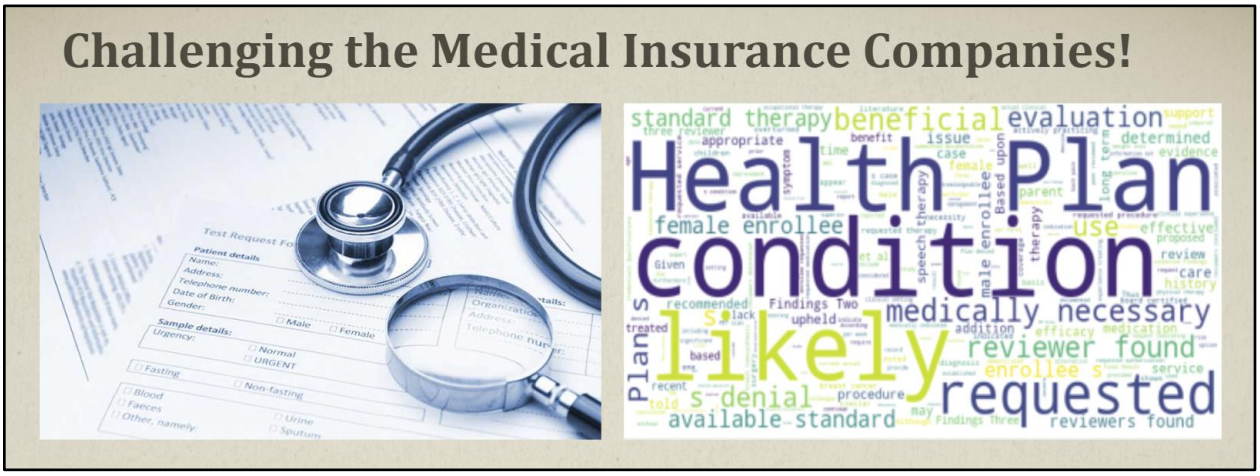
Introduction

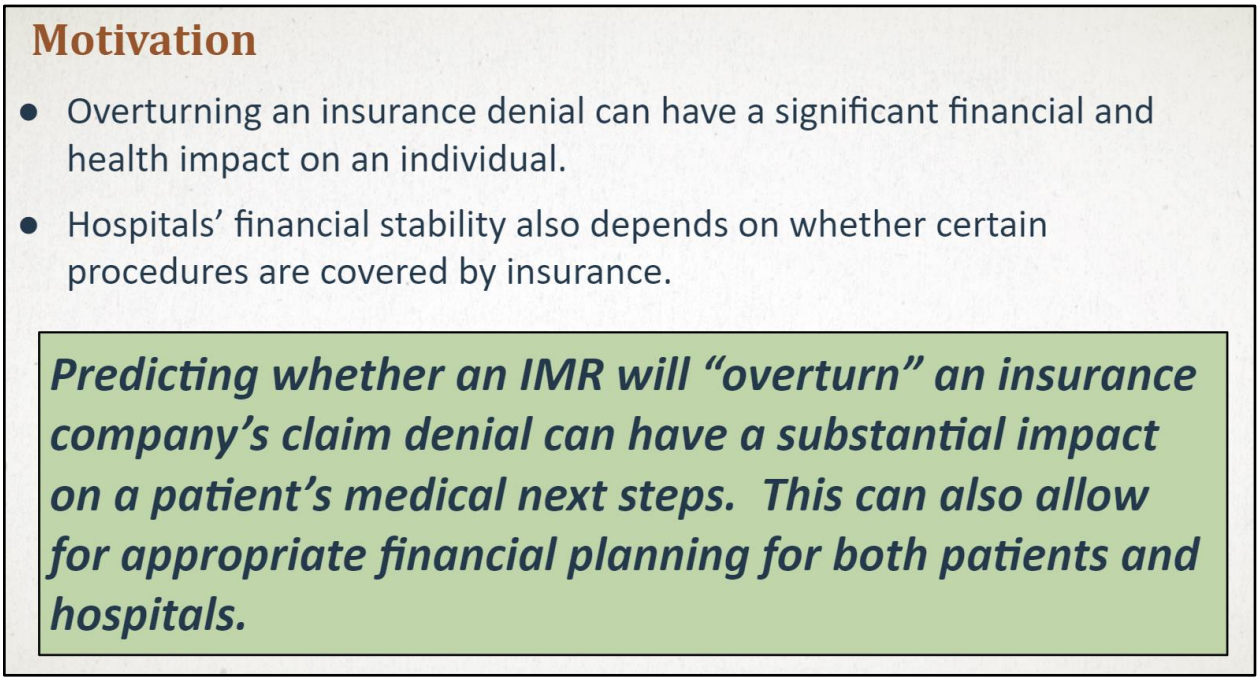


The healthcare industry in the US is one of the largest areas of spending in the country. For 2023, the national health expenditure is expected to top $4.7 Trillion, or about $14K per person. The US has a complicated payer system in which individuals largely rely on private insurance. Unfortunately, many patients have claims denied which requires extra resources to be assigned to fight these decisions.

The state of California permits review of an insurance company's denied, delayed, or modified service to a patient's health care plan. This is managed by the California Department of Managed Health Care (DMHC) via an Independent Medical Review (IMR) board. This process occurs when a patient contests a denied, delayed, or modified service by their health care plan. The IMR is carried out by independent physicians with no affiliation to the insurance company. During this process, the patient submits a request for review, and the independent reviewer carefully assesses the medical records, information provided by the patient, their physician, and the insurance company. Subsequently, the reviewer issues a binding decision that is applicable to both parties involved.

Problem Statement/Motivation

How can we develop an accurate and reliable classification model to predict the outcome of the IMR's decision with the goal of identifying potential biases or disparities in the review process?



Analysis and Models

Subsection 1: Data Preparation and Processing

The data was obtained from the California Department of Managed Health Care (DMHC) spanning from 2001 to 2023, which includes over 34,000 IMR decisions. The dataset consisted of both a structured and unstructured portion. The structured data provided information about each claim, such as the report year, diagnosis, treatment details, ruling outcome, gender, and days in the review process. On the other hand, the unstructured data consisted of extensive explanations provided by physicians for each case, averaging around 300 words each, totaling approximately 10 million words across the dataset.

1.1: Data Structure

Variable	Type	Levels	Description
Reference ID	Nominal	-	Unique identification assigned to each review
Report Year	Ordinal	23	Year in which the review was conducted
Diagnosis Category	Nominal	29	Categorization of patient diagnosis
Diagnosis Category	Nominal	416	Detailed diagnosis
Treatment Category	Nominal	34	Categorization of the sought treatment
Treatment Sub-Category	Nominal	378	Detailed treatment
Determination (Target)	Nominal	2	Ruling outcome by the IMR board (overturned or upheld)
Type	Nominal	3	Reason for denial of claim categorized as urgent care, medical necessity, or experimental
IMR Type	Nominal	2	Expedited vs Standard IMR review
Age Range	Ordinal	7	Various age groupings from 0-10 through 65+
Patient Gender	Nominal	3	Gender of the patient (Male/Female/Other)
Days to Review	Continuous	-	Number of days in physician review
Days to Adopt	Continuous	-	Number of days for the Administrative Director to adopt the determination of the physician reviewer
Findings	String	-	Describes the specific medical procedures or treatment requested by the enrollee. Describes the potential benefits or advantages of the requested procedure or treatment compared to standard therapies. Offers supporting evidence or citations to justify the request. States the reviewer's conclusion

Ultimately, the data was used in the following ways:

Field	Description	Use
Determination	Ruling outcome by the IMR board (overturned or upheld)	Target
Diagnosis Category	Categorization of patient diagnosis (29 levels)	Predictor
Diagnosis Sub-Category	Detailed diagnosis within the Diagnosis Category (415 levels)	Predictor
Treatment Category	Categorization of the sought treatment (34 levels)	Predictor
Treatment Sub-Category	Detailed treatment within the Treatment Category (377 levels)	Predictor
Age Range	Various patient age groupings from 0-10 through 65+	Predictor
Patient Gender	Gender of the patient (Male/Female/Other)	Predictor
Type	Urgent care, medical necessity, or experimental	Predictor
IMR Type	Expedited vs Standard Review	Predictor
Report Year	Year in which the review was conducted	EDA
Days to Review	Number of days in physician review	EDA
Days to Adopt	Number of days to adopt the review	EDA
Reference ID	Unique identification assigned to each review	Not Used

1.2: Cleaning and Preprocessing

The dataset was well organized and presented in a tabular format. However, there were 685 missing values in the Age and Gender features. Upon examination of the Findings feature, 99% of the missing values in the Gender feature and 93% of the missing values in the Age feature were identified.

Feature	Total NAs	NAs After Extraction
AgeRange	685	50
PatientGender	684	2

All Diagnosis and Treatment categories were converted to nominal factors, while Age Range was converted to an ordinal factor. The Days in Review variables were scaled in order to account for the varying degrees of magnitude, range, and units of the features.

For Random-Forest (RF) modeling, the model struggled with the high cardinality of the Diagnosis and Treatment sub-categories, and they were converted to a numerical format.

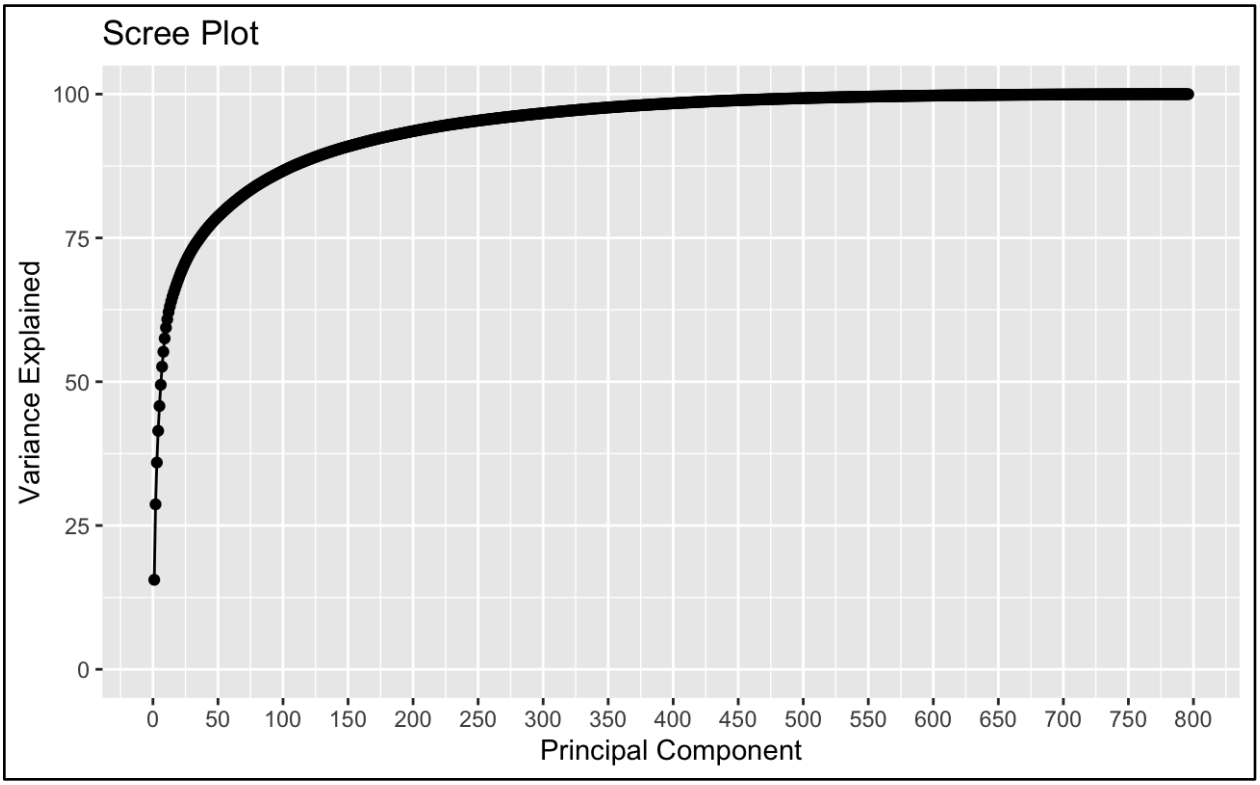
For SVM and KNN modeling, one-hot encoding was performed on the Diagnosis and Treatment categories, which converts variables into binary vectors. Due to the high cardinality of these features (415 diagnosis sub-categories and 377 treatment sub-categories), over 800 new features were created.

