|  |  |
| --- | --- |
| FINAL PROJECT | TOPIC  Healthcare Fraud Detection using Data Analytics  Bala Durga Sridevi Kesavarapu |

# B) Executive Summary

This paper will cover the use of business analytics in the healthcare industry with an emphasis on the security and privacy issues. While our proposal focused on utilizing predictive analytics to detect fraud initiated by the insured individual, throughout our studies we found that the overwhelming amount of committed fraud is a result of healthcare providers submitting fraudulent claims. Identity theft was also highlighted in the proposal; however, during our research this was not identified as a primary reason detected currently by the medical insurance providers. The actual observation discovered for medical insurance claims fraud will be further explored in the paper.

After analyzing the background and state of the industry, a conceptual model was then constructed with interesting observations captured and documented. This proof-of-concept focused on identifying a pattern to detect potentially fraudulent claims and use those rules to flag the claim for further review. For example, a provider making a claim for a procedure that is outside of his or her area of expertise may be further scrutinized as this behavior would be considered atypical and in violation of one of the rules.

# C) Introduction

Digital healthcare is now an emerging trend. Hospitals and other healthcare entities are providing internet services that give patients new ways to manage their health, wellness, and medical affairs online. These types of services are convenient for the patients, but they also lead to a significantly increased threat of identity fraud. Fraudulent claims continue to be a significant issue, costing policyholders billions of dollars. More companies are turning to analytics to help identify claim fraud. Identifying claim fraud using predictive analytics, however, represents a unique challenge.

For big-data initiatives to succeed, the healthcare system must undergo some fundamental changes. For instance, the old levers for capturing value, such as unit-price discounts based on contracting and negotiating leverage, do not take full advantage of the insights that big data provides and thus need to be supplemented or replaced with other measures. Stakeholders across the industry also need to protect patient privacy as more information becomes public and ensure that safeguards are in place to protect organizations that release information.

Privacy is viewed as a key governing principle of the patient-physician relationship. Patients are required to share information with their physicians to facilitate correct diagnosis and treatment, and to avoid adverse drug interactions. However, patients may refuse to divulge valuable information in cases of health problems such as psychiatric behavior and HIV, as their disclosure may lead to social stigma and discrimination. Over time, a patient’s medical record accumulates significant personal information including identification, history of medical diagnosis, digital renderings of medical images, treatments, medication history, dietary habits, sexual preference, genetic information, psychological profiles, employment history, income and physicians’ subjective assessments of personality and mental state.

Millions of Americans are insured in whole or in part by health insurance that is paid for with federal or state government funds although many rely on privately funded health insurance plans. Consumers sometimes do not recognize the prominence of protecting their digital health data. In this information age, individuals must take care of the sensitive data and should be mindful of the dangers lurking in cyberspace, particularly identity theft.

According to the Identity Theft Resource Center, 2.3 million theft cases are reported in 2014 which is a 22% increase from the year before. The leading cause of these incidents appear to be the data breaches.

Medical theft can be in any one of the following ways:

* When someone steals your personal information to receive the benefits such as free medical care, goods and prescription drugs.
* The treatment, health insurance, payment records and credit report may be affected if a thief’s health information is mixed with your medical records. This can disrupt a person’s life, damage their credit rating.
* The thief can make fraudulent claims against your own health insurance policy or Medicare to receive free healthcare.
* The thief can obtain illegal, free or bogus treatment by assuming your identity to begin their own medical care free. You can be billed for fake treatment claims.
* The thief secures the prescriptions such as addictive drugs and medical products which they can use or sell.

Additionally, medical providers and legitimate health insurance provider customers may also commit fraud in the following ways outside of identity theft:

* The patient relies on tenureship as a health insurance provider client to circumvent new customer checks and commit fraud claims
* The medical service provider upsells services provided to the client in a medical claim, indicating a fraudulent claim.
* The medical service provider has a relationship with the patient and collaborates to commit claims fraud through overprescription or billing for non-provided services.

