

# **PROJECT REPORT**

on

## ***Fraud Detection System using graphs (in Neo4j)***

*(Btech AI&DS IV Semester Mini project)*

2020-24



***Submitted to:***

*Department of computer Applications*

***Submitted by:***

*Ayush Rawat(05)*

*University Roll Number: 2017639*

*Under the guidance of Ms. Garima Sharma*

DEPARTMENT OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY  
**GRAPHIC ERA DEEMED TO BE UNIVERSITY, DEHRADUN**

## **DECLARATION**

*I, **Ayush Rawat** student of **B-tech, Semester 4**, Department of Computer Science and Engineering, Graphic Era Deemed To Be University, Dehradun, declare that the technical project work entitled “Fraud Detection System using graphs” has been carried out by me and submitted in partial fulfillment of the course requirements for the award of degree in B-tech of **Graphic Era Deemed To Be University** during the academic year **2021-22**. The matter embodied in this synopsis has not been submitted to any other university or institution for the award of any other degree or diploma.*



## **ACKNOWLEDGEMENT**

I would like to take this opportunity to express my gratitude to entire faculty at Department of Computer Science and Information Technology, Graphic Era Deemed To Be University, Dehradun who evaluated the project from time to time and gave me the valuable suggestions as to how to improve the project.

I am grateful to **Ms. Garima Sharma**, Graphic Era Deemed To Be University, for her supervision, encouragement, inspiration, and guidance. Working under her is being an enriched experience. In all, I found congenial work environment in Graphic Era University, Dehradun and this project completion will mark a new beginning for me in the coming days.

I am highly indebted to Graphic Era University for providing me the required infrastructure and facilities to accomplish the given task.

Ayush Rawat

Btech AI&DS

2020-24

Graphic Era University

# **INSURANCE CLAIM FRAUD DETECTION**

## **➤ Problem Statement**

The aim of this project is to make a model that can detect a fraud in vehicle insurance claims. Frauds are unethical and losses to a company. By building a model that can detect insurance fraud, a company can get rid of the losses due to fraud which in turn will maximize their economy and earnings.

This project deals with insurance claim fraud detection using neo4j graph representation.

## **➤ Motivation:**

- India is one of the biggest markets for the insurance companies across the world.
- It is estimated that the Indian Insurance companies lose close to \$6 billion due to fraud per year.
- Insurance fraud can be submitting claims for injuries or damage that never happens and false reports of stolen vehicles.
- Hence the insurance companies have an urgent need to have a model that can detect fraud with high accuracy.

## **➤ Tools Used:**

- **Neo4j**- Neo4j is one of the popular Graph Database Management systems that uses Cypher Query Language (CQL). Neo4j is written in Java Language. Neo4j creates a graph in the form of nodes from the dataset and helps the user to easily understand the relationship between different nodes.
- **Google colab**- Google colab or collaborative allows anybody to write and execute python code through the browser, and is mainly used to create machine learning models, data analysis and education.

### ➤ Header Files Used

- 1) **Numpy**- NumPy is a general purpose Python library used for working with arrays, linear algebra, and matrices.
- 2) **Pandas**- Pandas is an open-source Python package used mainly for working with relational or labeled data easily.
- 3) **Matplotlib.pyplot**- Pyplot is a Matplotlib module used for creating animated and visualization in python. The various plots we can do using Pyplot are Line Plot, Histogram, Scatter, 3D Plot, Image, Contour, and Polar.
- 4) **Seaborn**- It is a python library basically used for data visualization and analysis. Seaborn is used to customize the graphs created and work in dataframes.

