Healthcare Fraud Detection

Devaki Kalyan Chandra Yadav Podila
Computer Engineering
University of California Riverside
Riverside, CA, US
pyada015@ucr.edu
Student Id: 862468795

Rohith Reddy Kancharakuntla
Computer Science
University of California Riverside
Riverside, CA, US
rkanc004@ucr.edu
Student Id: 862466402

Shubham Mishra
Computer Engineering
University of California Riverside
Riverside, CA, US
smish040@ucr.edu
Student Id: 862467767

Omkar Kadam
Computer Engineering
University of California Riverside
Riverside, CA, US
okada001@ucr.edu
Student Id: 862467471

Yash Kathe
Computer Engineering
University of California Riverside
Riverside, CA, US
ykath001@ucr.edu
Student Id: 862464930

1. ABSTRACT

The project addresses the challenges posed by escalating healthcare costs and an aging population in the United States by implementing a Healthcare Fraud Detection system using Big Data analytics. The key idea involves leveraging advanced data analytics techniques such as feature engineering, geo-demographic metrics, machine learning techniques to scrutinize a vast and intricate healthcare dataset. The main result aims to construct a comprehensive data model, develop a specialized machine learning model for healthcare fraud pattern recognition, and derive actionable insights for proactive fraud prevention. By exploring information related to prescriptions, healthcare providers, and payments, the project aims to identify fraudulent practices, control rising healthcare expenses, safeguard system integrity, and enhance the industry's fiscal responsibility, ultimately contributing to the well-being of beneficiaries.

2. INTRODUCTION

This project aims to combat a significant challenge of healthcare fraud responding to the increasing difficulties arising from the growing healthcare expenses and aging demographic in the United States. With healthcare spending reaching unprecedented levels, accounting for a substantial portion of the national income, the urgent need to cut costs has never been more critical. This project introduces an initiative leveraging Big Data analytics to develop a Healthcare Fraud Detection system by exploring information related to prescriptions, healthcare providers, and payments, the goal is to identify fraudulent practices. This approach aims to control the increase in

healthcare fraud to safeguard the system's integrity, and enhance the industry's fiscal responsibility, ensuring the well-being of beneficiaries.

This report includes a review of the literature that was done in order to solve the difficulties that encountered during the project's implementation and to obtain knowledge about a range of technical fields that were relevant to the project. Along with a thorough presentation of the code implementation throughout the various phases, such as data preparation, transformation, pre-processing and the core methodology outlining the strategies used to meet the project's goals, it offers comprehensive information about the dataset used. Results visualizations are also included in the paper. A summary of the project's assessment is also provided, highlighting how well it satisfies the requirements of working on a big data project. The final sections provide a thorough conclusion by talking about the project's results we achieved.

3. RELATED WORK

3.1 Framework for Fraud Detection in Health Insurance

A hybrid framework combining rule-based systems capturing domain expertise, supervised machine learning models like decision trees and neural networks, and unsupervised techniques like outlier analysis and clustering to detect anomalies in big medical insurance claims data. It processes outstanding claims by first clustering services and diseases, then classifying claims as fraudulent or not. A weighted priority queue ranks claims are considered for investigation. Case studies with one of the country's insurers showed 209% better fraud hit rates. The challenges with manual fraud evaluation needing medical knowledge and unable to scale to large insurance claim databases. Big data mining and machine learning models enable automatic detection of fraud patterns. Their proposed hybrid system aims to harness the advantages of both supervised classification and unsupervised anomaly detection to identify suspicious claims. Flagging these claims with relevant remarks also assists investigation teams. Significant fraud identification gains are achieved in the case studies.