Principal Component Analysis (PCA) was used to reduce dimensionality of the data while preserving as much of the relevant information as possible. Below is how PCA general works:

- It calculates the covariance matrix of the high-dimensional data to find the directions (vectors) along which the data varies the most.
- It projects the high-dimensional data onto a new subspace defined by the principal components - the eigenvectors of the covariance matrix sorted by decreasing eigenvalues.
- By keeping only the top k principal components, it effectively reduces the dimensionality of the data from n to k, where k < n.

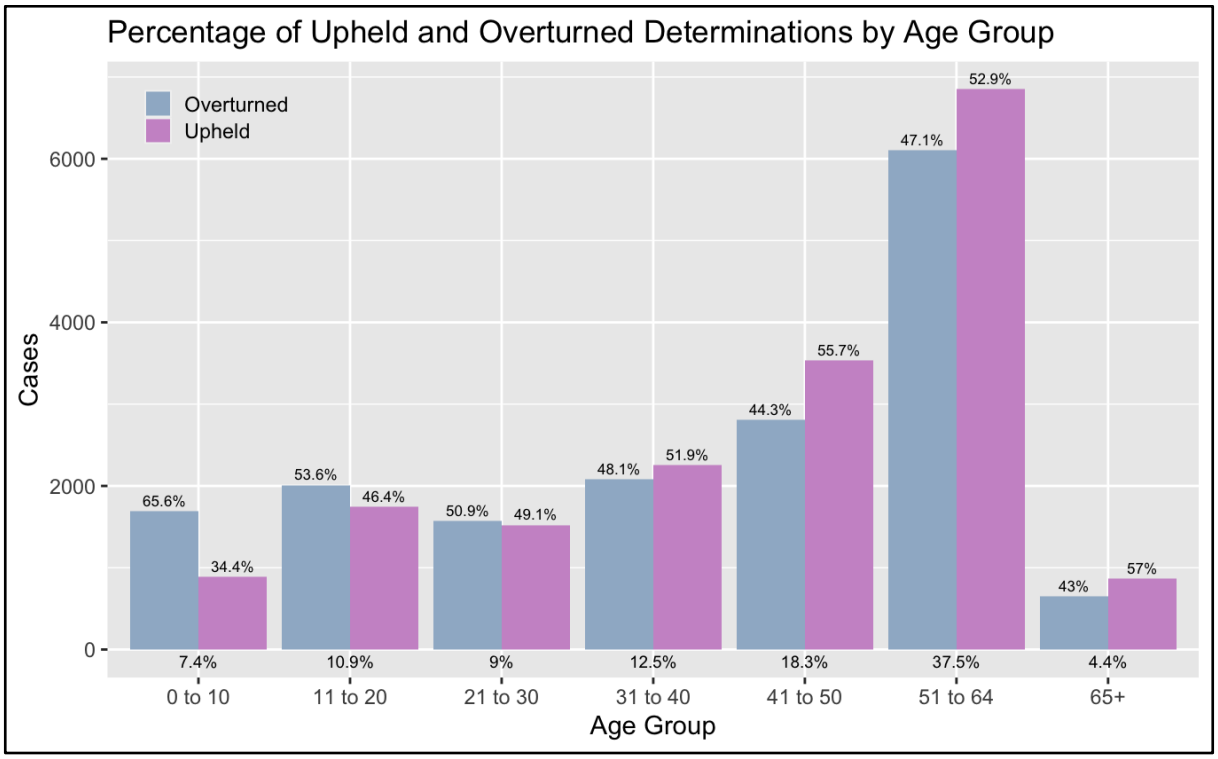
A scree plot was used to identify the point at which retaining additional components is no longer adding sufficient information to the model. These cutoff balances preserve the major patterns in the data while eliminating noise and redundancy.

While the elbow point is subjective, the scree plot provides a visual aid to select principal components judiciously based on the slope changes and explained variance cutoff desired. Based on the plot below, 200 dimensions were selected for optimization which accounted for approximately 93.6% of the variance in the data.

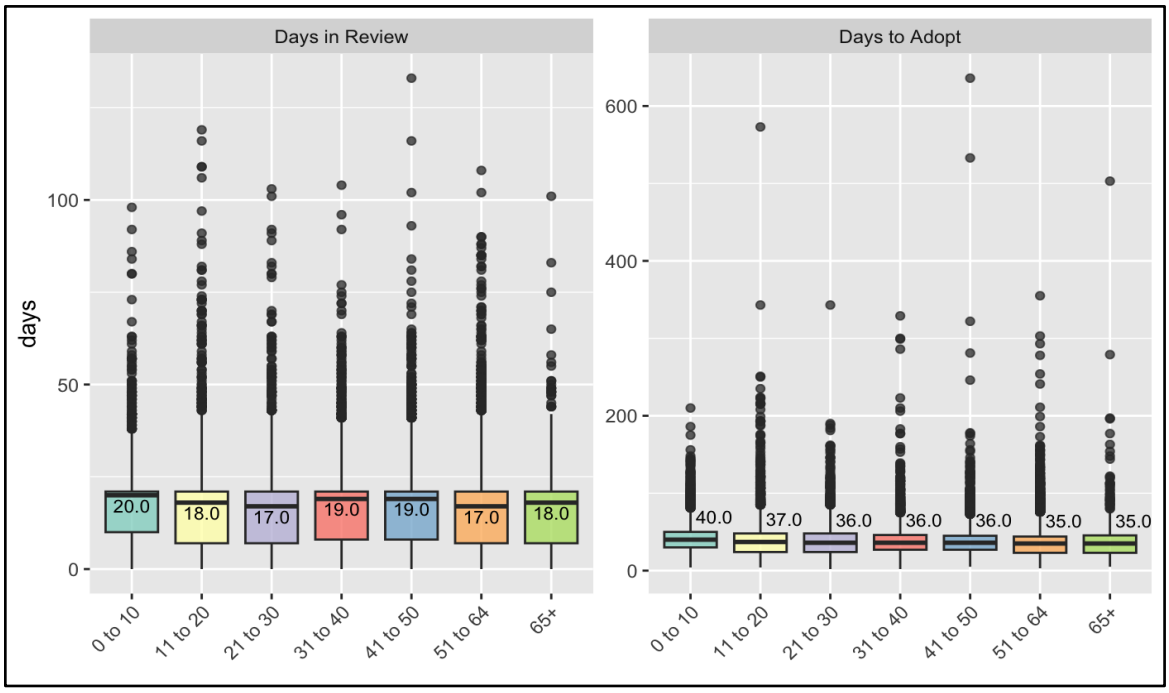


1.2: Exploratory Data Analysis (EDA)

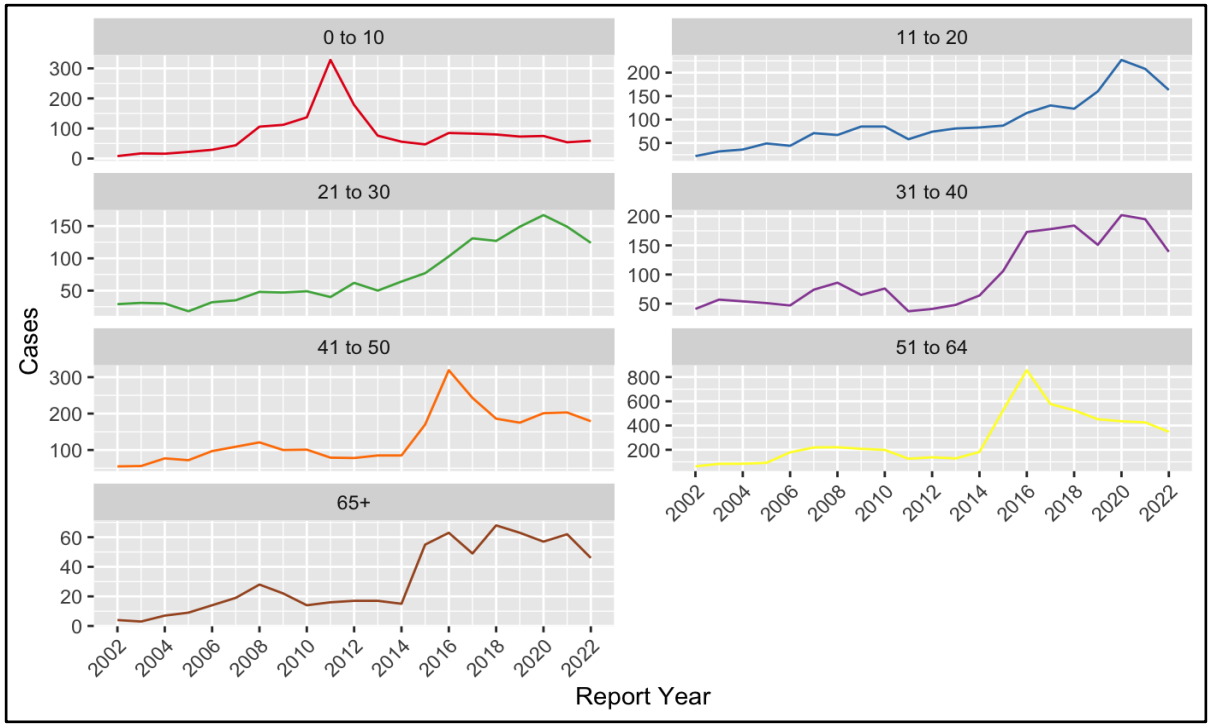
Most IMR review cases occurred within the 51 to 64-year-old age group. Curiously, the 65+ age group has comparatively minimal reviews which suggests that the IMR board does not review Medicare patients.



