesses. The underwriting technique is carried out by the industries to charge the policies appropriately. Due to the availability of data hugely, the automation of underwriting process can be done using data analytics technology. Due to this, the underwriting process becomes faster and therefore quickly processes a large number of applications. This study is carried to enhance risk assessment of the applicants of life insurance industries using predictive analytics. In this research, the Geographical Information Systems (GIS) system is used to collect the data such as Air pollution, Industrial area, Covid-19 and Malaria of various geographic areas of our country, since these factors attribute to the risk of an applicant of life insurance business. Thereafter, the research is carried out using this dataset along with another dataset containing more than 50,000 entries of normal attributes of applicants of a life insurance company. Artificial Neural Network (ANN), Decision Tree (DT), and Random forest (RF) algorithms are applied on both the datasets to predict the risks of the applicants. The results showed that random forest outperformed among all the algorithms, providing the more accurate result.**

Keywords—Risk prediction; data analytics; predictive analytics; underwriting; geographical information systems; random forest; artificial neural network; decision tree

I. INTRODUCTION

Assessment of risk plays a significant role in the classification of applicants in any life insurance industry. The life insurance businesses use underwriting procedure to choose applications and set insurance product prices accordingly. For quicker application processing, the underwriting procedure might be automated thanks to the growth of data and developments in data analytics [13]. In order to develop solutions that cater to the demands of various client and market groups, insurance companies are giving emphasis of using machine learning (ML) techniques on big data. The evaluation comprises estimating the company's risk factor and offering potential employees health insurance based on their medical histories.

Developing a scalable risk assessment model for the customer segment of the insurance domain using geospatial technologies alike Global Positioning System (GPS) is very much suitable, particularly for surveillance of areas appearing infectious disease or environmental health hazards such as air pollution, water pollution, etc.,. This research discusses how the insurance companies can use a GIS data model to analyze these type parameters across the globe while addressing their biggest bottleneck. Numerous tools and systems will be developed that enable the visualization of disease/health hazard data in location and time as a result of this ongoing public health burden and technological advancements with spatial data [5]. This geospatial technology will therefore offer insurance businesses and decision maker's visualization and analytical tools to conduct life insurance programs for clients in afflicted and/or suspected locations as well as analysis and forecasts that were previously technologically impractical.

The research problem addressed in this study is the development of an innovative framework leveraging Geographic Information Systems (GIS) and Machine Learning techniques for enhanced risk prediction in the Health Insurance Sector. The research question comes like as "What specific methodologies and algorithms within GIS and Machine Learning can enhance health insurance risk prediction?" The primary objectives of this study are to design and implement a novel GIS and Machine Learning framework for accurate health insurance risk prediction, assess its performance, and provide recommendations for practical implementation in the health insurance sector. The research holds significance by offering an innovative approach to health insurance risk prediction, potentially improving accuracy and decision-making in the sector through the integration of GIS and Machine Learning technologies. The main research contribution is the development of an advanced and novel framework that integrates GIS and Machine Learning to enhance the accuracy and effectiveness of risk prediction in the Health Insurance Sector.

GIS, GPS, and satellite-based technologies such as Remote Sensing (RS) are all examples of geospatial technology. GIS refers to the collection, input, updating, modification, transformation, analysis, query, modeling, and visualization of geographic data utilizing a collection of computer programs. The next Section II and Section III will discuss about the related research work and methodology respectively. The detailed results related discussion is mentioned in Section IV and finally conclusion given in Section V followed by the references section of this paper.

II. RELATED WORK

A map-based dashboard was proposed for visualizing the COVID19 pandemic in order to deliver information to individuals all around the world who want to safeguard themselves and their communities and how a complete GIS platform may aid in the surveillance, preparedness, and response of infectious diseases [3]. By collecting data from satellites can be made available to users. Satellites are terrain surveillance equipment that provides regional information on climatic parameters and terrain features. Additionally, GPS uses satellite data to provide locating, direction-finding, and timing services. As a result, while GPS and RS provide local and spatial information, GIS offers accurate geospatial analysis and real-time geospatial data integration. This study is conducted and found that machine learning algorithms like DT, RF, and ANN are effective in estimating the risk level of applicants in an insurance industry [4, 7].

In this paper, a potential risk variables was investigated those contribute to the cases of COVID-19 at the various districts of Bangladesh. In this work, three global models and one local model were built based on demographic, economic, built environment along with the factors like health and facilities which affect rates of COVID-19 occurrence cases [10]. It was found that the percentage of urban population of the districts was responsible for the COVID-19 occurrence rates. This is due to the fact that in high-density urban regions, movements of people as well as activities are higher than the non-urban regions. The researchers discovered that the higher the inhabitant’s density, there is more possibility that a person will come into contact with an infector. In this paper, a framework is presented which shows the different flood hazards in spatial hotspots areas and also assessed the vulnerability in the districts using MODIS data [16]. A ML approach was used to classify and find risk based on diabetics disease [17].

In the univariate analysis, population density was also revealed to be a significant variable in this study [2]. In this paper, authors found that in the OCHIN (Oregon Community Health Information Network) PBRN (practice-based research network) consist of community health centres. These networks were used to display of EHR (electronic health record) data using GIS web based mapping and thereby serving in detecting societies having higher number of patients without health insurance. The author suggested that this strategy might be adapted for use by PBRNs, primary care providers, public health officials, and others to recognize a wide range of practice and community needs and to correctly implement targeted interventions using additional EHR data pieces [5]. This paper work added current COVID-19 assessments by providing a geographical viewpoint. It's a collection that recognizes the themes and analyses that GIS and spatial-statistical tools are being used for [1]. The detailed comparative studies are mentioned below in the Table I which is done by the several researchers.

The main objective of this research work is to develop a software model that makes use of web GIS technology to calculate risk in a specific location using information for the health insurance industry. As a result, the insurance company might carry out a more thorough examination and come to better judgments that will benefit both their clients and the business. The insurance company may compare data from the previous year to help it make better decisions. Additionally, with the use of various hazard data, such as information on air pollution, malaria, COVID-19, industrial regions, etc., the companies may analyze the geographic area and make better decisions in order to provide consumers a variety of life insurance plans.

TABLE I. COMPARATIVE STUDY OF THE ALGORITHMS USED IN THE RELATED WORKS

Authors	Year	Objective	Methods used	Method giving the best performance
Rahman et al. [2]	2021	To identify the risk factors of Covid-19	GIS based spatial model	GIS model
Saran et al. [6]	2020	Reviewing of Geospatial Technology	GIS approach and dynamic modelling algorithms	GIS approach
Boulos et al. [3]	2020	Tracing and mapping of Covid-19 patients	GIS system	GIS system
Pardo et al. [1]	2020	Covid-19 analysis	GIS and spatial analysis	GIS method
Angier et al. [5]	2014	Identification of communities in need of health insurance	GIS web based	GIS web based
Boodhun et al. [4]	2018	Implementation of ML algorithms to classify the applicants risk levels	Multiple Linear Regression (MLR), ANN, RF, REPTree algorithms	REPTree performed better with Correlation Based Feature Selection (CFS) whereas MLR showed the best performance using Principal Components Analysis (PCA) method
Hutagaol et al. (2020) [11]		Examined risk level of customers using ML in life insurance companies	RF, Support Vector Machine (SVM) and Naive Bayesian algorithms	RF have highest precision in comparison to the SVM and Naive Bayesian (N.B.) algorithm
Mustika et al. [14]	2019	Applied a ML models to predict the risk level of applicants in life insurance	Extreme Gradient tree boosting (XGBoost), DT, RF and Bayesian ridge models	XGBoost model is provided more accurate result
Jain et al. [12]	2019	An ensemble learning method for assessing risk associated with a policy applicant	ANN and gradient boosting algorithm XGBoost	XGBoost provided the best result.
Biddle et al. [13]	2018	To automate the underwriting process	Logistic Regression, XGBoost and Recursive Feature Elimination	XGBoost is the most ideal one giving better accurac.
Dwivedi et. al. [15]	2020	ML algorithms are used to predict the risk levels of applicants	ANN, MLR, RT and RF	RF came out to be most efficient one.