3.2 Fraud detection in health insurance using data mining techniques

Application of big data analytics techniques like anomaly detection using support vector machines (SVM) and association rules mining to detect healthcare insurance fraud. Their techniques are evaluated on a real hospital transaction dataset to detect different types of anomalies in age, gender, service availing patterns etc. They proposed a framework that combines domain expertise rules, supervised machine learning (SVM), and unsupervised statistical anomaly detection (clustering, outlier detection) to identify fraudulent claims in big medical datasets. Their results demonstrate improved fraud detection hit rates compared to traditional methods. The problem of rising healthcare costs due to fraud, especially in developing countries where governments are initiating subsidies for underprivileged populations. Scalable and adaptive big data mining methods for effective fraud control are crucial for cost management and improved quality of healthcare delivery. The combination of techniques accounting for different fraud behaviors proves promising in the case study results.

3.3 Medicare Fraud Detection using Machine Learning Methods

Machine learning techniques like Bayesian classifiers, support vector machines, decision trees and neural networks for Medicare fraud detection in large medical claims datasets. Their methods are evaluated on a Medicare dataset with fraud labels from a national excluded individuals database. A comparative analysis handles class imbalance using oversampling and under-sampling. Different specialties and performance metrics are assessed. Their under-sampling method performs better across techniques, with supervised methods significantly outperforming unsupervised methods. They argue that rule-based healthcare fraud systems have difficulty given the complexity and scale of modern big health insurance programs like Medicare. Machine learning provides

automatic pattern recognition from medical claims data itself. Both supervised learning exploiting known fraud examples, and unsupervised anomaly detection are tested. The comparative assessment provides guidance on matching techniques, sampling methods and performance evaluation for effective data-driven Medicare fraud control.

3.4 A Framework For Fraud Detection in Government Supported National Healthcare Programs

A modular framework leveraging big data analytics and machine learning pipelines for detecting fraud in government funded national healthcare programs with large claims datasets. It generates time series traces for transactions to identify invalid procedures and claims based on anomalies in age, gender, service providing patterns etc. Their framework combines encoded domain expertise rules, supervised learning models like averaged perceptron classifiers and decision trees, unsupervised statistical outlier detection and clustering algorithms that scale to big medical data. Improved accuracy is demonstrated on real hospital data. They argue rule-based systems relying on human experts have difficulty scaling, detecting new fraud types and keeping updated regulations when applied to national healthcare datasets. An adaptive hybrid big data system blending rule-based, supervised machine learning and unsupervised techniques can enable automated fraud control on industry-scale medical claims data. Temporal pattern analysis with time series data is also an effective signal for suspicious activities. Their study results showcase detection improvements.

3.5 Identifying Medicare Provider Fraud with Unsupervised Machine Learning

Empirically evaluates several unsupervised outlier detection methods to identify Medicare fraud using Part B claims data labeled with exclusions from a General database. Along with established techniques like local outlier factor (LOF) and auto-encoders, newer approaches such as isolation forest and unsupervised random forest were also tested on the big medical claims dataset. Their performance is measured using ROC analysis and area under the curve (AUC). Results found LOF has highest AUC and detection capability while isolation forest has high sensitivity but low specificity. Their public Medicare data offers new fraud analysis opportunities but lacks direct fraud indicators. Outlier detection provides anomaly signals for suspicious providers without fraud examples. Comparing multiple techniques on Medicare data has been limited. Their empirical tests integrate exclusion database labels for validation. Performance variations showcase challenges recognizing Medicare fraud patterns in big Part B claims data. LOF has best discrimination capability, while isolation forest merits further tuning given high fraud sensitivity.

4. MAIN CONTENTS

Implementation of the Healthcare Fraud Detection Framework involves several key stages, each contributing to the overall success of the project. The following tasks outline the sequential steps required to bring the framework to fruition.

1. Data is collected from [1], [2], [3]. Then, it is required to perform some pre-processing.
2. For achieving the query analysis-based objectives, SparkSQL will be used.
3. MLlib will be used to develop the predictive models.
4. Evaluation of Performance of the predictive machine learning models based on different metrics.