# D) Background

The National Association of Insurance Commissioners defines insurance fraud as, “when an insurance company, agent, adjuster or consumer commits a deliberate deception in order to obtain an illegitimate gain. (D1)”

According to FBI estimates, between 3 and 10 percent of healthcare expenses were fraudulent in 2010 which equated from $77 billion to $259 billion. Insurance applicants, policyholders, and service providers all contributed to this fraud via charging for extra services not performed, lying on applications, charging for equipment not used, and even faking accidents(D2).

Another type of medical fraud on the rise is health identity theft in which victims’ identities and insurance information is stolen and then used to make false claims. In 2014, over 2.3 million Americans had their medical identity stolen. This information is often stolen by medical personnel providing services and resold on the black market(D3).

The amount of data generated by the millions of healthcare claims each year generate enormous amounts of data. As of 2013, EMC estimated the volume of healthcare data to be 153 exabytes and is expected to grow to 2.3 zettabytes by 2020((D4). Improved data analytics processes and the use of artificial intelligence are required to seek out the fraud needles in this haystack of data.

*Detection - Current Analytic Controls and Techniques*

Some of the common ways data analytics can be used to detect and prevent fraud, as identified by ACL Services, Ltd. are(D5):

* Find kickbacks paid in exchange for referring business.
* Identify charges posted outside of proper GL period.
* Highlight “upcoding” of procedures: Statistically outlying numbers.
* Match vendor names/addresses/tax IDs to payroll records for employees.
* Summarize large invoices without purchase orders, by amount, vendor, etc.
* Compare list of valid signed-up employees to list of people actually receiving health benefits from insurance company.
* Highlight billing for medically unnecessary tests.
* Identify false/invalid/duplicate Social Security numbers.
* Highlight excessive use of high risk DRGs (“Diagnosis-Related Groups”)
* Identify excessive billing by a single physician
* Identify employee overtime abuses.
* Report entries against authorization records for new or terminated employees.
* Identify multiple payroll deposits to the same bank account.

Analytic techniques useful to ACL in detecting fraud are(D5):

* Calculation of statistical parameters (e.g., averages, standard deviations, high/low values) – to identify outliers that could indicate fraud.
* Classification – to find patterns amongst data elements.
* Stratification of numbers – to identify unusual (i.e., excessively high or low) entries.
* Digital analysis using Benford’s Law – to identify unexpected occurrences of digits in naturally occurring data sets.
* Joining different diverse sources – to identify matching values (such as names, addresses, and account numbers) where they shouldn’t exist.
* Duplicate testing – to identify duplicate transactions such as payments, claims, or expense report items.
* Gap testing – to identify missing values in sequential data where there should be none.
* Summing of numeric values – to identify control totals that may have been falsified.
* Validating entry dates – to identify suspicious or inappropriate times for postings or data entry.

*Early Detection Means an Early Cure*

One of the largest issues with the healthcare system is that payment to providers is based on an “honor system.” This payment system is complicated by a large number of transactions as well as laws requiring prompt payments to healthcare providers (6). Claims are often paid before they can be reviewed (5), meaning that a predictive model that accurately detects fraud prior to payment can be used to reduce the number and overall cost of fraudulent claims.

Most healthcare claims are first sent to a clearinghouse with is a third-party entity that acts as a hub between healthcare providers and insurers. These clearinghouses receive the claims, clean up the data and fix errors, and then format them into industry standard formats before sending the claims on the insurers. Whereas individual insurers are only able to access regarding their claimants, the clearinghouses view of data from multiple claims, providers, and insurers make them a prime location for the implementation of fraud detection analytic tools(10).

*Collaboration*

Two of The National HealthCare Anti-Fraud Association’s (NHCAA) principles for policymakers focus on data sharing and collaboration between private insurers and government programs(D7):

1. The sharing of anti-fraud information between private insurers and government programs should be encouraged and enhanced.
2. Data consolidation and real-time data analysis must be at the forefront of health care fraud detection and prevention.