### ➤ Fraud in Automobile Insurance

Fraud in Automobile can be of various forms like:

- Fraudsters issue fake insurance claims
- Making fake reports of stolen vehicles
- Claiming money for pre-existing damages

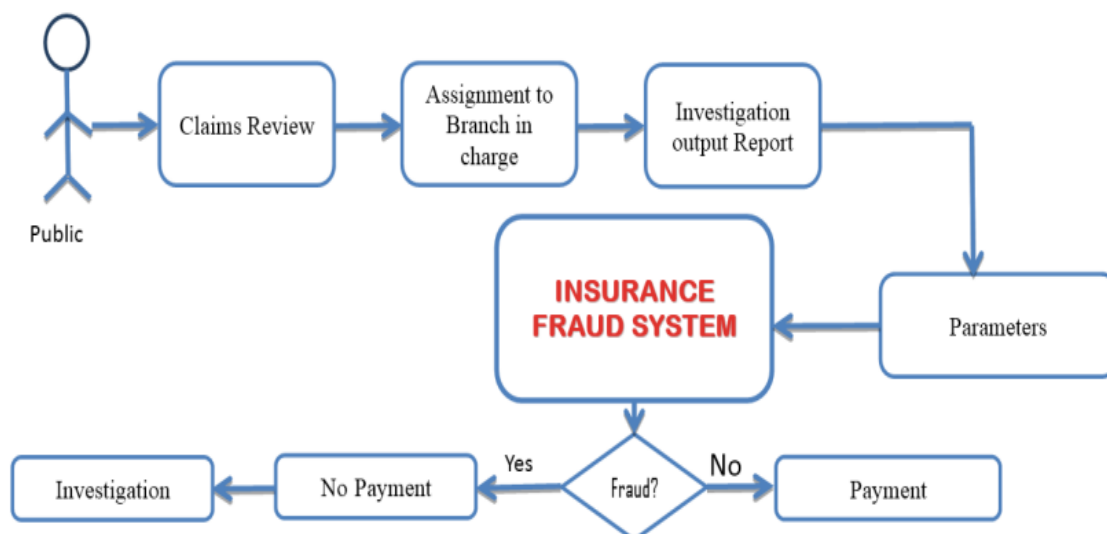


Figure 1: Architectural design of the proposed system

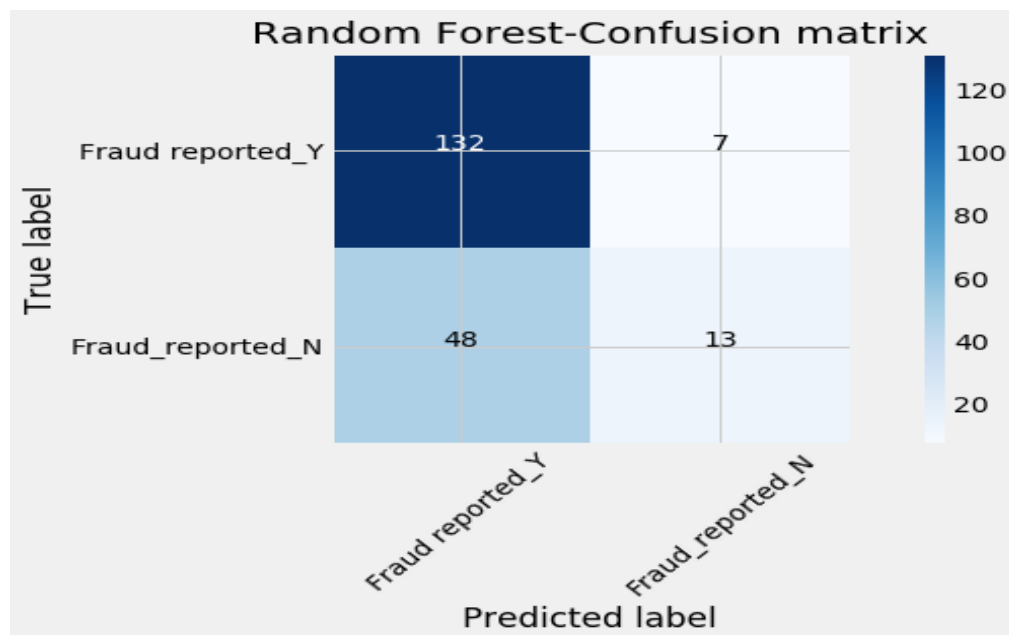
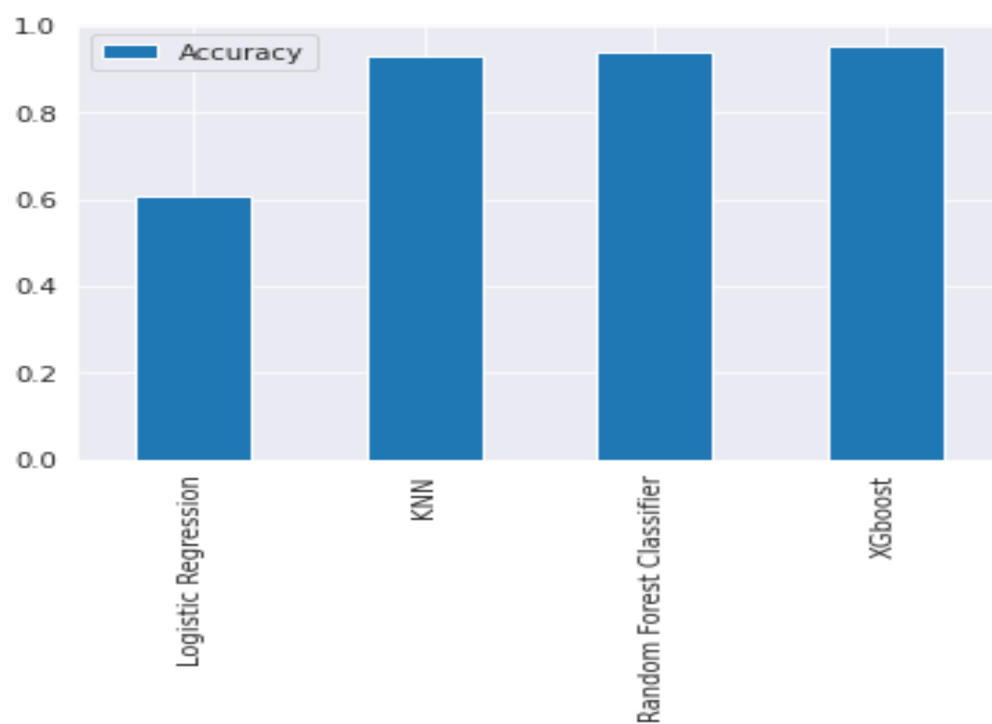
## ➤ Methodology