4.1 Data Collection and Preparation

4.1.1 CMS Data

- "Medicare Provider Utilization and Payment Data: Physician and Other Supplier" and "Medicare Provider Utilization and Payment Data: Part D Prescriber (Part D)" are publically available CMS datasets.
- Each row is structured with information describing a provider, primarily identified by the national provider identifier (NPI). Additional features include provider details like name, location (state and city) and demographics.
- The dataset includes information on drug prescriptions submitted to Medicare, including drug names, provider types, and aggregate statistics such as the total number of unique Medicare Part D beneficiaries and aggregate drug costs.

4.1.2 LEIE Data

- The List of Excluded Individuals and Entities (LEIE) data is updated monthly by the Office of Inspector General (OIG) and is also in CSV format.
- It contains information about healthcare providers prohibited from submitting claims to Medicare due to previous violations of Medicare rules. Important attributes include NPI and exclusion type.

4.2 Data Preprocessing

4.2.1 Labeling the Part D Dataset:

- The Part D dataset initially lacks labels. Labels are derived by merging the Part D dataset with the LEIE dataset using a left join on the NPI attribute.
- If an NPI from the Part D dataset is found in the LEIE dataset, the corresponding Part D records are labeled as fraudulent.
- We consider any records with NaN values after the merge as non-fraudulent.

4.2.2 Handling Missing Values:

- After merging, records with NaN values are treated as non-fraudulent, assuming that absence in the LEIE dataset indicates non-exclusion.

4.2.3 Feature Selection:

- The features used in the experiment are divided into numerical and categorical types, and it includes NPI, provider details, drug information, and claim related aggregate statistics.

4.2.4 Data Format:

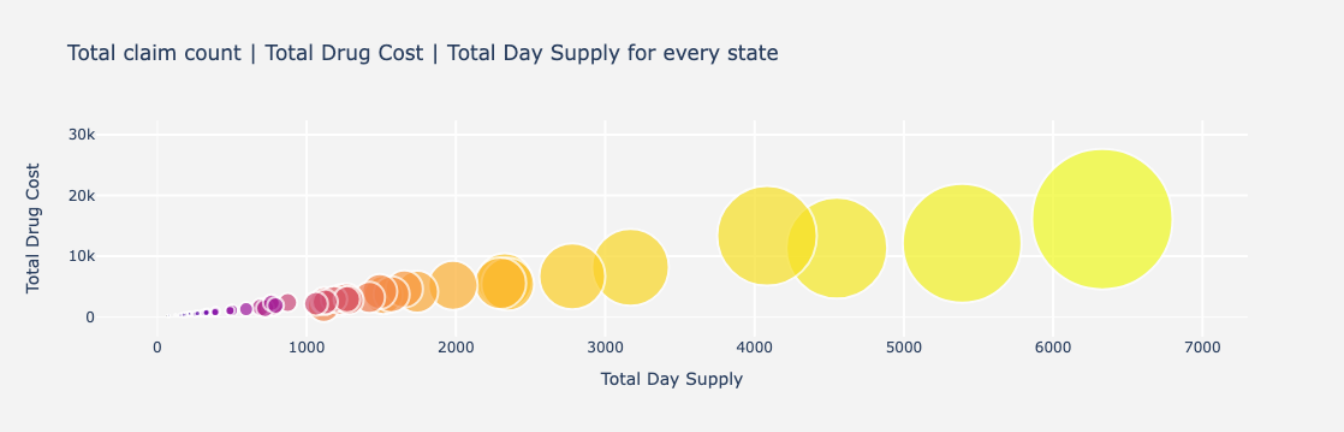
- We then ensure that the data is appropriately formatted for SparkSQL processing, considering the requirements and compatibility of the tools being used.

This sets the stage for further analysis using SparkSQL and MLlib for query analysis-based objectives and predictive modeling. The labeled dataset will be crucial for training and evaluating predictive models for fraud detection.