Most healthcare providers bill multiple private and government payers such as Medicare and private insurance companies; but when analyzing claims for fraud, these payers are unable to cross-reference their data with other payers in order to identify fraudulent activity. NHCAA asserts that an aggregated healthcare claims database whose content is accessible to all payers would allow data analysis tools to consider all related claim information as well as previous healthcare data that would lead to better identification of fraud patterns(D8).

*Use of AI*

Current fraud detection techniques are rule-based models that perform the same checks if-then-else checks on each transaction. Analytics companies such as AYASDI are branching out and using machine intelligence to develop algorithms to not only identify fraud but to train the AI to detect fraud on its own.

AI may also be used to provide confirmation as to the accuracy of existing models and even to improve those models. “Rules-based checks are typically linear and cannot recognize the increasing sophistication of many fraud schemes. Models created from standard machine learning algorithms tend to over-fit as they attempt to describe all of the underlying data (4).”

# E,F,G) Conceptual, Discussion, Results

In order to develop a proof of concept for fraud detection and; therefore, reduce the impact of fraudulent claims through an honor system; the group generated a theoretical dataset (see [reference](https://docs.google.com/document/d/1liq6kJGcOXKNokS6HUhwbGrUpPXvSMDeBBXWFeGEKKo/edit#bookmark=id.1j5w80ia61cb)). The shorthand descriptions are also captured for the dataset in the references (see [reference](https://docs.google.com/document/d/1liq6kJGcOXKNokS6HUhwbGrUpPXvSMDeBBXWFeGEKKo/edit#bookmark=id.3a6xm0ah4n07)). Unfortunately, while a significant amount of time was spent researching and attempting to find a real-world dataset to use for building and training a predictive model, a comprehensive set of attributes needed to target fraudulent claims data did not exist without a paywall to key attributes for analytics mining. Additionally, the data for healthcare and healthcare providers is constrained by HIPAA; therefore, any data sets found online with real data are neutered by the removal of important attributes useful for analytics. For example, <https://data.cms.gov/> and <https://catalog.data.gov/dataset> have datasets related to plans and related benefits along with some claim submission data (number of claims per region), but not much more information which can be used for a predictive model.

As a result, our training dataset was generated using a random number generated between 0 and 1, with 25% of the records flagged as fraudulent and 75% of the recorded information flagged as valid claims. The rest of the data was generated from an auto-claims dataset, with the direct replacement of auto-claims classifications with medical claims classifications typically submitted to providers.

In generating a model for determining if a claim should be flagged as potential fraud for analysis by a claims expert, the dataset was first imported into SAS Enterprise Miner with the roles outlined in the [appendix](https://docs.google.com/document/d/1liq6kJGcOXKNokS6HUhwbGrUpPXvSMDeBBXWFeGEKKo/edit#bookmark=kix.k5bzle3xwfju) below. Furthermore, any nominal variable with more than 10 classifications was marked as rejected, due to the complexities introduced in modeling decision trees with many branches for classifying a medical claim as fraud. Additionally, identification attributes unique to the row were also rejected, as these are unique identifiers to the claim and not necessarily useful in training a model to predict fraudulent claims. These rejected variables include the following:

* incident\_city
* insured\_hobbies
* policy\_number

Next, a StatExplore node was added to further explore the dataset utilized for this proof of concept. The initial standout from the generated statistics is the order of importance for the variables in generating a model; in other words, the strong correlation between a variable value and the fraud indicator value. In order of most important to least important, these included:

* incident\_severity
* claim\_amount
* negotiated\_amount
* care\_type
* policy\_annual\_premium
* insured\_occupation
* incident\_state
* months\_as\_customer
* drug\_prescribed
* age
* umbrella\_limit
* insured\_relationship
* num\_prescriptions
* copay
* insured\_education\_level
* policy\_deductable
* policy\_state
* insured\_sex
* preexisting\_phsyician\_relationship
* new\_prescription

Additional insights gleaned from the StatExplore node include that the mean claim age is 38, with 9.1 years as the standard deviation. Furthermore, the average age of a customer account submitting is 203 months, which works out to 16.9 years. In other words, fraudulent claims are not limited to a potential collaboration between a patient’s doctor and a new patient on an insurance plan, but if a collaborative effort, then the customers are long term customers. Finally, with a focus on the type of care provided, both urgent care and general practice have roughly the same number of fraudulent claims, resulting in a reduced worth on the overall model.