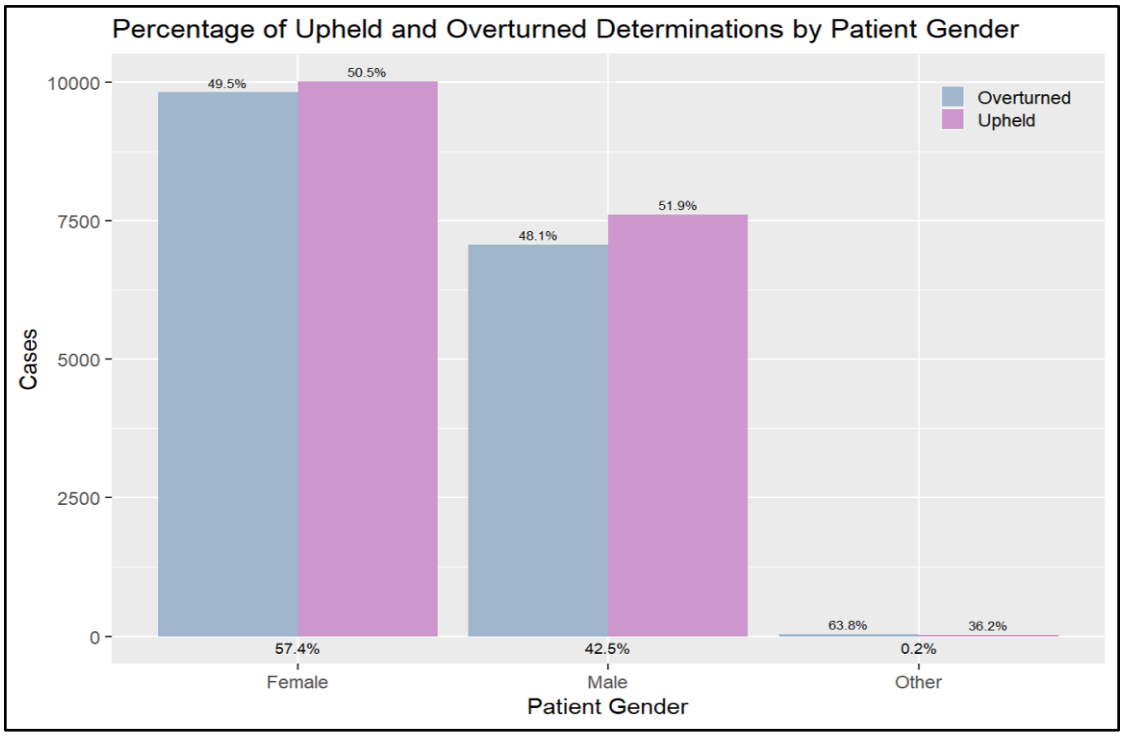
There are two metrics to assess the time it takes to reach a final binding decision. Days in Review (left) which are days in the physician's review and Days to Adopt (right) which are the days it takes for the legally binding decision.



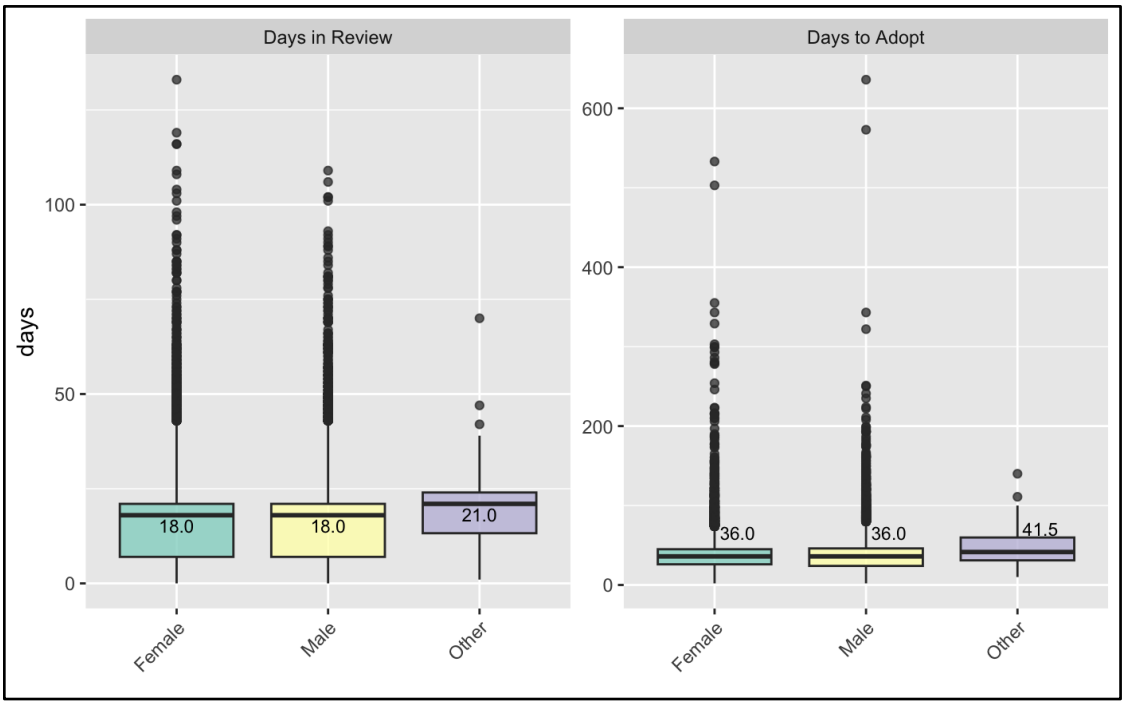
There is a notable uptick in overturned cases in 2016 (shown below) for all but the youngest age-group (0-10), which interestingly showed an uptick in the year 2011. Understanding the modeling implications of this uptick is performed subsequently during the Random Forest modeling.



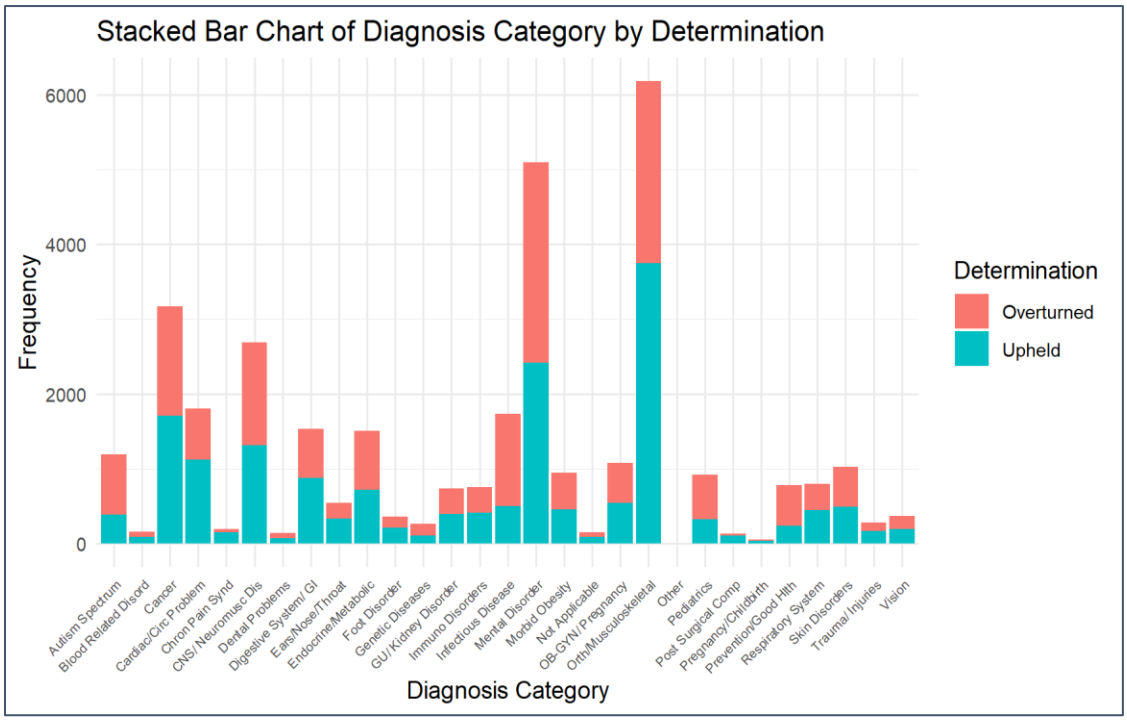
Females have over 5,000 more cases than the other genders and make up 57.4% of the entire cases reviewed. Curiously, the other gender has significantly more cases overturned but only make up 0.2% of the dataset.



The distributions between genders are visually similar to the distributions between the age-groups. Female and Males have heavily right skewed distributions, but the other gender is less variable but has more days in review.



A significant number of IMR cases were within Cancer, Orthopedic, Central Nervous System (CNS), and Mental Disorder as shown below:



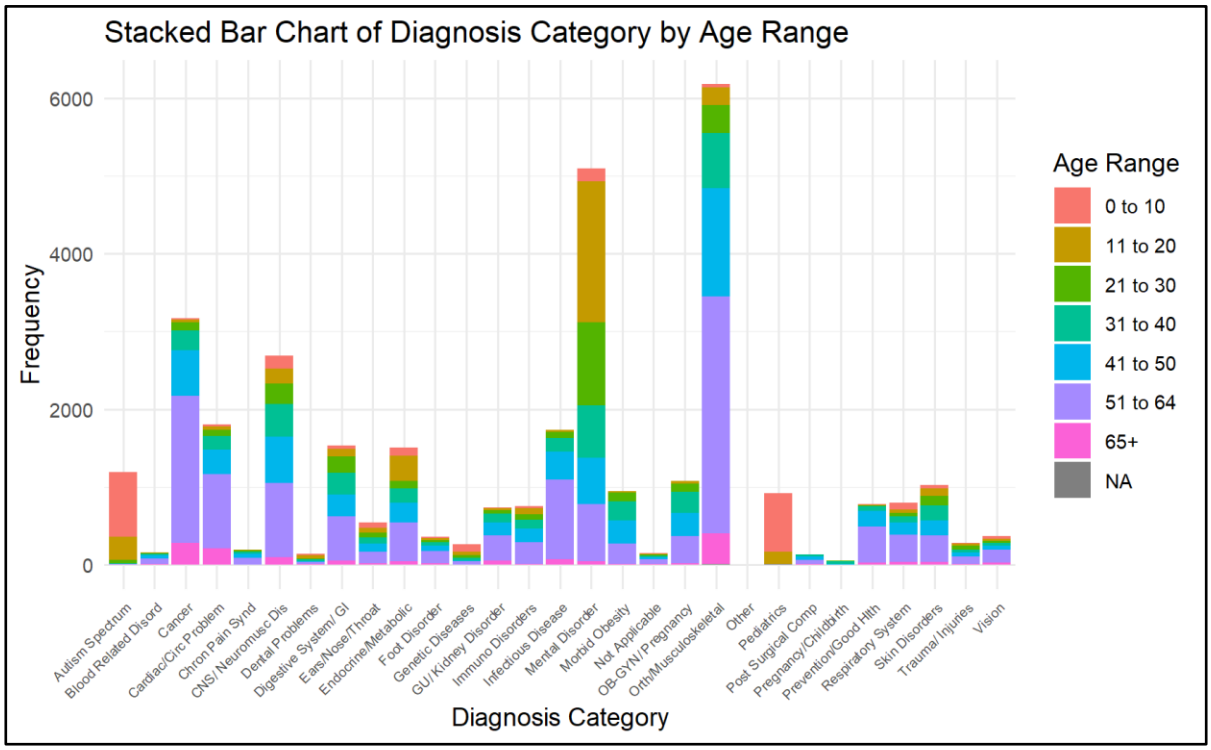
Orthopedic and Cardiac problems exhibited the lowest rates of overturned cases, while Infectious Disease, Autism, and Pediatrics had the highest rates of overturned cases per the patient's diagnosis.