III. Methodology

As part of the research methodology, data is acquired from online databases. The research paradigm adopts a positivist stance because the study is largely predictive and uses machine learning and some geospatial approaches to assist this research work goals. The major task of the gathered data is to use Quantum-GIS (QGIS) to turn all of the non-spatial data into spatial data. The process flow of the model which utilizes different techniques is shown in the Fig. 1. The details of the data, format, source and description are mentioned in the given below Table II.

A. Data Collection and Preparation

The 59,381 applications that make up the life insurance data collection each have 128 attributes that describe the traits of applicants for life insurance. The data set contains anonymised nominal, continuous, and discrete variables. Customer related sample dataset of the applicants is shown in the Table III. The data pre-processing, commonly referred to as data cleaning, which comprises removing erratic data or outliers from the target dataset. This step also includes developing any methods necessary to deal with the target data's inconsistencies. To facilitate analysis and interpretation in the event of disputes, certain variables will be changed. In this phase, the data will be cleaned to get rid of any missing values and make sure it can be used for the analysis. To determine the optimal imputation procedure for the dataset, it will be necessary to look into the structure and methodology of missing data.

B. Geospatial Technology

Geospatial technologies and services can help with the representation of spatio-temporal data and the analysis of dynamic spread when illnesses or hazards are present [8]. A “spatial health information infrastructure” must be built using geospatial technology and real time services. In order to better comprehend analysis the spread and outbreak of the phenomenon, a review of several geospatial technologies with enabled IT services will be conducted in this section using a case study on the COVID-19 pandemic, malaria illness, air pollution, and industrial region.

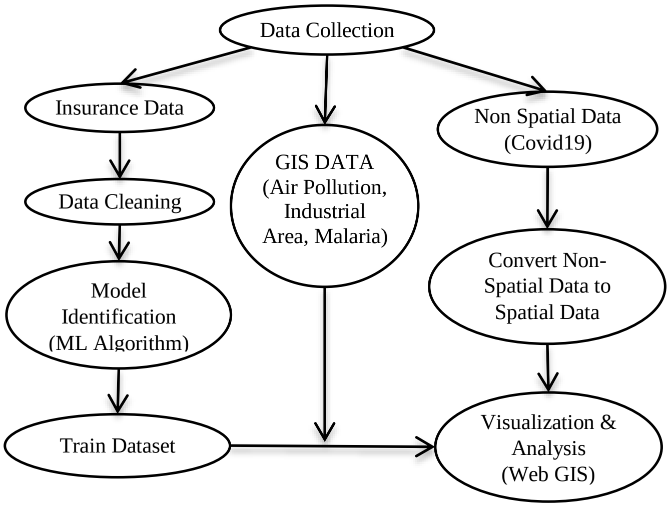

Data Collection
Insurance Data
Data Cleaning
Model Identification (ML Algorithm)
Train Dataset
GIS DATA (Air Pollution, Industrial Area, Malaria)
Non Spatial Data (Covid19)
Convert Non-Spatial Data to Spatial Data
Visualization & Analysis (Web GIS)

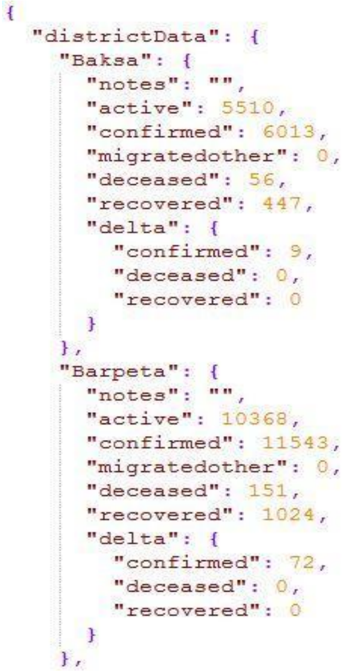
Fig. 1. Flowchart of the system.

TABLE II. Comparative Data Source and Description

Data	Source and Information	Format
Air Quality	MODIS	Raster Data
Industrial Area	NESDR	Vector Data
Covid19	covid19india.org	JSON
Malaria	NESDR	GEOJSON

TABLE III. Sample Dataset of Customer

Id	Product_Info_1	Ins_Age	Ht	Wt	BMI	Employment_Info_1
2	1	0.641791	0.581818	0.148536	0.323008	0.028
5	1	0.059701	0.6	0.131799	0.272288	0
6	1	0.029851	0.745455	0.288703	0.42878	0.03
7	1	0.164179	0.672727	0.205021	0.352438	0.042
8	1	0.41791	0.654545	0.23431	0.424046	0.027
10	1	0.507463	0.836364	0.299163	0.364887	0.325
11	1	0.373134	0.581818	0.17364	0.376587	0.11
14	1	0.61194	0.781818	0.403766	0.571612	0.12
15	1	0.522388	0.618182	0.1841	0.362643	0.165
16	1	0.552239	0.6	0.284519	0.587796	0.025
17	1	0.537313	0.690909	0.309623	0.521668	0.05
18	1	0.298507	0.690909	0.271967	0.45505	0.09
19	1	0.567164	0.618182	0.16318	0.320784	0.075
20	2	0.223881	0.781818	0.361925	0.507515	0.1
22	1	0.328358	0.636364	0.142259	0.264648	0.16
23	1	0.626866	0.672727	0.330544	0.581279	0.075



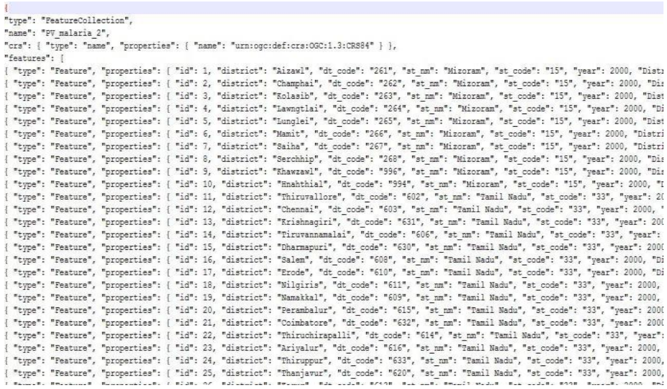
```
{
  "districtData": {
    "Baksa": {
      "notes": "",
      "active": 5510,
      "confirmed": 6013,
      "migratedother": 0,
      "deceased": 56,
      "recovered": 447,
      "delta": {
        "confirmed": 9,
        "deceased": 0,
        "recovered": 0
      }
    },
    "Barpeta": {
      "notes": "",
      "active": 10368,
      "confirmed": 11543,
      "migratedother": 0,
      "deceased": 151,
      "recovered": 1024,
      "delta": {
        "confirmed": 72,
        "deceased": 0,
        "recovered": 0
      }
    },
```

Fig. 2. COVID-19 sample data in JSON format.