After analyzing other variables in StatExplore, a replacement node was then added with a connection to the SAS table node. The replacement node used the default standard deviation replacement method to trim interval variable values not conforming to the bell curve except for a user defined interval override for claim\_amount, negotiated\_amount, and policy\_annual\_premium. The user overrides include the following:

* The claim\_amount lower limit was set to 16787.11 as primary concern in detection is for higher claim amounts and not necessarily “petty fraud”. Any claims lower than 1000 are considered for the report purpose as routine visits.
* The negotiated\_amount replacement lower limit set to 919.89 for the same reasons as claim\_amount. More attention was given to trimming the negotiated amount; however, as it is the more concerning amount since this is the amount paid to the provider and the potential motivation for a fraud payout.
* The policy\_annual\_premium lower limit set to 578.55 as anything lower considered bad data. Simply put, healthcare in the US even with a high deductible health plan would never be sold that cheap by any insurance provider.
* In regard to class variable replacements for nominal values, no transformation or replacement was applied.

After using a replacement node to trim off any interval outliers and bad data, an impute node was appended. This node handled imputing values for the preexisting\_physician\_relationship attribute which had 343 missing values,. That particular attribute suffered a data input integrity issue, indicating that it should further be reviewed by an application team in conjunction with the business analysts to determine a process for ensuring a proper data enrichment strategy. Other attributes such as insured\_hobbies would have benefited from imputing missing values, but that attribute was marked as rejected as it had too many nominal values without imputing to effectively utilize the data for the purpose of this paper.

Once the missing data was imputed to fill in the data density gaps, then the data was partitioned using a data partition node for a default allocation of 40% training, 30% validation, 30% test and connected to an auto-pruning decision tree. For a simple, easily understood model, the decision tree works perfectly to develop a predictive rule based model for flagging claims as potential fraud. Since, in our experience, results of interactive, manual pruning often results in a less efficient model, the generated tree was left with default conditions except for the following:

* Method was set to Largest to generate the most-accurate, full tree. As claims processing is generally not a speedy process, there is sufficient time to analyze for accuracy using the largest possible tree for accuracy
* Assessment Measure was set to misclassification in order to reduce the number of false positives for our boolean target.

As seen in [Appendix 2](https://docs.google.com/document/d/1liq6kJGcOXKNokS6HUhwbGrUpPXvSMDeBBXWFeGEKKo/edit#bookmark=kix.s3hu2gp1u11h), the incident severity of any non-high nominal value has a high likelihood of not being a fraudulent claim.  As we are interrogating the largest tree, for brevity the state context was also considered and for any state other than North Carolina there was a strong likelihood of non-fraudulent activity for non-high severity incidents. From the tree generated, it is also feasible to generate rules for auto-review by a claims adjuster. If the incident severity is high, that should require review; however, if in addition a new prescription is prescribed, then the claim should definitely require manual review and approval by a claims adjuster.

Using Enterprise Miner or another predictive analytics tool to analyze a large claims dataset provides easy, functional capabilities to generate models. As decision trees are easily understood, the focus of the POC utilized a decision tree model. For future refinement, the recommendation would be to analyze the data at least quarterly by an actuary working in-parallel with a business analyst to determine if the ruleset needs further refinement.

Alternatively, without a dataset of claims already known as fraudulent or without a definitive target variable, then a cluster analysis exploration would be used to generate classification group driven by rules. For each of the classification groups by demographics, the claims data would then be analyzed to find a risk factor associated with fraud. If any demographic clusters tend to file more fraudulent claims, then that demographic group, set of physician offices, or insured individuals would have claims scrutinized to avoid monetary losses paying out on invalid requests. However, given that there is already a target variable defined in the sample dataset converted from auto claims to healthcare claims, the need to partition out the data into clusters was not prevalent.