- 1) Data import from Dataset: The data stored in the CSV file need to be imported in our model.

POSSIBLE DATA LOSS: Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file format.

months	age	policy_nu	policy_bin	policy_sta	policy_cst	policy_dec	policy_anr	umbrella	insured_zi	insured	insured_e	insured_o	insured_h	insured_r	capital-gai	capital-los	incident	incident	collision	incident_s	authoritie	incident
1	328	48	521585	OH	250/500	1000	1406.91	0	466132	MALE	MD	craft-repa	sleeping	husband	53300	0	Single Veh Side Collis	Major Dar	Police	SC		
2	228	42	342868	IN	250/500	2000	1197.22	5000000	468176	MALE	MD	machine-c	reading	other-rela	0	0	Vehicle Th ?	Minor Dar	Police	VA		
3	134	29	687698	OH	100/300	2000	1413.14	5000000	430632	FEMALE	PhD	sales	board-gan	own-child	35100	0	Multi-vehi Rear Collis	Minor Dar	Police	NY		
4	256	41	227811	IL	250/500	2000	1415.74	6000000	608117	FEMALE	PhD	armed-for	board-gan	unmarriec	48900	-62400	Single Veh Front Collis	Major Dar	Police	OH		
5	228	44	367455	IL	500/1000	1000	1583.91	6000000	610706	MALE	Associate	sales	board-gan	unmarriec	66000	-46000	Vehicle Th ?	Minor Dar	None	NY		
6	256	39	104594	OH	250/500	1000	1351.1	0	478456	FEMALE	PhD	tech-supp	bungie-jur	unmarriec	0	0	Multi-vehi Rear Collis	Major Dar	Fire	SC		
7	137	34	413978	IN	250/500	1000	1333.35	0	441716	MALE	PhD	prof-speci	board-gan	husband	-77000	0	Multi-vehi Front Collis	Minor Dar	Police	NY		
8	165	37	429027	IL	100/300	1000	1137.03	0	603195	MALE	Associate	tech-supp	base-jum	unmarriec	0	0	Multi-vehi Front Collis	Total Loss	Police	VA		
9	27	33	489665	IL	100/300	500	1442.99	0	601734	FEMALE	PhD	other-serv	golf	own-child	0	0	Single Veh Front Collis	Total Loss	Police	WV		
10	212	42	636550	IL	100/300	500	1315.68	0	600983	MALE	PhD	priv-hous	camping	wife	-39300	0	Single Veh Front Collis	Total Loss	Other	NC		
11	235	42	543610	OH	100/300	500	1253.12	4000000	462283	FEMALE	Masters	exec-man	dancing	other-rela	38400	0	Single Veh Front Collis	Total Loss	Police	NY		
12	447	61	214618	OH	100/300	2000	1137.16	0	615561	FEMALE	High Scho	exec-man	skydiving	other-rela	-51000	0	Multi-vehi Front Collis	Major Dar	Fire	SC		
13	60	23	842643	OH	500/1000	500	1215.36	3000000	432220	MALE	MD	protective	reading	wife	0	0	Single Veh Rear Collis	Total Loss	Ambulanc	SC		
14	121	34	626808	OH	100/300	1000	936.61	0	464652	FEMALE	MD	armed-for	bungie-jur	wife	52800	-32800	Parked Ca ?	Minor Dar	None	SC		
15	180	38	644081	OH	250/500	2000	1301.13	0	476685	FEMALE	College	machine-c	board-gan	not-in-fam	41300	-55500	Single Veh Rear Collis	Total Loss	Police	SC		
16	473	58	892874	IN	100/300	2000	1131.4	0	558733	FEMALE	MD	transport-	movies	other-rela	55700	0	Multi-vehi Side Collis	Major Dar	Other	WV		
17	70	26	558938	OH	500/1000	1000	1199.44	5000000	619884	MALE	College	machine-c	hiking	own-child	63600	0	Multi-vehi Rear Collis	Major Dar	Other	NY		
18	140	31	275265	IN	500/1000	500	708.64	6000000	470610	MALE	High Scho	machine-c	reading	unmarriec	53500	0	Single Veh Side Collis	Total Loss	Police	WV		
19	160	37	921202	OH	500/1000	500	1374.22	0	472135	FEMALE	MD	craft-repa	yachting	other-rela	45500	-37800	Single Veh Side Collis	Total Loss	Other	NY		
20	196	39	143972	IN	500/1000	2000	1475.73	0	477670	FEMALE	High Scho	handlers-c	camping	own-child	57000	-27300	Multi-vehi Side Collis	Major Dar	Police	VA		
21	460	62	183430	IN	250/500	1000	1187.96	4000000	618845	MALE	JD	other-serv	bungie-jur	own-child	0	0	Multi-vehi Rear Collis	Minor Dar	Police	NY		
22	217	41	431876	IL	500/1000	2000	875.15	0	442479	FEMALE	Associate	machine-c	skydiving	own-child	46700	0	Multi-vehi Side Collis	Total Loss	Police	SC		
23	370	55	285496	IL	100/300	2000	972.18	0	443920	MALE	High Scho	prof-speci	paintball	other-rela	72700	-68200	Multi-vehi Rear Collis	Major Dar	Ambulanc	SC		
24	413	55	115399	IN	100/300	2000	1268.79	0	453148	MALE	MD	priv-hous	chess	own-child	-31000	0	Single Veh Front Collis	Total Loss	Ambulanc	WV		
25	237	40	736882	IN	100/300	1000	883.31	0	434733	MALE	College	craft-repa	kayaking	husband	-53500	0	Single Veh Rear Collis	Minor Dar	Other	VA		
26	8	35	699044	OH	100/300	2000	1266.92	0	613982	MALE	Masters	sales	polo	own-child	0	0	Multi-vehi Rear Collis	Major Dar	Other	OH		
27	257	43	867736	IN	100/300	2000	1322.1	0	436984	MALE	High Scho	prof-speci	only	own-child	-29200	0	Parked Ca ?	Minor Dar	Police	PA		