Fig. 2 shows the sample data of covid-19 for the districts of Assam in JSON format. Fig. 3 shows the sample data of malaria-PV (Plasmodium Vivax) virus and Fig. 4 shows the sample data of malaria-PF (Plasmodium Falciparum) virus in GEOJSON format of all the districts of India [9]. The red marks are shown in the Fig. 5 which denotes the industrial areas of NER for the 3rd Land Use Land Cover (LULC) cycle (2015-2016). Fig. 6 shows the MODIS data of NER air pollution for 15 January 2021.

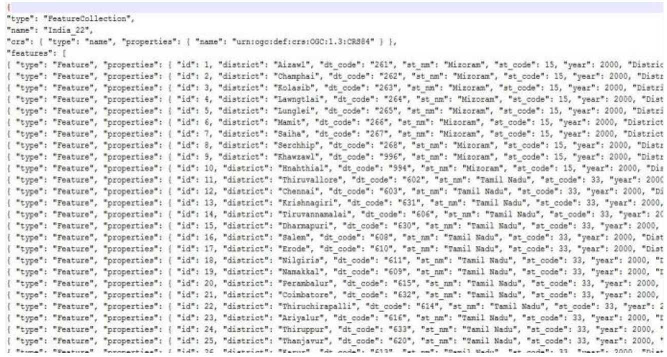
C. Model Development

The optimal model is chosen using a combination of decision tree and ensemble techniques because risk level is a multi-class variable. Combining data from a number of different models to increase accuracy and stability is referred to as an "ensemble." The many algorithms that were used to create the predictive models using the data set were covered in this section. Several machine learning methods, including DT, RF, and ANN have been applied on the dataset after pre-processing the data.



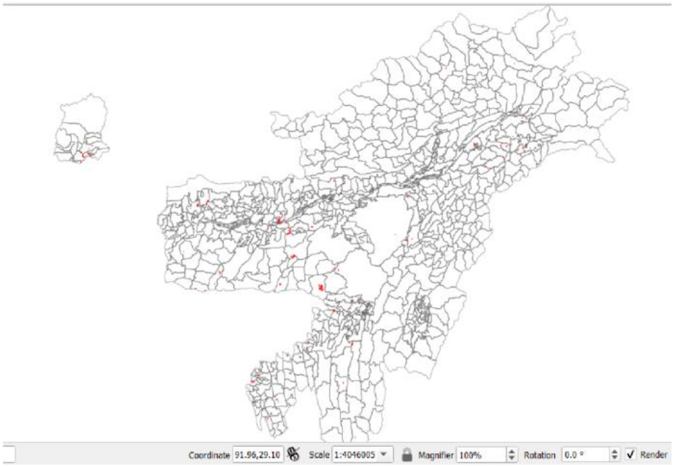
```
{
"type": "FeatureCollection",
"name": "PV_malaria_2",
"crs": { "type": "name", "properties": { "name": "urn:ogc:def:crs:OGC:1.3:CRS84" } },
"features": [
{ "type": "Feature", "properties": { "id": 1, "district": "Aizawl", "dt_code": "261", "st_nm": "Mizoram", "st_code": "15", "year": 2000, "Distr
{ "type": "Feature", "properties": { "id": 2, "district": "Champhai", "dt_code": "262", "st_nm": "Mizoram", "st_code": "15", "year": 2000, "Dis
{ "type": "Feature", "properties": { "id": 3, "district": "Kolasib", "dt_code": "263", "st_nm": "Mizoram", "st_code": "15", "year": 2000, "Dist
{ "type": "Feature", "properties": { "id": 4, "district": "Lawngtlai", "dt_code": "264", "st_nm": "Mizoram", "st_code": "15", "year": 2000, "Di
{ "type": "Feature", "properties": { "id": 5, "district": "Lunglei", "dt_code": "265", "st_nm": "Mizoram", "st_code": "15", "year": 2000, "Dist
{ "type": "Feature", "properties": { "id": 6, "district": "Mamit", "dt_code": "266", "st_nm": "Mizoram", "st_code": "15", "year": 2000, "Distri
{ "type": "Feature", "properties": { "id": 7, "district": "Saiha", "dt_code": "267", "st_nm": "Mizoram", "st_code": "15", "year": 2000, "Distri
{ "type": "Feature", "properties": { "id": 8, "district": "Serchhip", "dt_code": "268", "st_nm": "Mizoram", "st_code": "15", "year": 2000, "Dis
{ "type": "Feature", "properties": { "id": 9, "district": "Khawzawl", "dt_code": "996", "st_nm": "Mizoram", "st_code": "15", "year": 2000, "Dis
{ "type": "Feature", "properties": { "id": 10, "district": "Hnahthial", "dt_code": "994", "st_nm": "Mizoram", "st_code": "15", "year": 2000, "D
{ "type": "Feature", "properties": { "id": 11, "district": "Thiruvallore", "dt_code": "602", "st_nm": "Tamil Nadu", "st_code": "33", "year": 20
{ "type": "Feature", "properties": { "id": 12, "district": "Chennai", "dt_code": "603", "st_nm": "Tamil Nadu", "st_code": "33", "year": 2000, "
{ "type": "Feature", "properties": { "id": 13, "district": "Krishnagiri", "dt_code": "631", "st_nm": "Tamil Nadu", "st_code": "33", "year": 200
{ "type": "Feature", "properties": { "id": 14, "district": "Tiruvannamalai", "dt_code": "606", "st_nm": "Tamil Nadu", "st_code": "33", "year":
{ "type": "Feature", "properties": { "id": 15, "district": "Dharmapuri", "dt_code": "630", "st_nm": "Tamil Nadu", "st_code": "33", "year": 2000
{ "type": "Feature", "properties": { "id": 16, "district": "Salem", "dt_code": "608", "st_nm": "Tamil Nadu", "st_code": "33", "year": 2000, "Di
{ "type": "Feature", "properties": { "id": 17, "district": "Erode", "dt_code": "610", "st_nm": "Tamil Nadu", "st_code": "33", "year": 2000, "Di
{ "type": "Feature", "properties": { "id": 18, "district": "Nilgiris", "dt_code": "611", "st_nm": "Tamil Nadu", "st_code": "33", "year": 2000,
{ "type": "Feature", "properties": { "id": 19, "district": "Namakkal", "dt_code": "609", "st_nm": "Tamil Nadu", "st_code": "33", "year": 2000,
{ "type": "Feature", "properties": { "id": 20, "district": "Perambalur", "dt_code": "615", "st_nm": "Tamil Nadu", "st_code": "33", "year": 2000
{ "type": "Feature", "properties": { "id": 21, "district": "Coimbatore", "dt_code": "632", "st_nm": "Tamil Nadu", "st_code": "33", "year": 2000
{ "type": "Feature", "properties": { "id": 22, "district": "Thiruchirapalli", "dt_code": "614", "st_nm": "Tamil Nadu", "st_code": "33", "year":
{ "type": "Feature", "properties": { "id": 23, "district": "Ariyalur", "dt_code": "616", "st_nm": "Tamil Nadu", "st_code": "33", "year": 2000,
{ "type": "Feature", "properties": { "id": 24, "district": "Thiruppur", "dt_code": "633", "st_nm": "Tamil Nadu", "st_code": "33", "year": 2000,
{ "type": "Feature", "properties": { "id": 25, "district": "Thanjavur", "dt_code": "620", "st_nm": "Tamil Nadu", "st_code": "33", "year": 2000,
{ "type": "Feature", "properties": { "id": 26, "district": "Karur", "dt_code": "613", "st_nm": "Tamil Nadu", "st_code": "33", "year": 2000, "Di
```