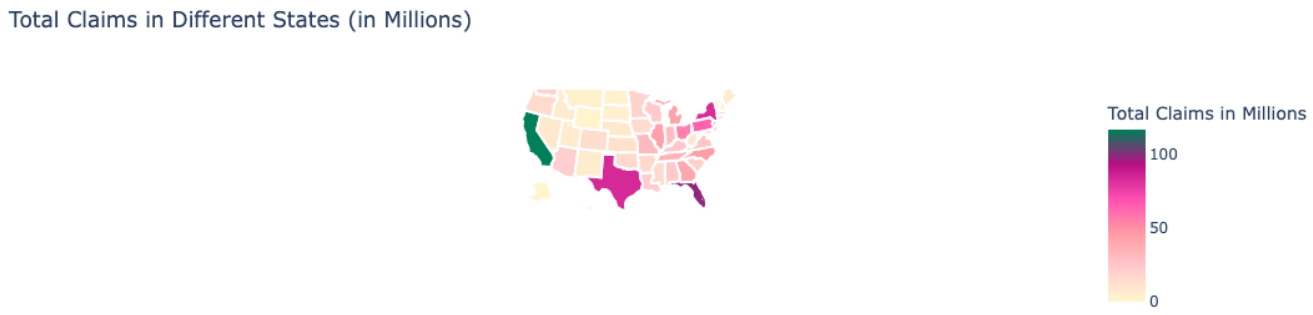
4.3 Choosing the Machine Learning Algorithm and Training Our Model

The implementation will incorporate the machine learning algorithms for training the model such as Gaussian Naive Bayes, Logistic Regression, Random Forest, Decision Trees. The choice of these algorithms aligns with the problem's approach, nature of the dataset and the complexities inherent in identifying fraudulent activities. The subsequent steps involve comprehensive evaluation using metrics such as accuracy, precision, recall, F1-score, and AUC score.