From the test dataset, a few results were observed. First, with the decision tree method as assessment, the misclassification rate is at [20](https://docs.google.com/document/d/1liq6kJGcOXKNokS6HUhwbGrUpPXvSMDeBBXWFeGEKKo/edit#bookmark=id.g1xkqenjlc48)% during validation. Furthermore, while the node rules are not as precise in this dataset for predicting definitive fraud, it is possible to predict with 92% certainty non-fraud if the incident occurs in SC, NY, WV, VA, PA and the severity of the incident low, urgent, or trivial for the 250 observations. For predicting fraud, 71% of the new prescriptions with high severity will be classified as fraudulent; while only 28% of the claims will be fraudulent for NC claims of low, urgent, or trivial severity.

# H) Conclusion

Given the various types of fraud covered, from identity theft based claims to invalid physician submitted claims, the industry has a wide range of avenues to utilize predictive analytics for fraud detection. As the conceptual approach described, even a simple decision tree can be utilized to handle exception based scenarios requiring further review and minimizing the scrutiny needed to analyze the details of every claim. Furthermore, if such a process was implemented within a clearinghouse with access to aggregated information across multiple insurance providers, then an extremely accurate model can be developed using all of that data available. With the minimum impact (3%) of fraudulent claims estimated at $77 billion and a rough maximum impact (10%) at $259 billion, all insurance providers need to mitigate this damage through an applied predictive analytics model.

***Dataset***

*CSV File*

[*https://drive.google.com/open?id=1hGdKmZkv2pGczDcX0Jy8q7uoCkYwNPHi*](https://drive.google.com/open?id=1hGdKmZkv2pGczDcX0Jy8q7uoCkYwNPHi)