Description: df [29 × 4]

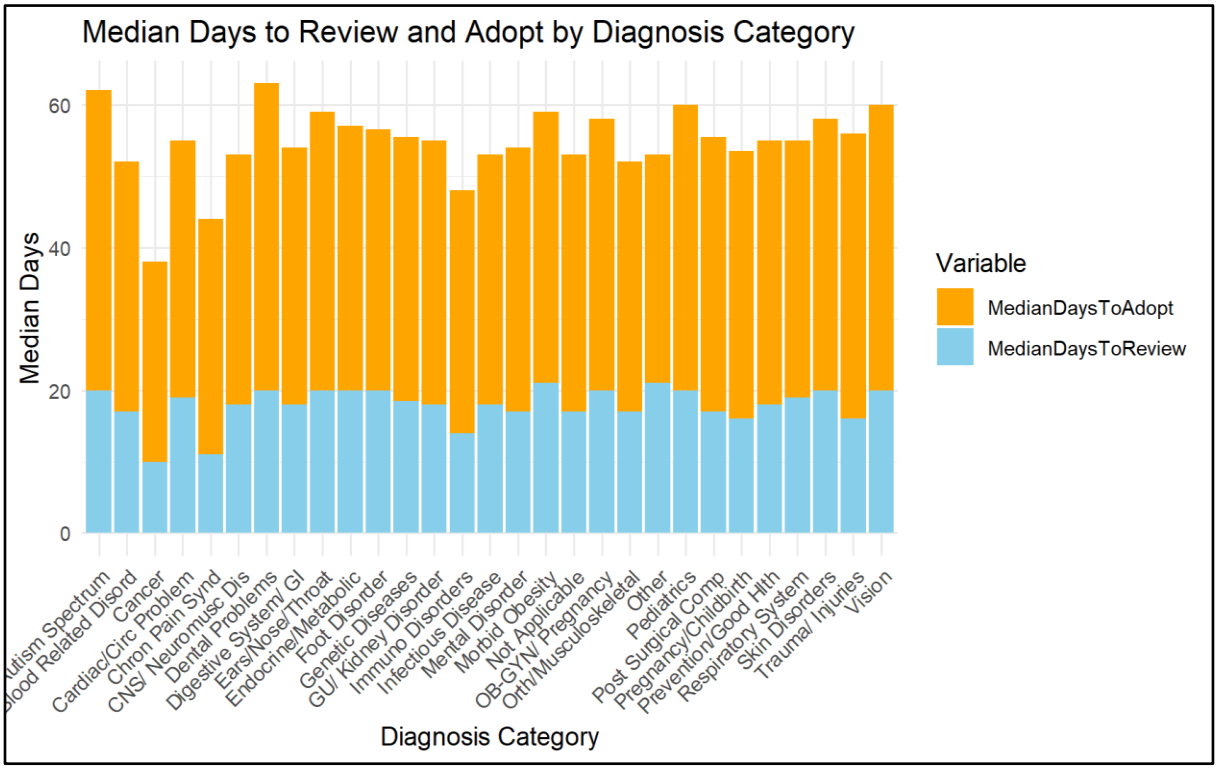
DiagnosisCategory <fctr>	OverturnedCount <int>	UpheldCount <int>	PercentOverturned <dbl>
Mental Disorder	2675	2419	52.51276
Orth/Musculoskeletal	2426	3752	39.26837
Cancer	1463	1709	46.12232
CNS/ Neuromusc Dis	1377	1315	51.15156
Infectious Disease	1235	501	71.14055
Autism Spectrum	807	384	67.75819
Endocrine/Metabolic	783	724	51.95753
Cardiac/Circ Problem	681	1124	37.72853
Digestive System/ GI	658	880	42.78283
Pediatrics	592	328	64.34783

1-10 of 29 rows

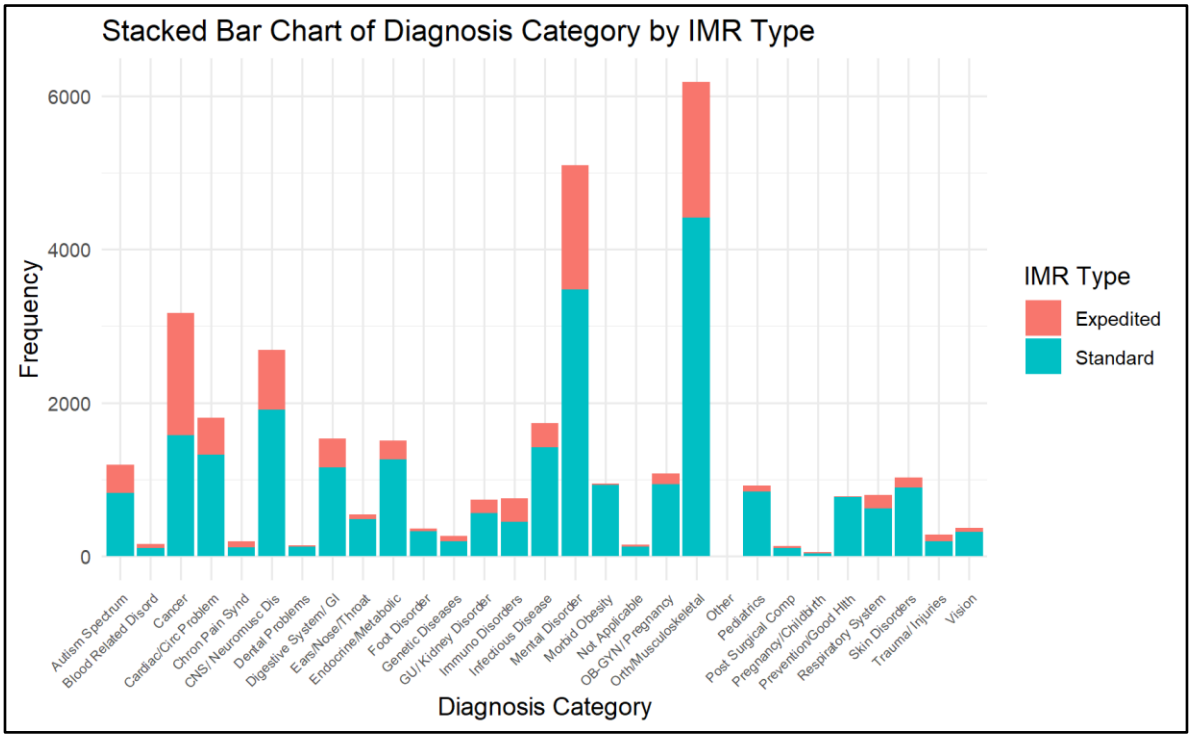
Visually, it's clear that the 51-64 age group makes up the vast majority of IMR reviews across all Diagnosis categories. Noteworthy, the 11-20 age group makes up a significant portion of the Mental Disorder IMR reviews.



Cancer and Chronic Pain have the lowest IMR review through adoption time. Autism and Dental problems have the highest IMR review through adoption time.



Most IMR requests follow the standard process. Approximately 50% of Cancer requests, understandably, are expedited.



Subsection 2: Modeling

2.1: Naïve Bayes

Naïve Bayes classifiers are a family of simple “probabilistic classifiers” based on applying Bayes’ theorem with strong (naïve) independence assumptions between the features. They are among the simplest Bayesian network models. However, when coupled with kernel density estimation higher accuracy levels are achieved.

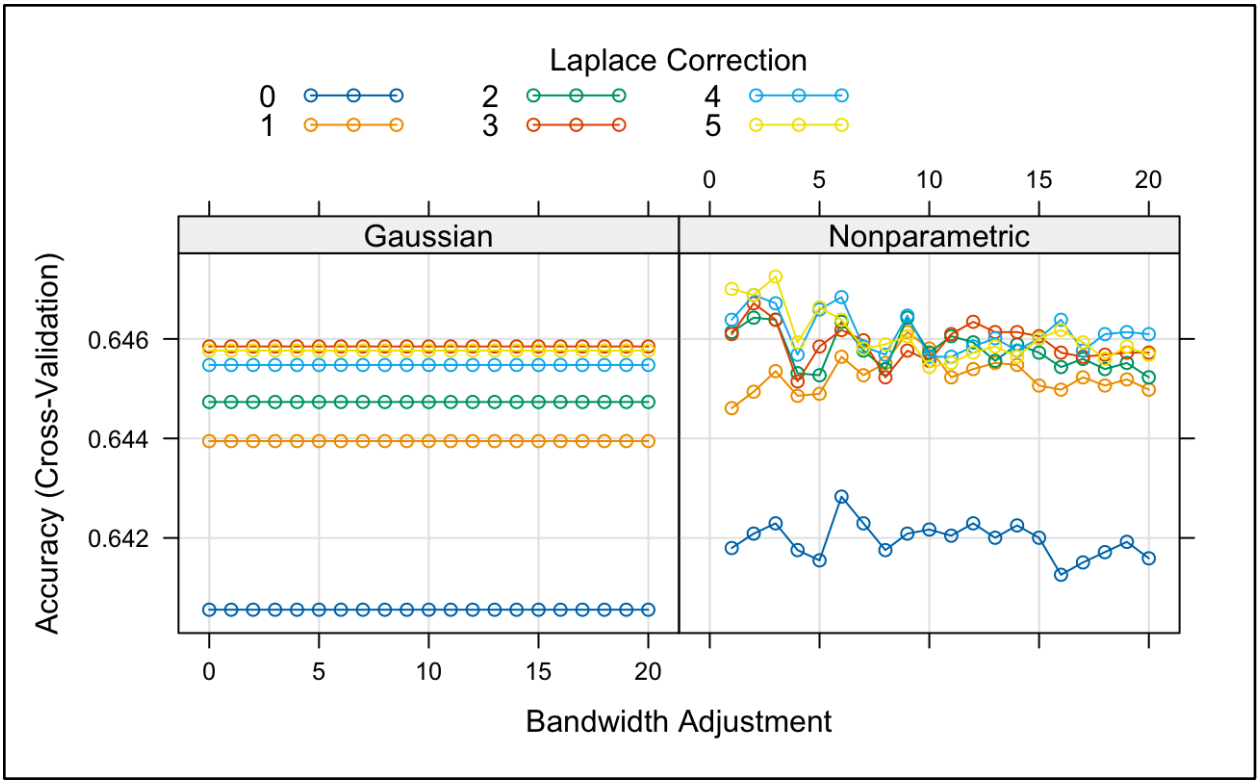
Naive Bayes classifiers are a family of simple probabilistic machine learning models that are based on applying Bayes' theorem with strong (naïve) independence assumptions between the features.

They are called naïve because they assume that each input feature is conditionally independent from every other input feature for a given value of the class variable. This independence assumption is clearly a simplification of reality but can work surprisingly well when coupled with Kernel Density Estimation (KDE).

KDE is specifically used when dealing with continuous predictor variables that may not follow a standard parametric distribution like Gaussian. For a given continuous predictor variable X, instead of assuming it follows a Gaussian (normal) distribution, KDE is used to estimate the probability density p(X) directly from the training data samples.