Fig. 3. Malaria- PV (Plasmodium Vivax) sample data in GEOJSON format.



```
{
"type": "FeatureCollection",
"name": "India_22",
"crs": { "type": "name", "properties": { "name": "urn:ogc:def:crs:OGC:1.3:CRS84" } },
"features": [
{ "type": "Feature", "properties": { "id": 1, "district": "Aizawl", "dt_code": "261", "st_nm": "Mizoram", "st_code": 15, "year": 2000, "Distric
{ "type": "Feature", "properties": { "id": 2, "district": "Champhai", "dt_code": "262", "st_nm": "Mizoram", "st_code": 15, "year": 2000, "Distr
{ "type": "Feature", "properties": { "id": 3, "district": "Kolasib", "dt_code": "263", "st_nm": "Mizoram", "st_code": 15, "year": 2000, "Distri
{ "type": "Feature", "properties": { "id": 4, "district": "Lawngtlai", "dt_code": "264", "st_nm": "Mizoram", "st_code": 15, "year": 2000, "Dist
{ "type": "Feature", "properties": { "id": 5, "district": "Lunglei", "dt_code": "265", "st_nm": "Mizoram", "st_code": 15, "year": 2000, "Distri
{ "type": "Feature", "properties": { "id": 6, "district": "Mamit", "dt_code": "266", "st_nm": "Mizoram", "st_code": 15, "year": 2000, "District
{ "type": "Feature", "properties": { "id": 7, "district": "Saiha", "dt_code": "267", "st_nm": "Mizoram", "st_code": 15, "year": 2000, "District
{ "type": "Feature", "properties": { "id": 8, "district": "Serchhip", "dt_code": "268", "st_nm": "Mizoram", "st_code": 15, "year": 2000, "Distr
{ "type": "Feature", "properties": { "id": 9, "district": "Khawzawl", "dt_code": "996", "st_nm": "Mizoram", "st_code": 15, "year": 2000, "Distr
{ "type": "Feature", "properties": { "id": 10, "district": "Hnahthial", "dt_code": "994", "st_nm": "Mizoram", "st_code": 15, "year": 2000, "Dis
{ "type": "Feature", "properties": { "id": 11, "district": "Thiruvallore", "dt_code": "602", "st_nm": "Tamil Nadu", "st_code": 33, "year": 200(
{ "type": "Feature", "properties": { "id": 12, "district": "Chennai", "dt_code": "603", "st_nm": "Tamil Nadu", "st_code": 33, "year": 2000, "Di
{ "type": "Feature", "properties": { "id": 13, "district": "Krishnagiri", "dt_code": "631", "st_nm": "Tamil Nadu", "st_code": 33, "year": 2000,
{ "type": "Feature", "properties": { "id": 14, "district": "Tiruvannamalai", "dt_code": "606", "st_nm": "Tamil Nadu", "st_code": 33, "year": 2(
{ "type": "Feature", "properties": { "id": 15, "district": "Dharmapuri", "dt_code": "630", "st_nm": "Tamil Nadu", "st_code": 33, "year": 2000,
{ "type": "Feature", "properties": { "id": 16, "district": "Salem", "dt_code": "608", "st_nm": "Tamil Nadu", "st_code": 33, "year": 2000, "Dist
{ "type": "Feature", "properties": { "id": 17, "district": "Erode", "dt_code": "610", "st_nm": "Tamil Nadu", "st_code": 33, "year": 2000, "Dist
{ "type": "Feature", "properties": { "id": 18, "district": "Nilgiris", "dt_code": "611", "st_nm": "Tamil Nadu", "st_code": 33, "year": 2000, "I
{ "type": "Feature", "properties": { "id": 19, "district": "Namakkal", "dt_code": "609", "st_nm": "Tamil Nadu", "st_code": 33, "year": 2000, "I
{ "type": "Feature", "properties": { "id": 20, "district": "Perambalur", "dt_code": "615", "st_nm": "Tamil Nadu", "st_code": 33, "year": 2000,
{ "type": "Feature", "properties": { "id": 21, "district": "Coimbatore", "dt_code": "632", "st_nm": "Tamil Nadu", "st_code": 33, "year": 2000,
{ "type": "Feature", "properties": { "id": 22, "district": "Thiruchirapalli", "dt_code": "614", "st_nm": "Tamil Nadu", "st_code": 33, "year": 2
{ "type": "Feature", "properties": { "id": 23, "district": "Ariyalur", "dt_code": "616", "st_nm": "Tamil Nadu", "st_code": 33, "year": 2000, "I
{ "type": "Feature", "properties": { "id": 24, "district": "Thiruppur", "dt_code": "633", "st_nm": "Tamil Nadu", "st_code": 33, "year": 2000, "
{ "type": "Feature", "properties": { "id": 25, "district": "Thanjavur", "dt_code": "620", "st_nm": "Tamil Nadu", "st_code": 33, "year": 2000, "
{ "type": "Feature", "properties": { "id": 26, "district": "Karur", "dt_code": "613", "st_nm": "Tamil Nadu", "st_code": 33, "year": 2000, "Dist
```

Fig. 4. Malaria- PF (Plasmodium Falciparum) sample data in GEOJSON format.



Coordinate 91.96,29.10 Scale 1:4046005 Magnifier 100% Rotation 0.0 ° ✓ Render

Fig. 5. Industrial area depiction in NER region using QGI software environment.

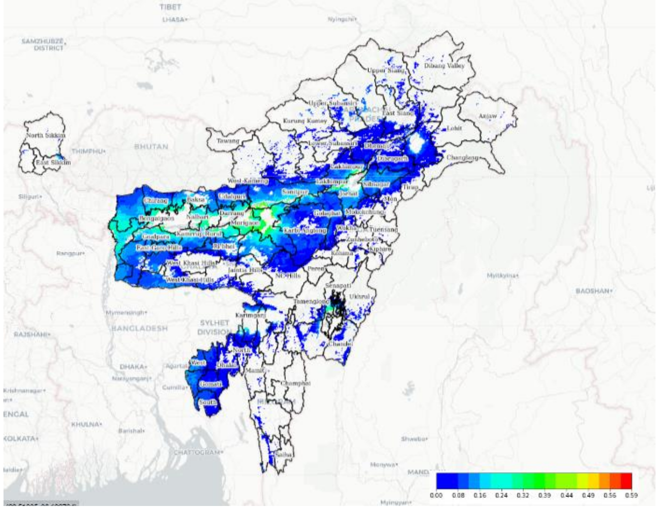


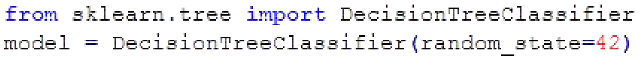
Fig. 6. MODIS data of NER air pollution.