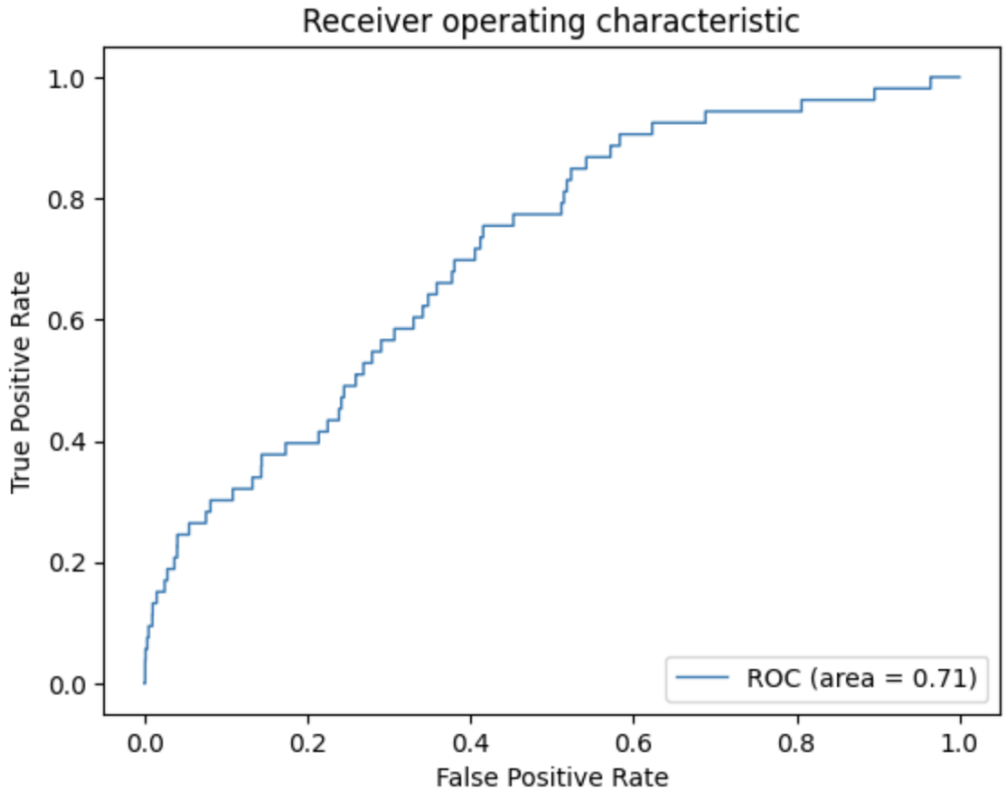
5. EVALUATION



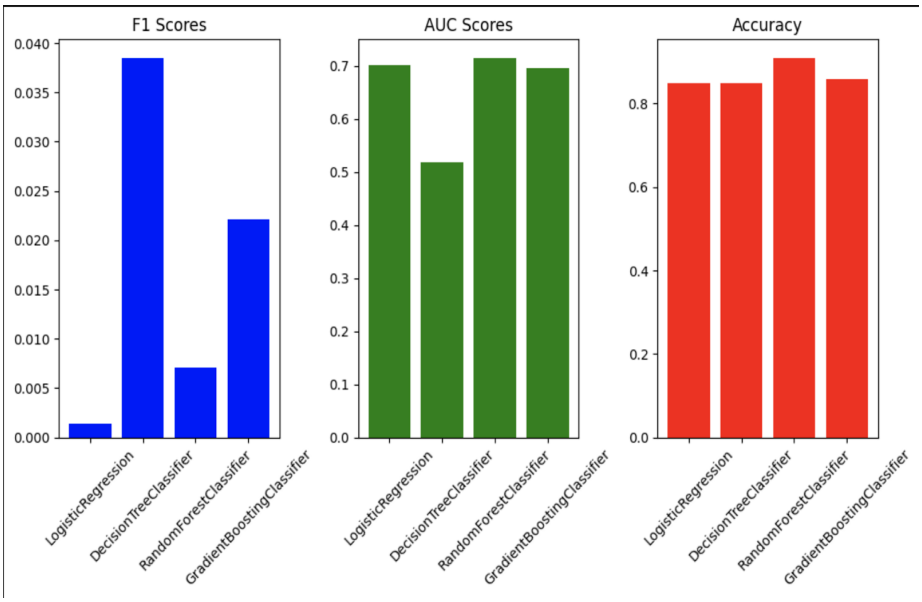
The above figure highlights the relationship between total day supply and total drug cost for different states. The size of the bubbles represent the total claim count. This visualization helps in understanding how drug supply and costs vary across states, with larger bubbles indicating a higher number of claims.



The Map of the United States showing the total number of claims in different states, measured in millions. The color gradient represents the scale of total claims, with darker colors indicating a higher number of claims.



The above image represents Receiver Operating Characteristic (ROC) curve with an area under the curve (AUC) of 0.71. The ROC curve plots the true positive rate against the false positive rate at various threshold settings. An AUC of 0.71 indicates a good level of model performance in distinguishing between the positive and negative classes. Generally, an AUC closer to 1 signifies a better-performing model, while an AUC closer to 0.5 suggests no discriminative power.



Comparative Analysis of four ML models against F1, AUC and accuracy where Logistic Regression emerges with the smallest F1 score, whereas the Random Forest Classifier leads with the highest, indicating a superior balance between precision and recall. The AUC Scores are relatively uniform across the models, suggesting that each has a comparable ability to differentiate between classes. When considering Accuracy, the Random Forest Classifier

performs better than others, marking it as the most precise in this evaluation.