- 2) Data Preprocessing:
  - a) Drop the column not required for prediction.
  - b) In the dataset, there may be NULL values or '?' in the columns. These need to be replaced by NaN values.
  - c) Check for the categorical values and if present replace them with numeric values.
  - d) Merge the pre-existing numeric columns and converted numeric columns to the main dataset.
- 3) Clustering: KMeans algorithm is used to create clusters in the preprocessed data. The KMeans model is trained over this data, and the model is used for further prediction.
- 4) Model Selection: After the clusters have been created, we find the best model for each cluster by checking the accuracy given by different algorithms like K nearest neighbour, Logistic Regression, XGboost. In this, I have used "Random Forest" classifier for my model.



## ➤ How a graph database can fit into fraud detection

If we consider a fraud that we are taking into account, different subject and the relationship between them like how they are linked together can give us the valuable insights out of our data. For representing these subjects and their relationships, the best suited is graphs.

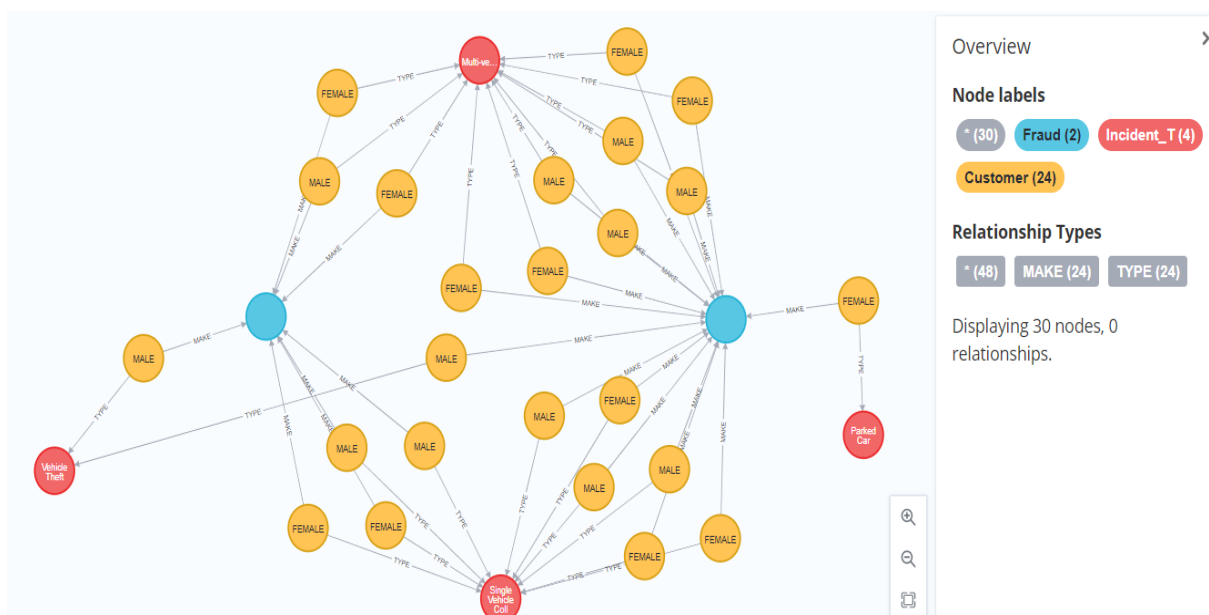
For building a fraud detection system that detects and prevents fraud as they happen, real-time traversal of a complex and highly interconnected dataset is essential.

Another important aspect to take in account is that traditional fraud systems looks for outliers(noise) but fraudsters try to act normal to avoid detection. Therefore, we need to detect fraudulent by linking different points and analysing normal behaviours.

## ➤ Why Neo4j?

Neo4j stores interconnected data that is neither purely linear nor purely hierarchical and creates graphs and gives us the relationship in our data by making different nodes instead of rows and columns , making it easier to detect links of fraudulent activity regardless of the size and shape of the data.