1) Decision Tree: The SKlearn module in Python is used to build the DT as shown in Fig. 7 and also the information gain is calculated with the help of sample estimation which are shown below in Eq. (1), (2) and (3). The accuracy of this algorithm is coming 80%.

$$\mathrm{E(s)} = \sum_{i=1}^{c}(-p_i \log_2 p_i) \quad (1)$$

$$\mathrm{E(T, X)} = \sum_{c \in X} P(c)E(c) \quad (2)$$

$$\mathrm{InforGain\ (T, X) = E(T) - E(T, X)} \quad (3)$$



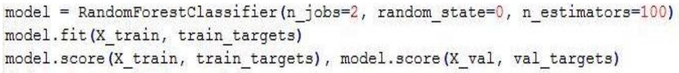
```
from sklearn.tree import DecisionTreeClassifier
model = DecisionTreeClassifier(random_state=42)
```

Fig. 7. Decision tree code snippet.

2) Random Forest (RF): The SKlearn module in Python is used to build the Random Forest technique as shown in Fig. 8, and 100 decision trees are employed to provide the final output for classification. The feature importance value for RF is calculated with the help of normalized importance of feature which are shown below in Eq. (4) and Eq. (5). The accuracy of the algorithm is coming 95%.

$$\mathrm{normfi}_i = fi_i / \sum_{j \in all\ features} fi_j \quad (4)$$

$$\mathrm{RFfi}_i = \sum_j normfi_{ij} / \sum_{j \in all\ features, k \in all\ trees} normfi_{jk} \quad (5)$$

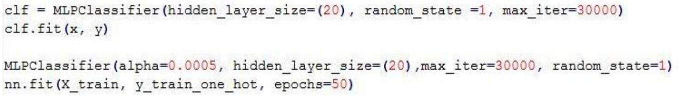


```
model = RandomForestClassifier(n_jobs=2, random_state=0, n_estimators=100)
model.fit(X_train, train_targets)
model.score(X_train, train_targets), model.score(X_val, val_targets)
```

Fig. 8. Random forest code snippet.

3) Artificial Neural Network (ANN): The input, hidden, and output layers of neurons are the three that are chosen in this procedure. There are 20 units in the hidden layer and a 0.1 starting random weight, for a total of 30000 iterations. The ANN code snippet is mentioned in the below Fig. 9. The accuracy of the algorithm is 90%. Eq. (6) shows the output layer calculation. Fig. 10 shows the following diagram illustrates the neural network used in this application. The neural network is too large to be plotted.

$$\mathrm{f(x)} = \sum_{i=1}^{m}(w_{ij}x_i) + bias \quad (6)$$



```
clf = MLPClassifier(hidden_layer_size=(20), random_state =1, max_iter=30000)
clf.fit(x, y)

MLPClassifier(alpha=0.0005, hidden_layer_size=(20),max_iter=30000, random_state=1)
nn.fit(X_train, y_train_one_hot, epochs=50)
```

Fig. 9. ANN code snippet.

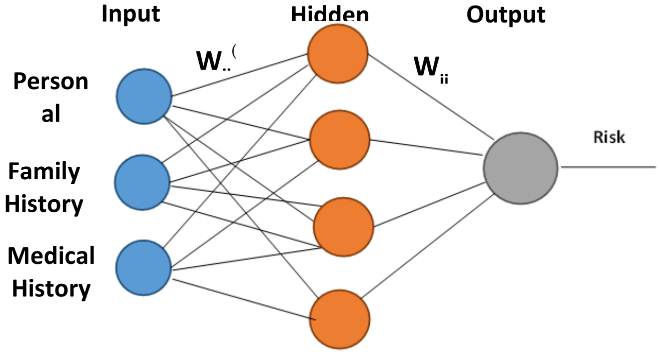
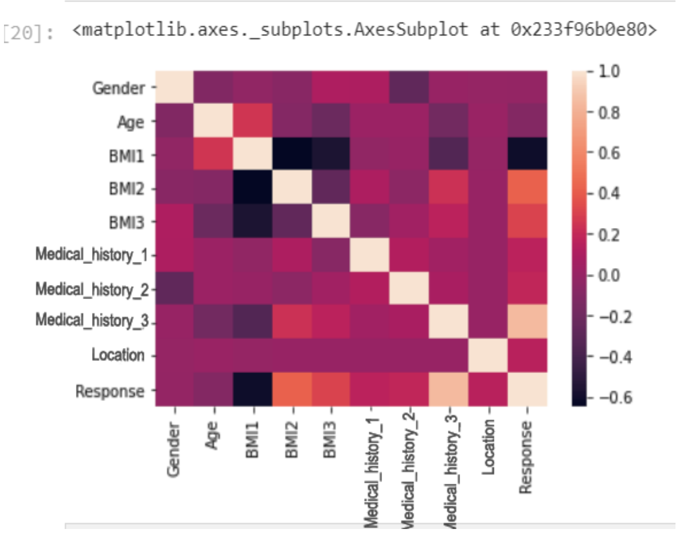


Fig. 10. A diagram of neural network for risk related application.

4) Heat Map: A heat map is a representation of two-dimensional data which shows the values having different intensities of colors. A simple heat map provides an immediate visual summary of information which is shown below in the Fig. 11. The heat map helps the viewer to understand and interpret the complex data sets. The color variation based on the different intensities provides understandable visual indications and also signifies the phenomenon of cluster or non-cluster over space.



```
[20]: <matplotlib.axes._subplots.AxesSubplot at 0x233f96b0e80>
```

Fig. 11. Heat Map of the insurance data.

IV. Results and Analysis

Here I have developed a web GIS-Dashboard using scripting languages like HTML, CSS, JavaScript as front end and PHP, python as back end. Geoserver is used here to publish different layers for displaying in this portal. Different base maps like Google map, Cartodb, Bhuvan map and Open street map were used here. In addition to its state boundary and district boundary of North East India were used here as overlay layers. Moreover, census data and three cycles of LULC data viz., LULC 1st cycle (2005-06), LULC 2nd cycle (2011-12) and LULC 3rd cycle (2015-16) are provided for analyzing and comparing trends over time. Different hazards like malaria (2019), covid-19 (2020-2021), MODIS air pollution data (2021) and industrial area (LULC 3rd cycle) were shown on this dashboard. Fig. 12 shows the UI of the dashboard. The default map that appears here is the Bhuvan map and NER district boundary on top of it. Different types of platforms are used for developing the web GIS-Dashboard which is as follows:

- Front End: HTML, CSS, JavaScript
- Back End : PHP, Python
- Web Server : XAMP, Geoserver
- Database : Postgres

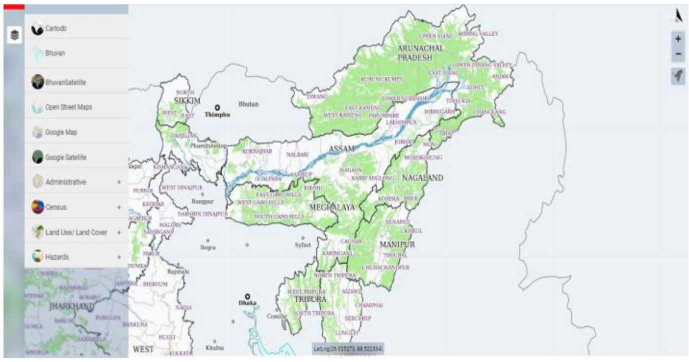


Fig. 12. Web portal designed using script language based environment.