The core idea is to use the joint probabilities of features and classes to estimate the conditional probabilities of classes given a particular observation of features. Bayes' theorem calculates the conditional probability from the joint probability. To classify a new instance, the Naive Bayes classifier calculates the posterior probability of each class given the feature values of the new instance. The class with the highest posterior probability is then assigned as the predicted class.

Using 10-fold cross-validation, 20 bandwidth adjustments, and 5 Laplace smoothers were adjusted for both kernel and non-kernel density estimations. The best accuracies were identified using kernel density estimations with Bandwidth set to three and Laplace set to five for non-parametric approximations. The "Days in Review" feature was severely right-skewed, so the non-parametric approximation worked best in this scenario.



2.1.1: Naïve Bayes Limitations

1. The Naive Bayes model assumes that all features are conditionally independent given the class variable. This assumption is often violated in real-world datasets, where features may be correlated or dependent on each other. Violating this assumption can lead to inaccurate probability estimates and suboptimal model performance.
2. Due to the independence assumption, the Naive Bayes model cannot learn feature interactions or capture complex relationships between features. It treats each feature independently, which may not be suitable for problems where feature interactions play a significant role.
3. When dealing with categorical features, the Naive Bayes model can encounter the zero-frequency problem, where a feature value has zero occurrences in the training data for a particular class. This can lead to a probability of zero, making it impossible to make predictions for new instances with that feature value. Techniques like Laplace smoothing, or additive smoothing are typically used to address this issue.
4. The Naive Bayes model assumes that continuous features follow a specific distribution, typically Gaussian (normal) distribution. If the features do not follow this distribution, the model's performance may be affected. In such cases, techniques like kernel density estimation or data transformation may be necessary.
5. The Naive Bayes model can be sensitive to imbalanced data, where one class is significantly underrepresented compared to others. This can lead to biased probability estimates and poor performance on the minority class.
6. The Naive Bayes model does not have an inherent regularization mechanism, making it prone to overfitting, especially when dealing with high-dimensional data or when there is noise in the training data.
7. The Naive Bayes model assumes linear decision boundaries between classes, which may not be suitable for problems with non-linear decision boundaries or complex decision regions.
8. The Naive Bayes model can be sensitive to the scale of the features, as it assumes that all features have equal importance. Feature scaling or normalization may be required to ensure that features with larger scales do not dominate the model's predictions.

2.2: K-Nearest Neighbor (KNN)

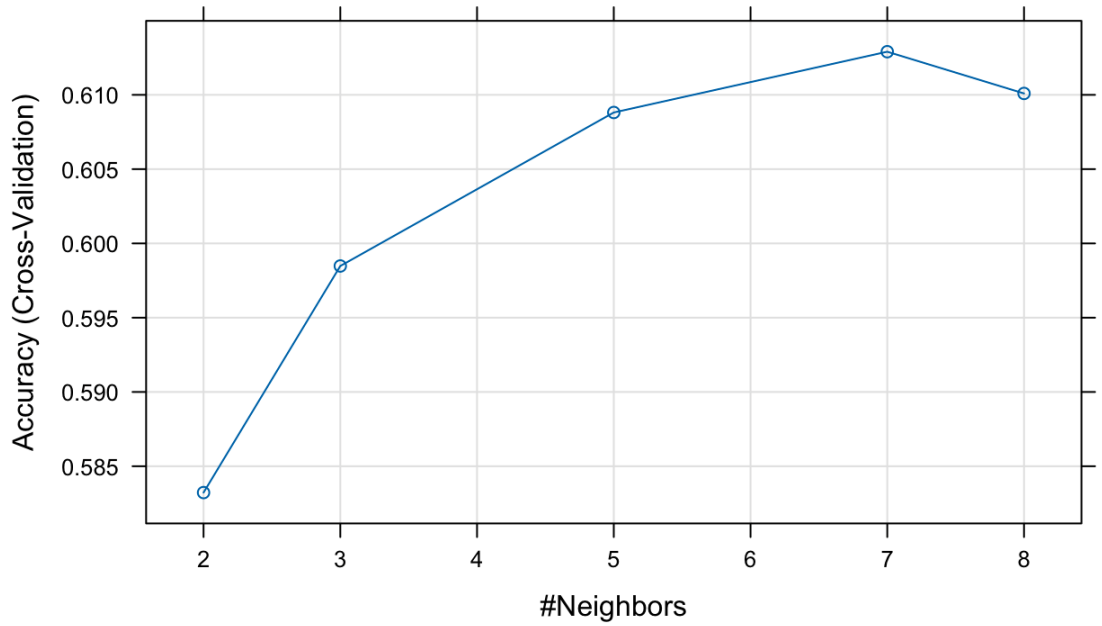
K-Nearest Neighbor (KNN) is a non-parametric, supervised machine learning algorithm used for both classification and regression tasks. It is a type of instance-based learning where the algorithm makes predictions based on the similarity or distance between the new data point and the available training data points.

Here's how the KNN algorithm works:

1. **Feature Similarity**: For a new data point, the algorithm calculates the distance or similarity between that data point and all the other training data points. The distance can be calculated using various metrics such as Euclidean distance, Manhattan distance, or Minkowski distance, depending on the problem and the type of data.
2. **K Value**: The user specifies the value of K, which represents the number of nearest neighbors to consider for making a prediction.
3. **Nearest Neighbors**: The algorithm selects the K training data points that are closest or most similar to the new data point based on the distance metric.
4. **Prediction**:
    - **For Classification**: The algorithm assigns the class label that is most frequent among the K nearest neighbors to the new data point.
    - **For Regression**: The algorithm takes the average or median of the target variable values of the K nearest neighbors and assigns that value as the prediction for the new data point.

The choice of the K value and the distance metric can significantly impact the performance of the KNN algorithm. A smaller value of K can lead to overfitting, where the model becomes too sensitive to noise or outliers in the training data. On the other hand, a larger value of K can lead to underfitting, where the model becomes too generalized and fails to capture the nuances of the data.

Based on a 10-fold cross-validation, the top accuracy of 64.7% was achieved by setting the number of neighbors to seven (k=7).



2.2.1: KNN Limitations

1. KNN models are computationally expensive, especially for large datasets, as it needs to calculate the distance between the new data point and all training data points.
2. KNN models are sensitive to the scale of the features and the presence of irrelevant features.
3. KNN models are required to store the entire training dataset for making predictions.
4. KNN models are Sensitive to the choice of K value and distance metric.

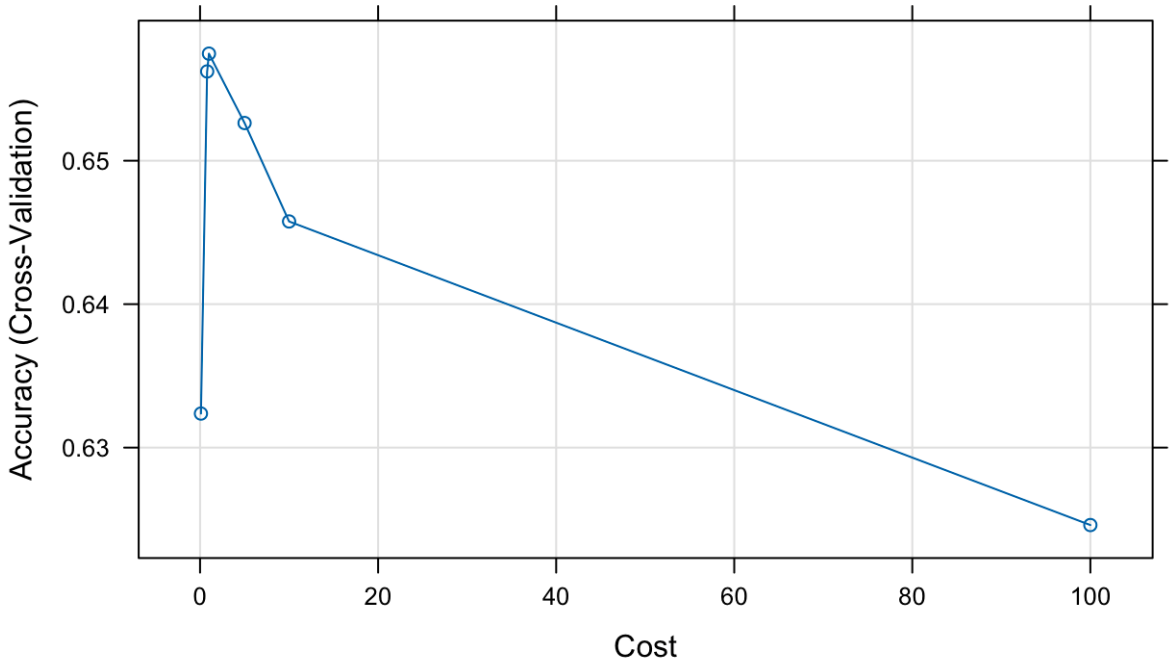
2.3: Support Vector Machine (SVM)

Support Vector Machines (SVMs) are a powerful and widely used machine learning algorithm for both classification and regression tasks. SVMs are based on the idea of finding the optimal hyperplane that separates different classes in a high-dimensional feature space.

Here's how SVMs work:

1. **Mapping Data to Higher Dimensions**: The data is mapped from the original input space to a higher-dimensional feature space using a kernel function. This transformation allows the SVM to find a separating hyperplane that maximizes the margin between the classes, even if the data is not linearly separable in the original input space.
2. **Finding the Optimal Hyperplane**: The SVM algorithm seeks to find the hyperplane that maximizes the margin between the closest data points of different classes. These closest data points are called support vectors, and they define the decision boundary.
3. **Kernel Functions**: To map the data to a higher-dimensional space, SVMs use kernel functions, such as linear, polynomial, radial basis function (RBF), or sigmoid kernels. The choice of kernel function depends on the problem and the characteristics of the data.
4. **Prediction**: Once the optimal hyperplane is determined, new data points can be classified by mapping them to the higher-dimensional feature space and determining on which side of the hyperplane they fall.

Based on a 10-fold cross-validation, the top accuracy of 65.7% was achieved by setting the cost parameter to one (C=1).



2.3.1: SVM Limitations

1. The choice of kernel function and its parameters can significantly impact the performance of the SVM, and finding the optimal kernel and parameters can be a challenging task.
2. Training an SVM can be computationally expensive, especially for large datasets or when using complex kernel functions.
3. SVMs are considered "black-box" models, making it difficult to interpret the learned decision boundaries or extract feature importance.
4. The performance of SVMs can be sensitive to the choice of regularization parameters and kernel parameters, which may require extensive tuning.