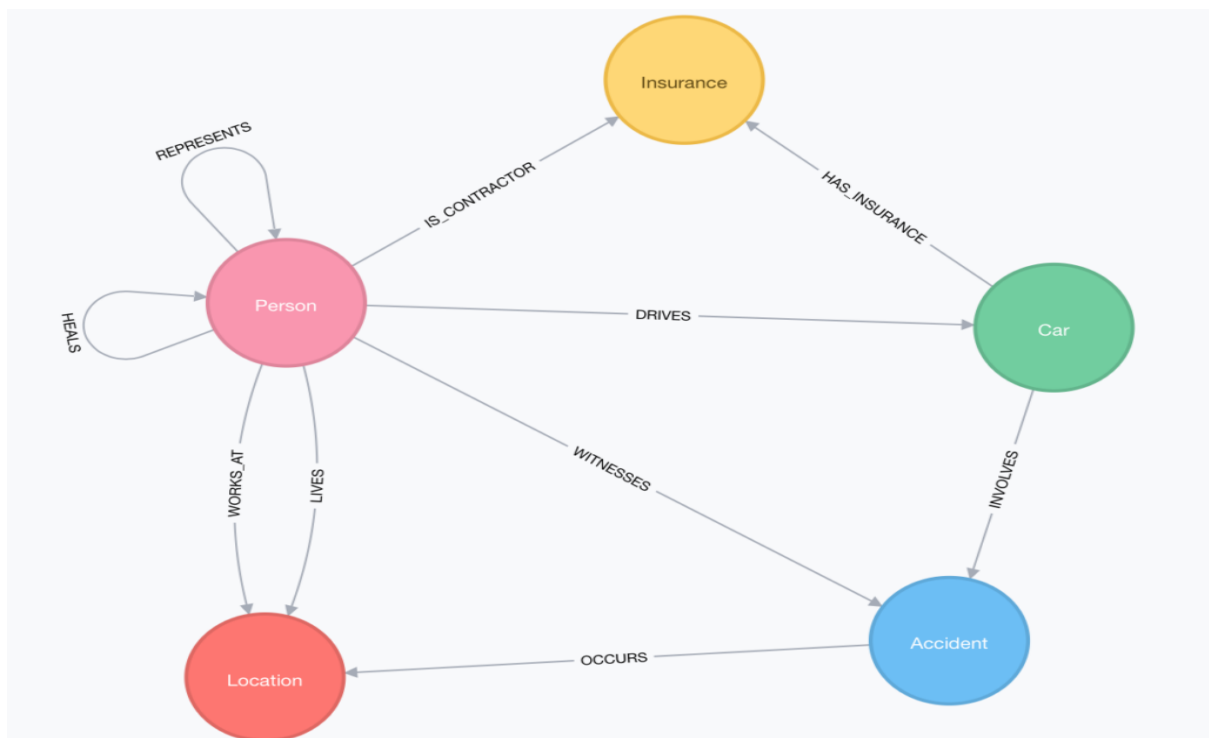
Neo4j's versatile property makes it easier for organizations to build fraud detection data models and rules, and apply it in the real world applications.





## ➤ Fraud insurance scenarios

Let's consider a simple model where we want to analyze two common scenarios: various people that are involved in different accidents but with different role for each accident and relationship between first aid location, where injured people live and also checks if there is any link of doctors , lawyers and witnesses with the person that met with the accident.



All of these subjects seem to act normal but what if two or more of them are involved in more than one accident cases with different roles. What if a doctor lives near a driver who gets first aid in a healthcare facility where that doctor works and also the healthcare facility is far from driver's home.

These scenarios are difficult to find through old models but are easy to understand through graph database. Hence, relationships become a key factor in identifying fraud rings or fraudsters.

## ➤ **Conclusion:**

Machine learning and Neo4j helped a lot to build this model in which queries are clear, concise and fast enabling real-time detection preventing fraud before they happen.

Modern fraud detection tools can improved by looking at the connections between different points and linking them together to obtain a output rather than looking at the individual point. The use of **native graph database** like Neo4j is the best solution to achieve this.

The completion of this project went quite well, I have learned many new things like about the Neo4j, graph database, Machine Learning, Python language and how hard work helps to build a good project. This project can be used to build a major project out of it which can help the society to tackle the fraud easily