*SAS Table*

<https://drive.google.com/open?id=1ECn1QB9pKHV1COtEEHhwba_rNPrgZsOP>

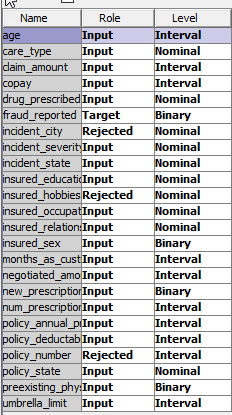
*EM File*

<https://drive.google.com/open?id=1MU15SySI8dBUTJ-NgOK_k8voYxhrRgok>

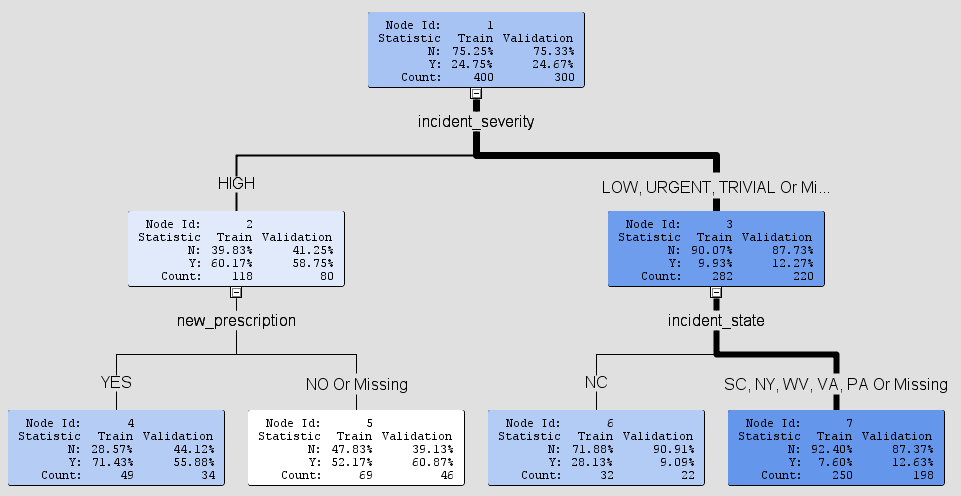
Data Label



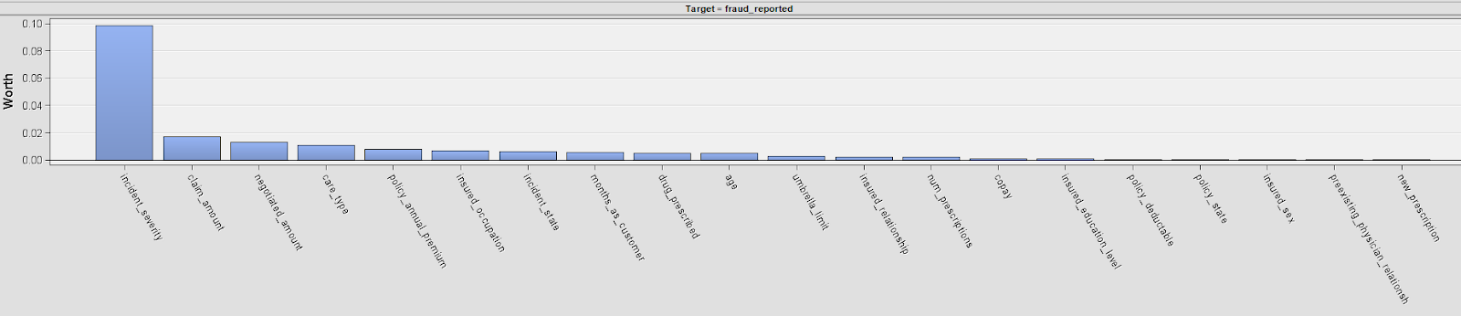
**Appendix 1**



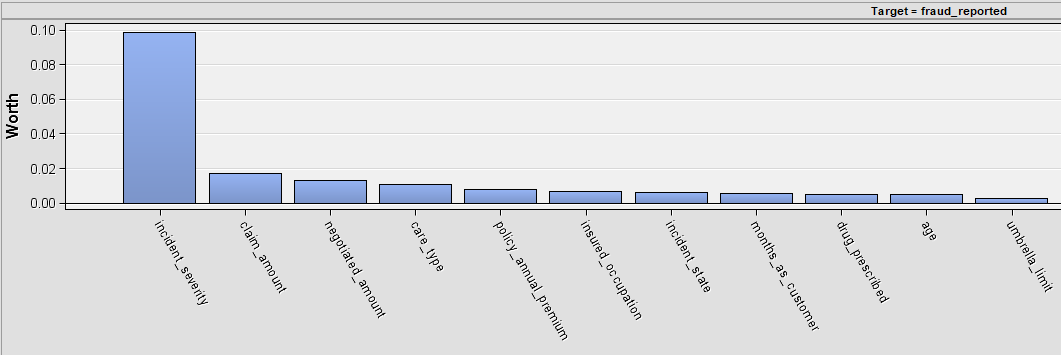
**Appendix 2**



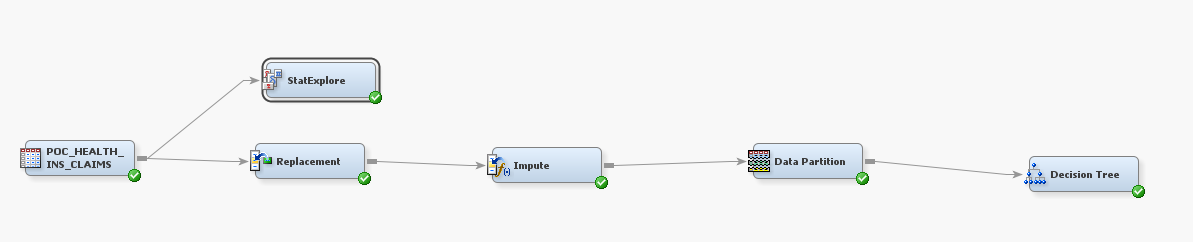
**Appendix 3**



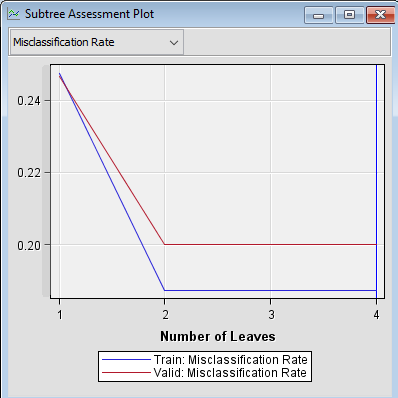
**Appendix 4**



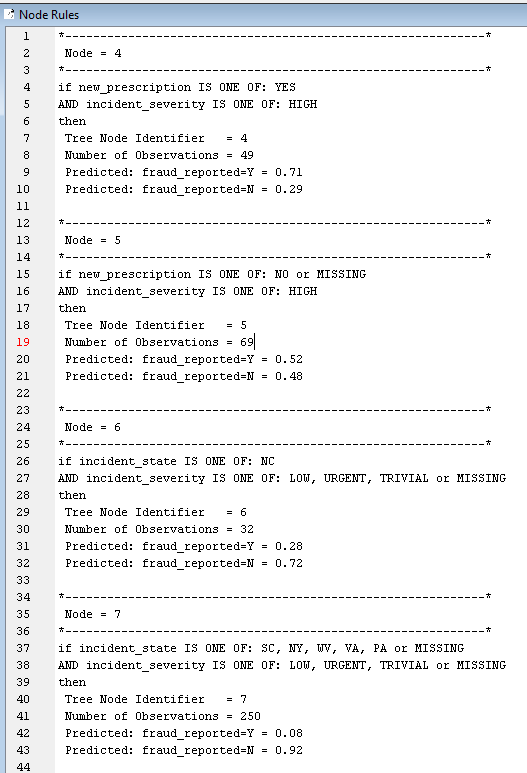
**Appendix 5**



**Appendix 6**



**Appendix 7**



# I) References

1. Advanced Analytics for Fraud, Waste, and Abuse Detection. (2018). Retrieved from <https://s3.amazonaws.com/cdn.ayasdi.com/wp-content/uploads/2015/06/02151543/AYASDI-Advanced-Anaylytics-for-Fraud-Waste-and-Abuse-Detection-WP-071217V01.pdf>
2. Background on: Insurance fraud | III. (2018). Retrieved from <https://www.iii.org/article/background-on-insurance-fraud>
3. By the numbers: fraud statistics. (2018). Retrieved from <https://www.insurancefraud.org/statistics.htm#10>