2.4: Random Forest (RF)

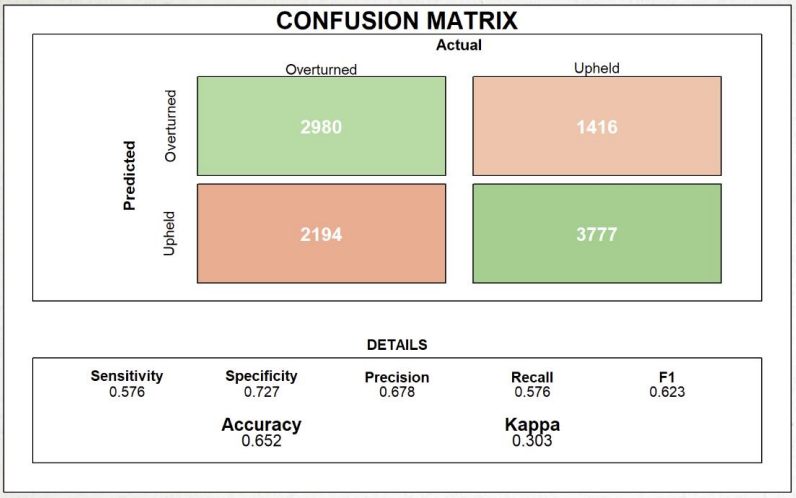
Random Forest is a powerful machine learning algorithm used for both classification and regression tasks. It is an ensemble learning method that combines multiple decision trees to improve predictive accuracy and control overfitting.

Random Forest description:

1. **Bootstrap Aggregating (Bagging)**: Random Forest uses the bagging technique, which involves creating multiple subsets of the training data by randomly sampling from the original dataset with replacement. Each subset is used to train an individual decision tree model.
2. **Random Feature Selection**: In addition to bagging, Random Forest also introduces randomness in the feature selection process. When building each individual decision tree, the algorithm selects a random subset of features from the total available features. This random feature selection helps to reduce correlation between the trees and improve model diversity.
3. **Tree Construction**: For each bootstrap sample, a decision tree is grown to its maximum depth without pruning. This process is repeated for a specified number of trees, typically hundreds or thousands.
4. **Prediction**: To make a prediction for a new instance, the instance is passed through each individual tree in the forest. For classification tasks, each tree casts a vote for the predicted class, and the final prediction is the class with the most votes. For regression tasks, the predictions from all trees are averaged to obtain the final prediction.

Random Forest modeling was turned via 1944 iterations given a parameter grid. The results were reasonable at 65.2% accuracy and a 62.3% F1, but not quite as significant as the KNN model. Noteworthy, running these 1944 iterations was computationally intensive, taking more than 6 hours to complete.

Random Forest Results after Tuning



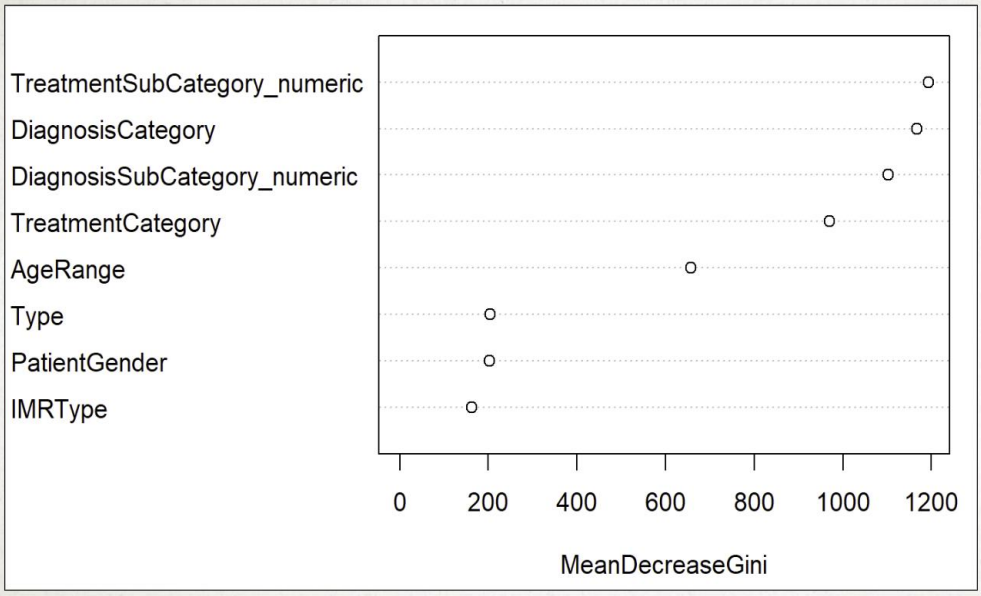
To improve the Random Forest model, 8 parameters were adjusted. Ultimately, the training was done 1,944 times (all parameter combinations). Accuracy was measured against the test data in every case.

Tuning Parameter Grid

```
# Defining parameter values over which to iterate
ntree_values <- c(100, 200, 500)
mtry_values <- c(2, 4, 6)
max_depth_values <- c(10, 20, 30)
min_samples_split_values <- c(2, 5, 10)
min_samples_leaf_values <- c(1, 2, 5)
max_features_values <- c("sqrt", "log2")
criterion_values <- c("gini", "entropy")
bootstrap_values <- c(TRUE, FALSE)
```

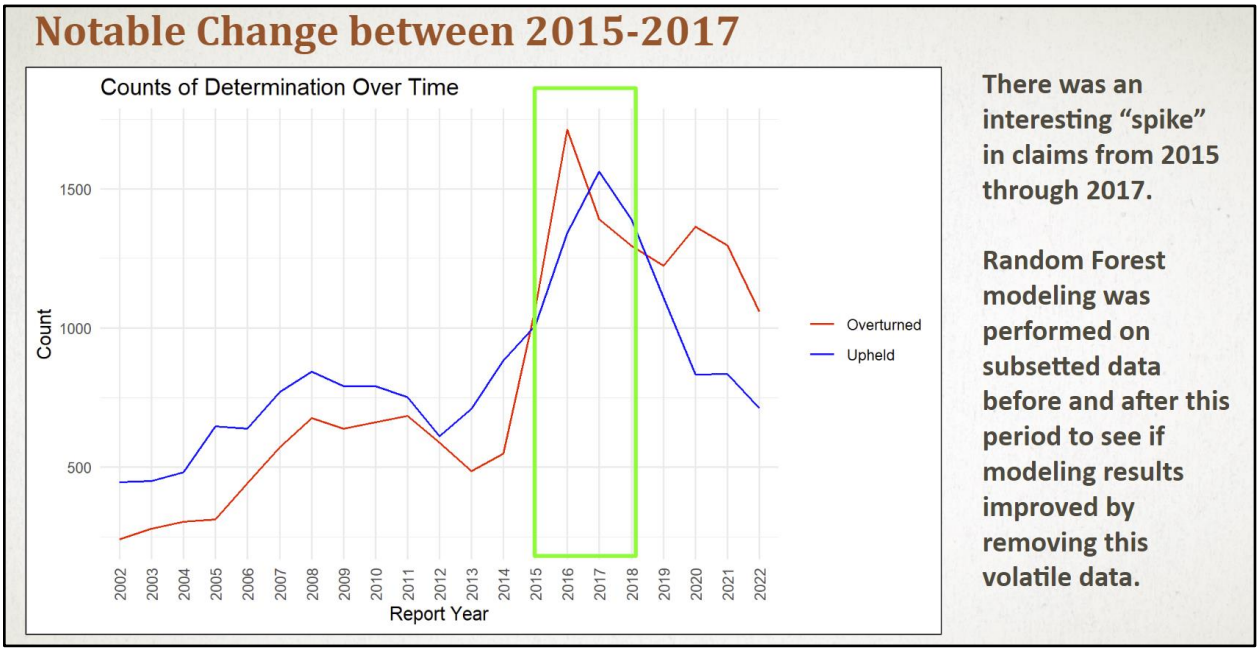
The Treatment and Diagnosis categories/sub-categories were highly important in terms of the Random Forest modeling. With AgeRange next, followed by Type, Gender, and IMR Type (standard or expedited processing).

Variable Importance

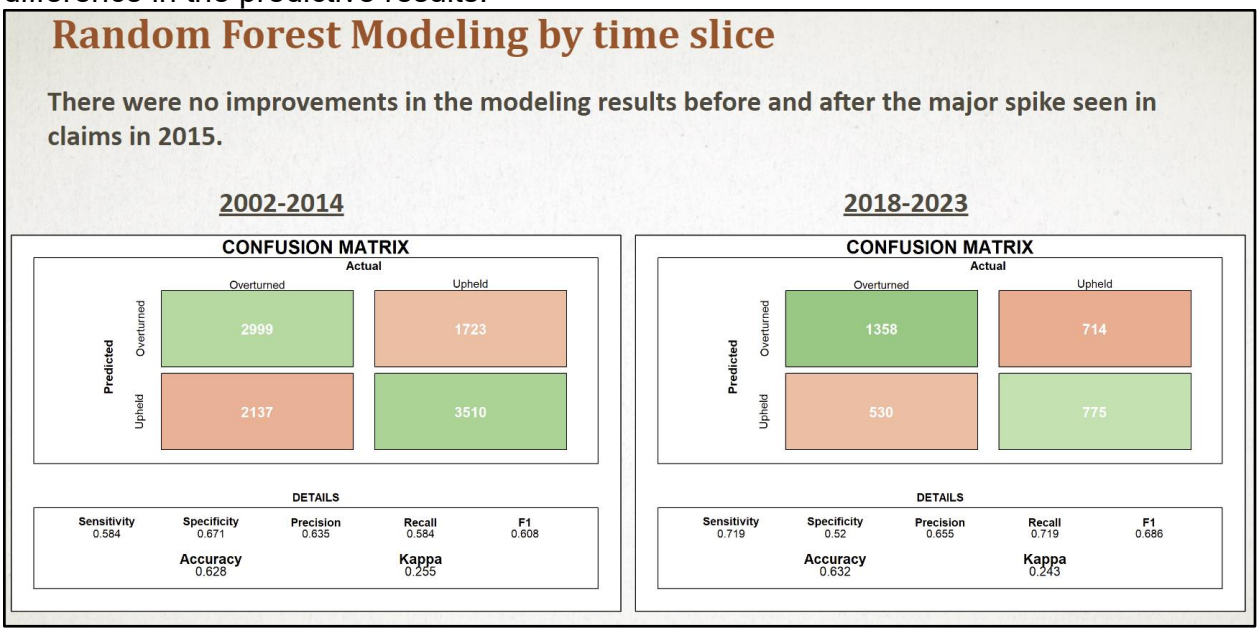


8 key predictors were used in the Random Forest modeling.

There was a notable “spike” in the IMR reviews between 2015 and 2017. Additional modeling was performed before and after this period to understand any differences that might occur if this data is removed.



After Random forest analysis by “time slice” was performed, there was no significant difference in the predictive results.



2.4.1: RF Limitations

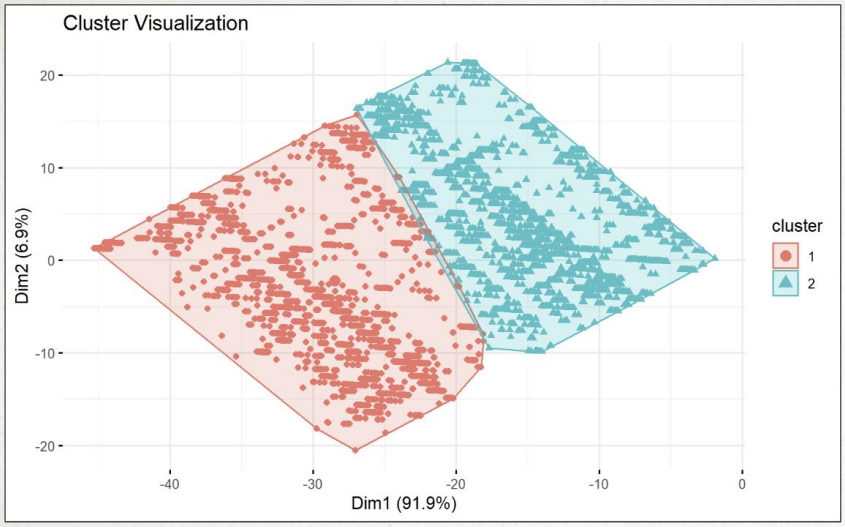
1. While feature importance can be derived, Random Forest models are generally considered "black boxes," making it difficult to interpret the individual decision rules or the relationship between features and predictions.
2. Random Forest can be biased towards selecting correlated features, which may lead to redundant or irrelevant features being included in the model.
3. Training Random Forest models can be memory-intensive, especially for large datasets or when using a large number of trees.