Fig. 13 shows different base maps like Cartodb, Bhuvan, Open Street and Google map. In addition to it different overlay layers like state boundary, district boundary, census, LULC and different hazard layers are shown in this figure. The census data for Barpeta district is shown in Fig. 14, which includes total population, total work population, literacy, and the number of households.

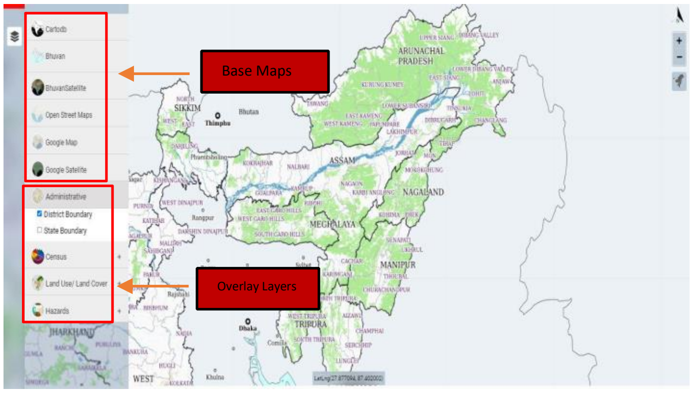


Fig. 13. Different base maps and overlay layers.

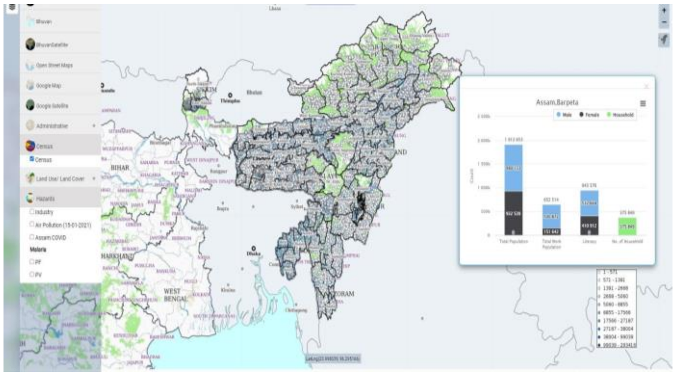


Fig. 14. Representation of census data for barpeta district.

Fig. 15 depicts the various LULC 1st cycle sectors for the Karbi Anglong district, including agriculture, built-up, forest, wasteland, and waterbodies. The analysis and comparison of all three LULC cycles in different sectors viz., agriculture, built up, waterbodies, snow, shifting cultivation, wastelands; forest for the Karbi Anglong area is shown in Fig. 16. In the Fig. 17, the risk level of different districts of Assam is depicted based on the insurance data and industrial data for the LULC 3rd cycle. The industrial area is shown in red marks for the NER region.

Fig. 18 displays the air pollution MODIS data for the NER region on January 15, 2021. Red indicates a high level of pollution, whereas blue indicates a low level of pollution. The figure shows various risk levels for insurance company applicants broken down by district. The following Figure 19 shows the corona virus count for the Assam district from January 2020 to October 2021. The graph shows the varied risk levels for insurance applicants in various districts of Assam.

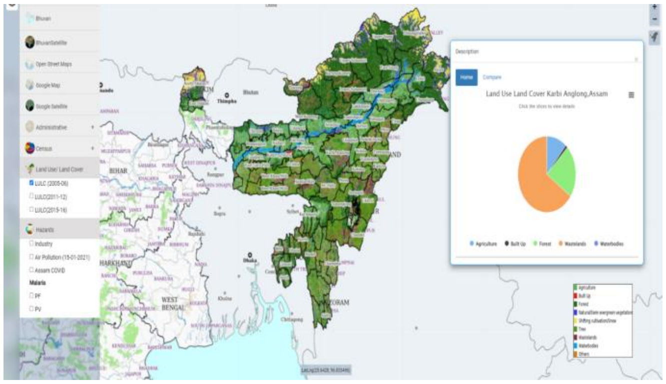


Fig. 15. Representation of LULC 1st cycle data for karbi anglong district.

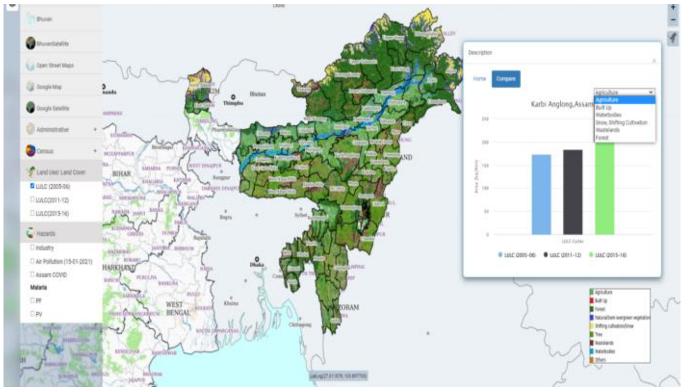
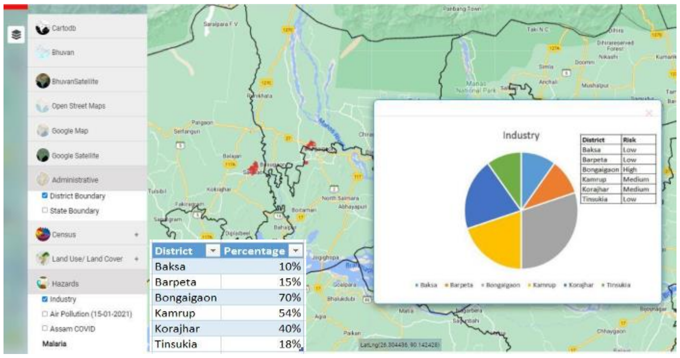


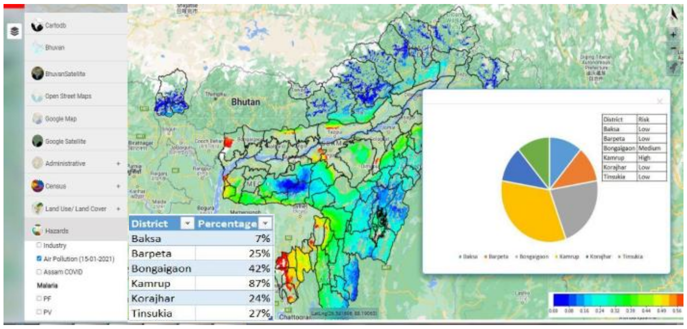
Fig. 16. Analysis and comparison of the LULC 3 cycle in several sectors for the karbi anglong district.



District	Percentage
Baksa	10%
Barpeta	15%
Bongaigaon	70%
Kamrup	54%
Korajhar	40%
Tinsukia	18%

District	Risk
Baksa	Low
Barpeta	Low
Bongaigaon	High
Kamrup	Medium
Korajhar	Medium
Tinsukia	Low

Fig. 17. Red marks on the map represent the industrial area for Assam district with risk level in various districts of Assam.