6. CONCLUSION

We have successfully identified fraudulent activities within the healthcare sector by analyzing three different datasets i.e. Part-D, payment, and excluded individual records. We have performed data transformation, feature extraction and engineering processes and applied machine learning classification models on top of it and successfully demonstrated different evaluation metrics such as accuracy, f1 and roc curve

7. AUTHOR CONTRIBUTIONS

Author	**Contributions**
Devaki Kalyan Chandra Yadav Podila	● Finding the dataset ● Literature Survey ● Report Writing ● Data pre-processing ● Understanding the part-d prescriber by providerad drug dataset ● Understanding payment dataset ● Understanding list of excluded individual dataset ● Transforming datatype of columns ● Joining Datasets on NPI ● Training and Testing ML Models ● Data visualization
Rohith Reddy Kancharakuntla	● Finding the dataset ● Literature Survey ● Report Writing ● Data pre-processing ● Understanding the part-d prescriber by providerad drug dataset ● Understanding payment dataset ● Understanding list of excluded individual dataset ● Transforming datatype of columns ● Joining Datasets on NPI ● Training and Testing ML Models ● Data visualization
Shubham Mishra	● Finding the dataset ● Literature Survey ● Report Writing ● Data pre-processing ● Understanding the part-d prescriber by providerad drug dataset ● Understanding payment dataset ● Understanding list of excluded individual dataset ● Transforming datatype of columns ● Joining Datasets on NPI ● Training and Testing ML Models ● Data visualization

Omkar Kadam	• Finding the dataset • Literature Survey • Report Writing • Data pre-processing • Understanding the part-d prescriber by providerad drug dataset • Understanding payment dataset • Understanding list of excluded individual dataset • Transforming datatype of columns • Joining Datasets on NPI • Training and Testing ML Models • Data visualization
Yash Kathe	• Finding the dataset • Literature Survey • Report Writing • Data pre-processing • Understanding the part-d prescriber by providerad drug dataset • Understanding payment dataset • Understanding list of excluded individual dataset • Transforming datatype of columns • Joining Datasets on NPI • Training and Testing ML Models • Data visualization

8. REFERENCES

[1] "Exclusions List," U.S. Department of Health & Human Services, Office of Inspector General. [Online]. Available: https://oig.hhs.gov/exclusions/exclusions_list.asp.[Accessed: 9 October, 2023]

[2] "Drugs@FDA: FDA Approved Drug Products," U.S. Food and Drug Administration. [Online]. Available: https://www.fda.gov/Drugs/InformationOnDrugs/ucm079750.htm#collapseOne .[Accessed: 9 October , 2023]

[3] "Medicare Provider Charge Data - Part D Prescriber Data, "Centers for Medicare & Medicaid Services. [Online]. Available: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Part-D-Prescriber.html .[Accessed: 9 October, 2023]

[4] "Dataset Downloads," Centers for Medicare & Medicaid Services. [Online]. Available: https://www.cms.gov/OpenPayments/Explore-the-Data/Dataset-Downloads..[Accessed: 16 October , 2023].

[5] R. Bauder, R. da Rosa and T. Khoshgoftaar, "Identifying Medicare Provider Fraud with Unsupervised Machine Learning," 2018 IEEE International Conference on Information Reuse and Integration (IRI), Salt Lake City, UT, USA, 2018, pp. 285-292, doi: 10.1109/IRI.2018.00051.

[6] R. A. Bauder and T. M. Khoshgoftaar, "Medicare Fraud Detection Using Machine Learning Methods," 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), Cancun, Mexico, 2017, pp. 858-865, doi: 10.1109/ICMLA.2017.00-48.