4. Combating Health Care Fraud in a Post-Reform World. (2010). Retrieved from <https://www.nhcaa.org/docs/nhcaa/PDFs/Member%20Services/WhitePaper_Oct10.pdf>
5. Corbin, K. (2014). How CIOs Can Prepare for Healthcare ‘Data Tsunami’. Retrieved from <https://www.cio.com/article/2860072/healthcare/how-cios-can-prepare-for-healthcare-data-tsunami.html>
6. Course 5: Medical Billing Insurance Claims Process. Retrieved from <https://www.medicalbillingandcodingonline.com/medical-billing-claims-process/>
7. Fighting Health Care Fraud: An Integral Part of Health Care Reform. (2009). Retrieved from <https://www.nhcaa.org/docs/nhcaa/PDFs/Member%20Services/Fighting%20Health%20Care%20Fraud_NHCAAJune2009.pdf>
8. Fraud Detection Using Data Analytics in the Healthcare Industry. (2018). Retrieved from <https://www.acl.com/pdfs/DP_Fraud_detection_HEALTHCARE.pdf>
9. Insurance Fraud. (2018). Retrieved from <https://www.naic.org/cipr_topics/topic_insurance_fraud.htm>
10. Medical Identity Theft. (2018). Retrieved from <https://www.consumer.ftc.gov/articles/0171-medical-identity-theft>