2.5: K-Means Clustering

The K-means algorithm was utilized to partition the data into two cluster centroids. These centroids represent the centers (“means”) of the clusters. Each data point is then assigned to the nearest centroid based on a distance metric. The distance metrics used include Euclidean, Manhattan, Cosine, and Binary distances. After assigning all data points to a cluster, the algorithm updates the centroids by computing the mean of all the data points assigned to each cluster. These updated centroids become the new centers of the clusters. This iterative process continues until convergence is achieved or when the centroids no longer change significantly over the iterations.

A difficulty arises in selecting parameters for clustering algorithms, particularly when small changes in these parameters can drastically affect the clustering results. Consequently, parameter selection often becomes a process of trial and error. Various cluster optimization techniques have been employed to address this challenge, with the results summarized below:

K-Means Clustering Analysis



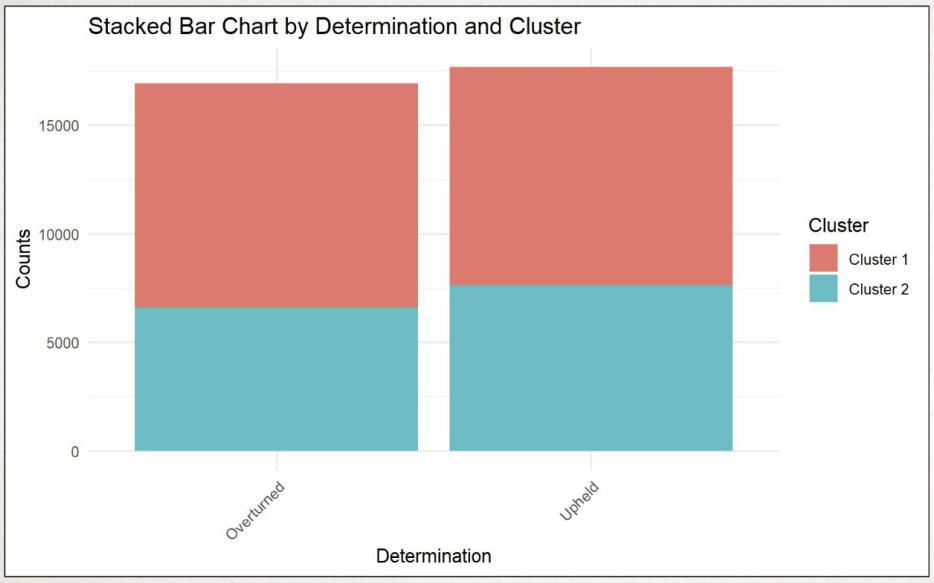
K-Means Clustering analysis with 2 clusters was done to see if the clusters align with IMR Determination. The clusters themselves look compact with the distance between centroids being much higher than the intra-cluster mean distance.

```
Cluster Centroid distance:
15.8719

Mean intra-cluster difference
from cluster centroid:
Cluster 1: 9.793901
Cluster 2: 11.1265
```

While the cluster analysis did show distinct clusters, there was not a relationship between these clusters and the IMR determination as shown below:

Clusters compared to IMR Determination



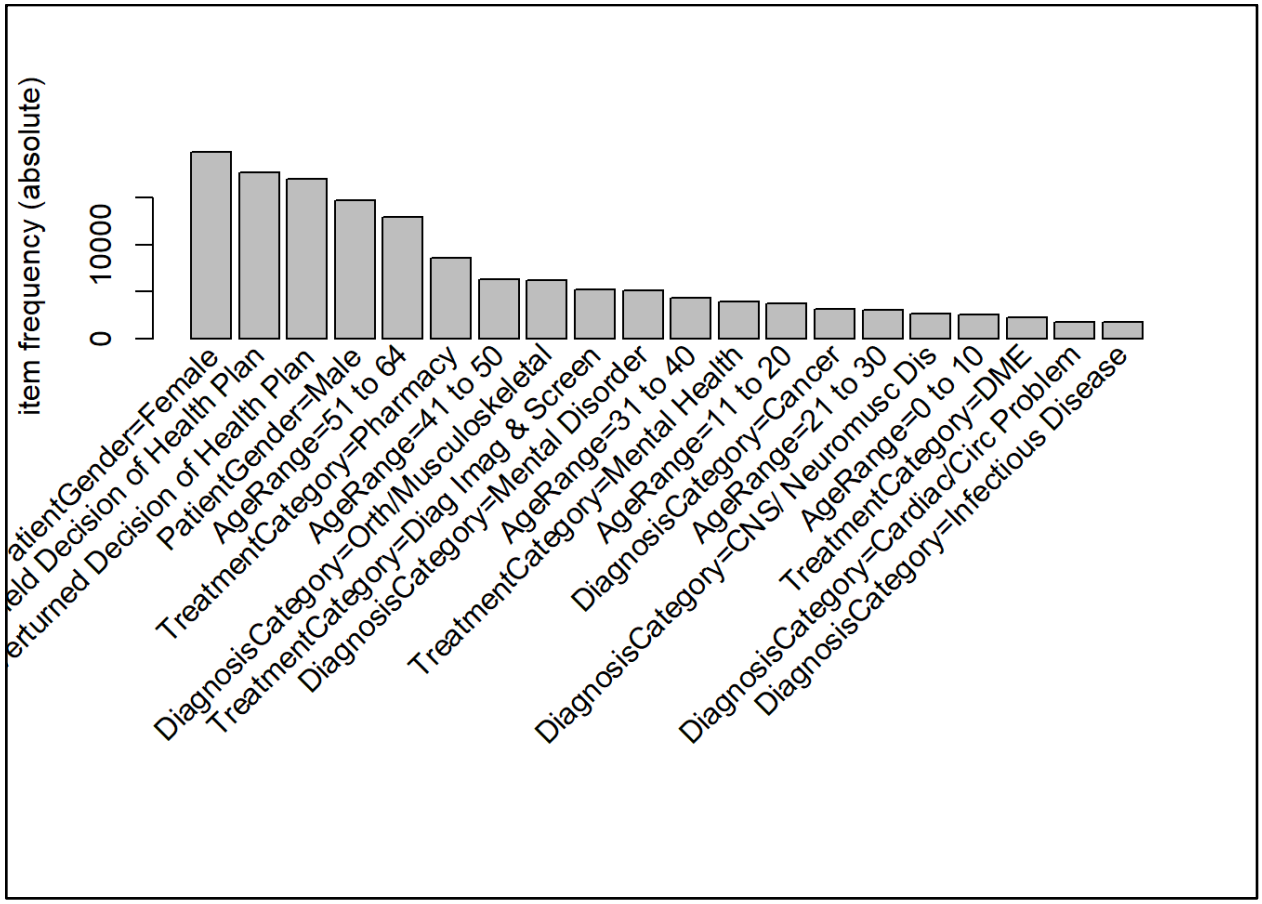
While the cluster analysis did reveal 2 distinct clusters, there is no relationship between these clusters and the IMR Determination.

2.5.1: K-Means Clustering Limitations

1. The K-means model suffers from outliers because they increase SSE and distort centroids.
2. The K-means model has difficulty handling different sizes and shapes of clusters.
3. The K-means algorithm requires the user to specify the number of K clusters, in advance. This makes the selection arbitrary and thus reproducibility decreased.
4. K-means separation continues to increase as K increases until there is one cluster per point. In other words, SSE continues to decrease as K increases. This increases the complexity of the model.

2.6: Apriori Algorithm (Association Rules)

The Apriori algorithm was utilized to derive association rules, shedding light on specific patient characteristics that are more likely to have a case overturned. This algorithm is designed to identify frequent itemsets within a dataset, enabling the identification of patterns based on patient characteristics and determination of case outcome. It systematically evaluates subsets of the dataset, comparing and scoring them based on their frequency across the entire dataset. Therefore, we can first identify the most frequent items in the dataset to better understand the association rules that will be generated. The most frequent items are presented in the graph below:



Three key metrics are employed to assess the association rules:

1. **Support:** is a measure of how frequently a specific item-set appears in the dataset.
2. **Confidence:** For a rule A=>B, it is the proportion of times in which A and B occurred together in the dataset. The number of transactions with both A and B

divided by the total number of transactions having A. (i.e. the proportion of a specific patient characteristic that had a case overturned)

3. **Lift:** is a ratio that indicates how much more likely A and B occur together, compared to if they were independent of each other in the context of a rule A => B. For instance, it can demonstrate the probability of specific patient characteristics (A) and overturned case (B) occur together, relative to the probability of these events occurring if they were independent of each other. It is a measure of dependency.
    - > 1 indicates that the specific patient characteristic (A) and overturned case (B) occur together more often than independently of each other.
    - = 1 implies that A and B are independent of each other.
    - < 1 suggests that A and B occur less often than expected by chance, indicating a negative or non-existent association.

Support and confidence measures gauge the significance of a rule. They are determined by the minimum support and minimum confidence thresholds, which aid in comparing the strength of rules.

The support parameter threshold was fixed at **0.001** to identify patient characteristics with higher rates of overturned cases. The confidence parameter threshold was set between **0.8** and **0.9**, with a final decision made at **0.8**. This determination assumed that a **0.8** rate represents a strong proportion in which A and B occur together, thus providing actionable insights.

2.6.1: Apriori Limitations

1. **High computational complexity:** Apriori algorithm becomes impractical for large datasets due to multiple scans and candidate itemset generation.
2. **Memory usage:** Requires substantial memory, posing constraints and performance issues, especially on systems with limited memory.
3. **Generation of numerous candidate itemsets:** Initial stages generate many candidates, leading to inefficiency in computation and memory usage.
4. **Potential for redundant rules:** May produce redundant or uninteresting rules, requiring filtering and post-processing for meaningful results.
5. **Difficulty in handling infrequent items:** Struggles to identify association rules involving infrequent items, potentially overlooking important but rare patterns.
6. **Inability to handle continuous variables:** Designed for categorical data, necessitating preprocessing like discretization for datasets with continuous attributes.

Results

Subsection 3: Model Results

The below is a description of the key confusion matrix statistics that will be used to quantify the performance for each model. The positive class for all models is Overturned and negative is Upheld.