[7] N. Rayan, "Framework for Analysis and Detection of Fraud in Health Insurance," 2019 IEEE 6th International

Conference on Cloud Computing and Intelligence Systems (CCIS), Singapore, 2019, pp. 47-56, doi: 10.1109/CCIS48116.2019.9073700.
[8] I. Matloob and S. Khan, "A Framework for Fraud Detection in Government Supported National Healthcare Programs," 2019 11th International Conference on Electronics, Computers and Artificial Intelligence (ECAI), Pitesti, Romania, 2019, pp. 1-7, doi: 10.1109/ECAI46879.2019.9042126.
[9] V. Rawte and G. Anuradha, "Fraud detection in health insurance using data mining techniques," 2015 International Conference on Communication, Information & Computing Technology (ICCICT), Mumbai, India, 2015, pp. 1-5, doi: 10.1109/ICCICT.2015.7045689.
[10] Herland, M., Bauder, R. A., & Khoshgoftaar, T. M. (2020). Approaches for identifying U.S. medicare fraud in provider claims data. Health care management science, 23(1), 2–19. https://doi.org/10.1007/s10729-018-9460-8
[11] Herland, M., Bauder, R.A. & Khoshgoftaar, T.M. The effects of class rarity on the evaluation of supervised healthcare fraud detection models. J Big Data 6, 21 (2019). https://doi.org/10.1186/s40537-019-0181-8
[12] Batko, K., Ślęzak, A. The use of Big Data Analytics in healthcare. J Big Data 9, 3 (2022). https://doi.org/10.1186/s40537-021-00553-4
[13] Karthika, I., and K. P. Porkodi. "Fraud Claim Detection Using Spark." International Journal of Innovations in Engineering Research and Technology, vol. 4, no. 2, 2017, pp. 1-4
[14] M. Dash, H. Liu, Feature selection for classification, Intelligent Data Analysis, Volume 1, Issues 1–4, 1997, Pages 131-156, ISSN 1088-467X, https://doi.org/10.1016/S1088-467X(97)00008-5
[15] Ortega, Pedro & Figueroa, Cristián & Ruz, Gonzalo. (2006). A Medical Claim Fraud/Abuse Detection System based on Data Mining: A Case Study in Chile. DMIN. 6. 224-231.
[16] Kumari, A., Punn, N.S., Sonbhadra, S.K., Agarwal, S. (2023). Impact of the Composition of Feature Extraction and Class Sampling in Medicare Fraud Detection. In: Tanveer, M., Agarwal, S., Ozawa, S., Ekbal, A., Jatowt, A. (eds) Neural Information Processing. ICONIP 2022. Lecture Notes in Computer Science, vol 13625. Springer, Cham. https://doi.org/10.1007/978-3-031-30111-7_54
[17] R. Bauder and T. Khoshgoftaar, "Medicare Fraud Detection Using Random Forest with Class Imbalanced Big Data," 2018 IEEE International Conference on Information Reuse and Integration (IRI), Salt Lake City, UT, USA, 2018, pp. 80-87, doi: 10.1109/IRI.2018.00019.
[18] Karca Duru Aral, Halil Altay Güvenir, İhsan Sabuncuoğlu, Ahmet Ruchan Akar, A prescription fraud detection model, Computer Methods and Programs in Biomedicine,
Volume 106, Issue 1, 2012, Pages 37-46, ISSN 0169-2607, https://doi.org/10.1016/j.cmpb.2011.09.003.
[19] Kumaraswamy N, Markey MK, Ekin T, Barner JC, Rascati K. Healthcare Fraud Data Mining Methods: A Look Back and Look Ahead. Perspect Health Inf Manag. 2022 Jan 1;19(1):1i. PMID:
35440932; PMCID: PMC9013219.
[20] Herland, M., Khoshgoftaar, T.M. & Bauder, R.A. Big Data fraud detection using multiple medicare data sources. J Big Data 5, 29 (2018). https://doi.org/10.1186/s40537-018-0138-3
[21] Li J, Huang KY, Jin J, Shi J. A survey on statistical methods for health care fraud detection. Health Care Manag Sci. 2008 Sep;11(3):275-87. doi: 10.1007/s10729-007-9045-4. PMID: 18826005.
[22] Apache Software Foundation. (2023, October 15). "Spark Python API Documentation." Apache Spark. [Online]. Available: https://spark.apache.org/docs/latest/api/python/index.html