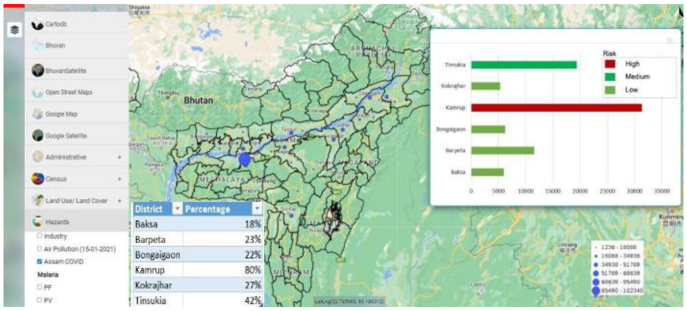


□ Industry
☑ Air Pollution (15-01-2021)
□ Assam COVID
Malaria
□ PF
□ PV

District	Percentage
Baksa	7%
Barpeta	25%
Bongaigaon	42%
Kamrup	87%
Korajhar	24%
Tinsukia	27%

District	Risk
Baksa	Low
Barpeta	Low
Bongaigaon	Medium
Kamrup	High
Korajhar	Low
Tinsukia	Low

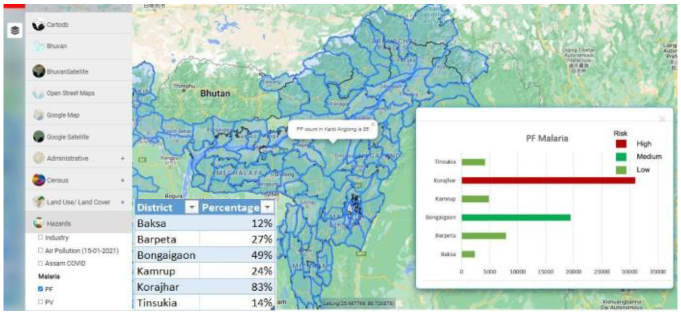
Fig. 18. MODIS data of NER air pollution for 15 January 2021 with risk level in various districts of Assam.



□ Industry
□ Air Pollution (15-01-2021)
☑ Assam COVID
Malaria
□ PF
□ PV

District	Percentage
Baksa	18%
Barpeta	23%
Bongaigaon	22%
Kamrup	80%
Kokrajhar	27%
Tinsukia	42%

Fig. 19. Corona virus count from January 2020 to October 2021 with the risk level in various districts of Assam.

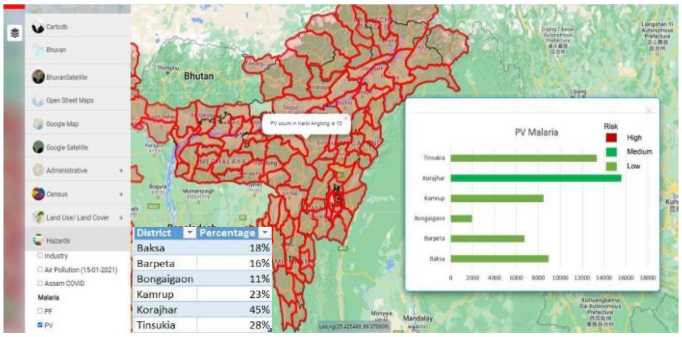


□ Industry
□ Air Pollution (15-01-2021)
□ Assam COVID
Malaria
☑ PF
□ PV

District	Percentage
Baksa	12%
Barpeta	27%
Bongaigaon	49%
Kamrup	24%
Korajhar	83%
Tinsukia	14%

Fig. 20. Analysis of malaria- PF count in the infected area of India with different risk levels.

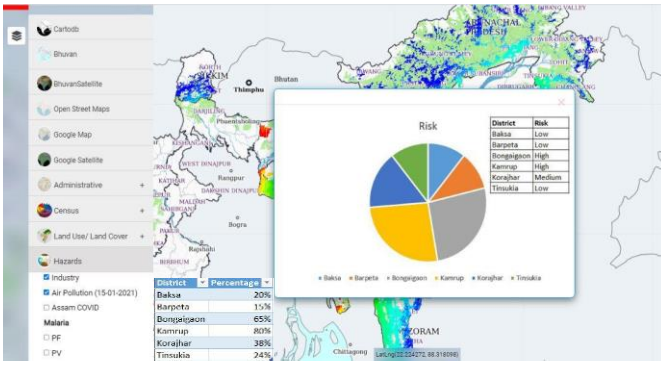
Fig. 20 displays the malaria-PF virus counts for various districts in India for the year 2019, and below Fig. 21 displays the malaria-PV virus numbers for various districts for the same year, with different risk levels for insurance company applicants for various districts of Assam. Fig. 22 shows the risk of the applicants in different districts of Assam based on two factors Industrial area and Air pollution. The result is depicted as pie chart and table as shown in the image. The risk of the applicants is based on three factors Industry, Air pollution and Covid for different districts of Assam as shown in Fig. 23. Fig. 24 shows the total risks of applicants in the insurance dataset based on three categories i.e., High, Medium and Low. The main insight and opinion offered by this research is that the integration of GIS and Machine Learning presents a promising and transformative approach for improving risk prediction in the Health Insurance Sector, providing valuable insights for industry stakeholders and policymakers.



□ Industry
□ Air Pollution (15-01-2021)
□ Assam COVID
Malaria
□ PF
☑ PV

District	Percentage
Baksa	18%
Barpeta	16%
Bongaigaon	11%
Kamrup	23%
Korajhar	45%
Tinsukia	28%

Fig. 21. Analysis of malaria- PV count in the infected area of India with different risk.

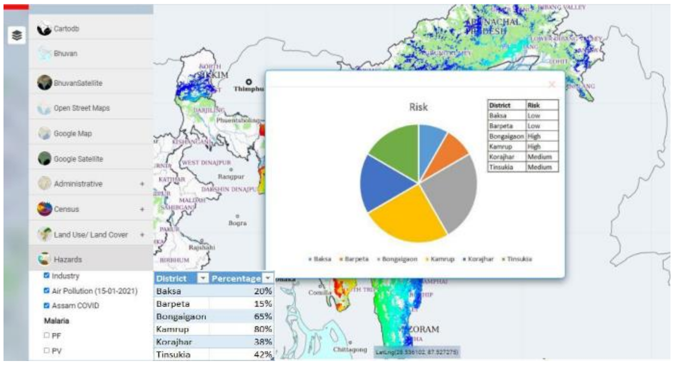


☑ Industry
☑ Air Pollution (15-01-2021)
□ Assam COVID
Malaria
□ PF
□ PV

District	Percentage
Baksa	20%
Barpeta	15%
Bongaigaon	65%
Kamrup	80%
Korajhar	38%
Tinsukia	24%

District	Risk
Baksa	Low
Barpeta	Low
Bongaigaon	High
Kamrup	High
Korajhar	Medium
Tinsukia	Low

Fig. 22. Total risk of the applicants is calculated for different districts considering two factors industry and air pollution.



☑ Industry
☑ Air Pollution (15-01-2021)
☑ Assam COVID
Malaria
□ PF
□ PV

District	Percentage
Baksa	20%
Barpeta	15%
Bongaigaon	65%
Kamrup	80%
Korajhar	38%
Tinsukia	42%

District	Risk
Baksa	Low
Barpeta	Low
Bongaigaon	High
Kamrup	High
Korajhar	Medium
Tinsukia	Medium

Fig. 23. Total risk of the applicants is calculated for different districts considering three factors industry, air pollution and covid.

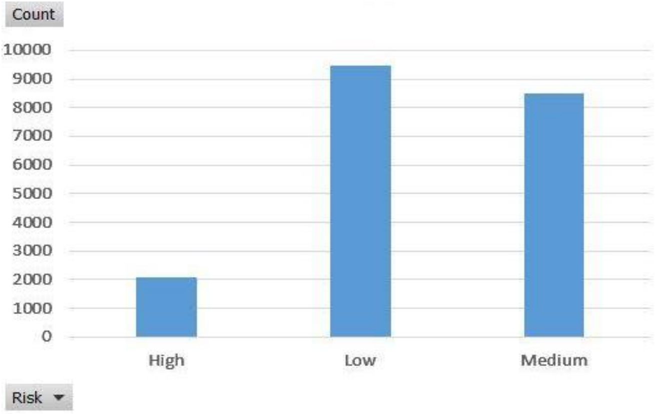


Fig. 24. Total risk of the applicants in 3 categories.

V. CONCLUSION

A GIS software model that makes use of online GIS technologies for analysis and visualization purposes which have been developed in this paper. The insurance provider might therefore be in a position to conduct a more thorough investigation and come to better judgments for its customers and business. Additionally, the corporate environment will specifically benefit from this research endeavour. Data visualization and analysis are becoming more and more common among enterprises all around the world. In this paper, several factors are incorporated in QGIS which provides risk analysis in better interpretable form. When compared to conventional methods, predictive modeling using GIS technologies can have a substantial impact on how business is performed in the life insurance market. Previously, lengthy and difficult actuarial calculations were used to evaluate risk for life underwriting. With the use of a web GIS map, the task may now be accomplished more quickly and with better results. As a result, it would benefit the company by enabling quicker customer service, thereby boosting satisfaction and loyalty.

The future research development is being planned in the following ways:

- Calculating premium for the customers in a specific geographic location based on the prediction.
- Collecting data in large quantities in order to increase the accuracy and usability of the model in real life situations.

REFERENCES

[1] I. Franch-Pardo, B. M. Napoletano, F. Rosete-Verges, and L. Billa, “Spatial analysis and GIS in the study of COVID-19. A review,” Science of the Total Environment, vol. 739, 2020.

[2] M. H. Rahman, N. M. Zafri, Fajle Rabbi Ashik, Md Waliullah, A. Khan, “Identification of risk factors contributing to COVID-19 incidence rates inBangladesh: A GIS-based spatial modeling approach,” Heliyon, vol. 7, pp. e06260, 2021.

[3] N. Maged, K. Boulos, E.M. Geraghty, “Geographical tracking and mapping of coronavirus disease COVID 19/severe acute respiratory syndrome coronavirus 2 (SARS CoV 2) epidemic and associated events around the world: how 21st century GIS technologies are supporting the global fightagainst outbreaks and epidemics,” International Journal ofHealth Geographics, vol. 19, 2020.

[4] N. Boodhun, Jayabalan, “M. Risk prediction in life insurance industry using supervised learningalgorithms,” *Complex & Intelligent Systems*, vol. 4, pp. 145-154, 2018.

[5] H. Angier, S. Likumahuwa, S. Finnegan, T. Vakarcs, C. Nelson, A. Bazemore, M. Carrozza, J. E. DeVoe, Using Geographic Information Systems (GIS) to Identify Communities in Need of Health Insurance Outreach: An OCHIN Practice-based ResearchNetwork (PBRN) Report. *JABFM* 27, 804-810, 2014.

[6] S. Saran, P. Singh, V. Kumar, P. Chauhan, “Review of Geospatial Technology for Infectious Disease Surveillance: Use Case on COVID-19,” *Journal of the Indian Society of Remote Sensing*, vol. 48, pp. 1121–1138, 2010.

[7] Changing face of the Insurance Industry. Available online: https://www.infosys.com/industries/insurance/whitepapers/ Documents/changing-face-insurance-industry,pdf (accessed on 09 November 2023).

[8] K. M. Al Kindi, A. Alkharusi, D. Alshukaili, N. A. Nasiri, T. A. Awadhi, Y. Charabi, A. M. E. Kenawy, “Spatiotemporal Assessment of COVID 19 Spread over Oman Using GIS Techniques,” *Earth Systems and Environment*, vol. 4, pp. 797-811, 2020.

[9] A. Das, A.R. Anvikar, L. J. Cator, R. C. Dhiman, A. Eapen, N. Mishra, B. N. Nagpal, N. Nanda, K. Raghavendra, A. F. Read, S. K. Sharma, O. P. Singh, V. Singh, P. Sinnis, H. C. Srivastava, S. A. Sullivan, P. L. Sutton, M. B. Thomas, J. M. Carlton, N. Valecha, “Malaria in India: The Center for the Study of Complex Malaria in India,” *Acta Trop.*, vol. 121, pp. 267-273, 2012.

[10] M. R. Rahman, A.H.M.H. Islam, M. N. Islam, “Geospatial modelling on the spread and dynamics of 154 day outbreak of the novel coronavirus (COVID 19) pandemic in Bangladesh towards vulnerability zoning and management approaches,” *Modeling Earth Systems and Environment*, vol. 7, pp. 2059-2087, 2021.

[11] B.J. Hutagaol, T. Mauritsius, “Risk level prediction of life insurance applicant using machine learning,” International Journal of Advanced Trends in Computer Science and Engineering, vol. 9, pp. 2213-2220, 2020.

[12] R. Jain, J.A. Alzubi, N. Jain, P. Joshi, “Assessing risk in life insurance using ensemble learning,” Journal of Intelligent & Fuzzy Systems, vol. 37, pp. 2969-2980, 2019.

[13] R. Biddle, S. Liu, P. Tilocca, G. Xu, “Automated underwriting in life insurance: Predictions and optimisation.&quot; Databases Theory and Applications,” *In the proceedings of 29th Australasian Database Conference (ADC)*, Gold Coast, QLD, Australia, 2018.

[14] W. F. Mustika, H. Murfi, Y. Widyaningsih, “Analysis Accuracy of XGBoost Model for Multiclass Classification - A Case Study of Applicant Level Risk Prediction for Life Insurance,” *In Proceedings of the 5th International Conference on Science in Information Technology* (ICSITech), Yogyakarta, Indonesia, pp. , 2019.

[15] S. Dwivedi, A. Mishra, A. Gupta, “Risk Prediction Assessment in Life Insurance Company through Dimensionality Reduction Method,” International Journal of Scientific & Technology Research, vol. 9, pp. 1528-1532, 2020.

[16] S. Roy, S.K. Ojah, N. Nishant, P.P. Singh, D. Chutia, “Spatio-temporal Analysis of Flood Hazard Zonation in Assam,” In: Gupta, D., Goswami, R.S., Banerjee, S., Tanveer, M., Pachori, R.B. (eds.): LNEE, vol. 888, Springer, Singapore, pp. 521-531, 2022.

[17] P. P. Singh, B. Das, U. Poddar, D. R. Choudhury, and S. Prasad, “Classification of diabetic’s patient data using machine learning techniques,” In: Perez, G., Tiwari, S., Trivedi, M., Mishra, K. (eds.): Ambient Communications and Computer Systems. Advances in Intelligent Systems and Computing, vol. 696, Springer Singapore, pp. 427-436, 2017.