- **No Information Rate (NIR)** represents the accuracy achieved by always predicting the majority class, which is Upheld at 51.1%.
- **McNemar Test** evaluates the difference between the proportion of False Positives and False Negatives. A significant p-value indicates the model is biased towards one class.
    - *Null Hypothesis:* Classifiers have a similar proportion of errors on the test set.
    - *Alt Hypothesis*: Classifiers have a different proportion of errors on the test set.
- **Sensitivity (Recall)** measures proportion of actual Overturned cases that were correctly predicted as Overturned. Denominator is the population of all Overturned cases.
- **Specificity** measures the proportion of actual Upheld cases that were correctly predicted as Upheld.
- **Positive Predictive Value (Precision)** measures the correctly predicted Overturned cases from all the predicted Overturned cases. Denominator is the population of all predicted Overturned cases.
- **Negative Predictive Value** measures the correctly predicted Upheld cases from all the predicted Upheld cases.
- **Prevalence** is the proportion of Overturned cases
- **F1** is a harmonic mean of Precision and Recall and gives a better measure of the incorrectly classified cases than the Accuracy Metric.
    - Accuracy is used when the True Positives and True negatives are more important while F1-score is used when the False Negatives and False Positives are crucial.
    - Accuracy can be used when the class distribution is similar while F1-score is a better metric when there are imbalanced classes as in the above case.
    - In most real-life classification problems, imbalanced class distribution exists and thus F1-score is a better metric to evaluate our model on.

As a general guideline:

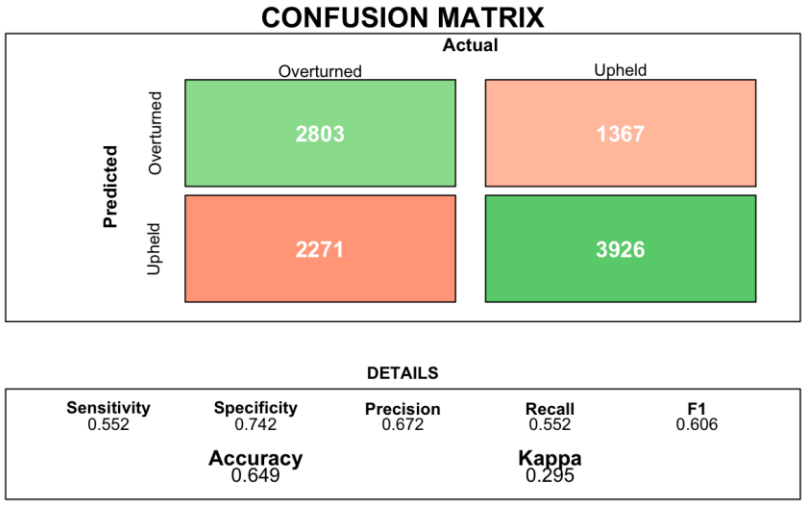
- An F1 score close to 1 indicates a high overall performance, with both high precision and high recall. It suggests that the model achieves a good balance between minimizing false positives and false negatives.
- An F1 score around 0.5 indicates a moderate performance, where precision and recall are roughly balanced but may not be optimal.
- An F1 score close to 0 indicates poor performance, suggesting that the model has low precision and/or low recall.

3.1: Naïve Bayes

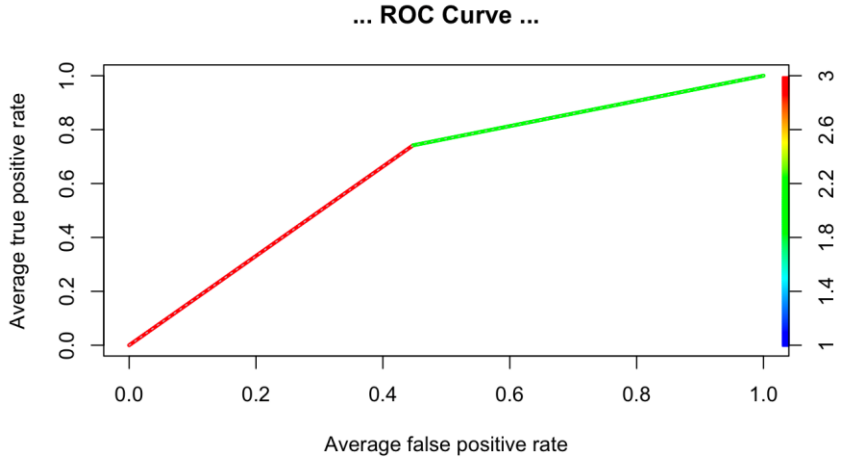
The Naive Bayes model has relatively low recall (55.2%), meaning it struggles to identify a significant portion of the overturned cases correctly. This could lead to potentially missing many cases that should have been overturned.

The precision value is better at 67.2%, but not exceptionally high. This means that while the model is reasonably accurate when it predicts a case as overturned, it still makes a considerable number of false positive errors, predicting 32.8% of upheld cases as overturned.

The overall accuracy of the model was 64.9%, with an F1 score of 60.6%.



The ROC curve indicates that the Naïve Bayes model is not able to discriminate between Overturned and Upheld cases effectively. The ideal curve is bow-shaped and curves closer to the upper left-hand corner. This curve is almost linear.

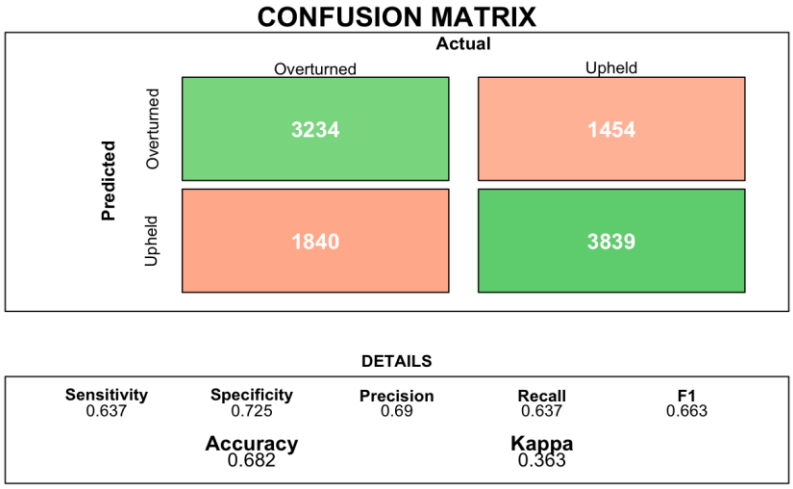


3.2: KNN

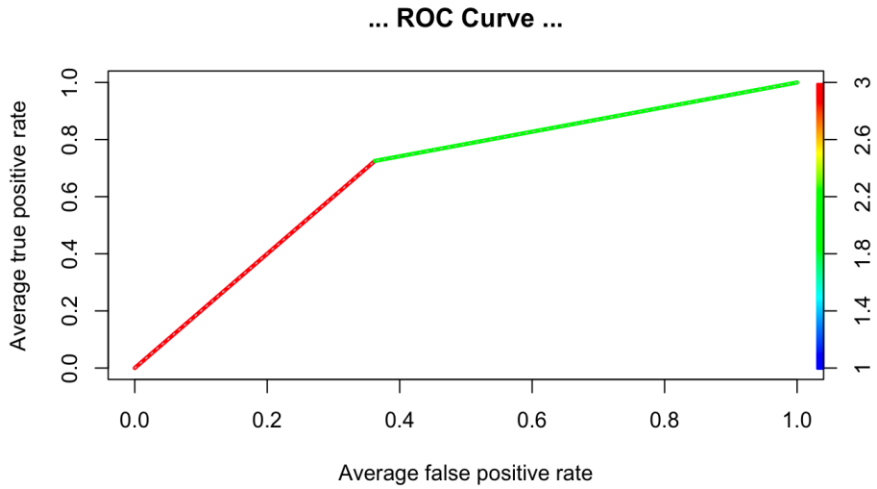
The same patterns are observed utilizing the K-Nearest Neighbors (KNN) model. The model's recall (63.7%) indicates that it struggled to identify a significant portion of the Overturned cases correctly, while its specificity (72.5%) suggests a notable improvement at identifying Upheld cases correctly.

Again, the precision value (69.0%) is higher than the recall value, but not significantly, with a 31.0% False Positive rate. However, the KNN model performed better than Naive Bayes, with an average performance increase of 3.5% across the accuracy metrics.

The overall accuracy of the model was 68.2%, with an F1 score of 66.3%.



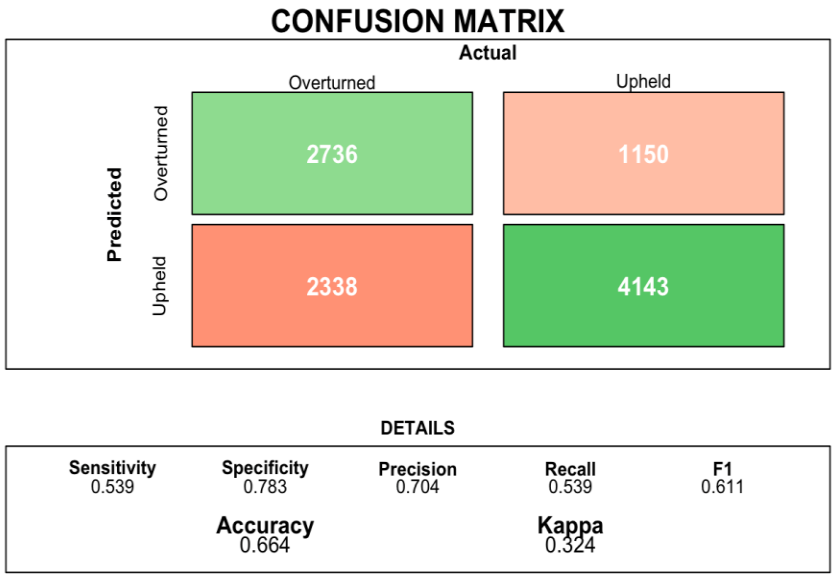
Like the KNN model, the ROC curve is almost linear indicating poor performance in classifying cases as Overturned.



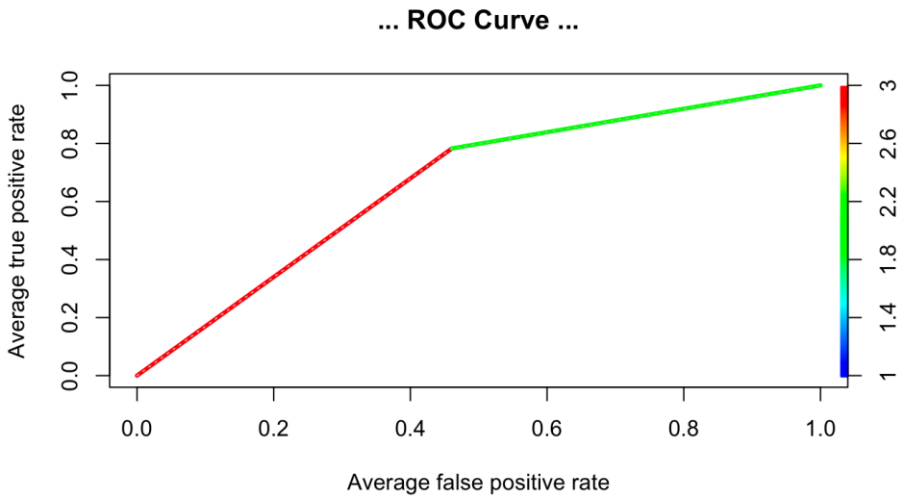
3.3: SVM

The same patterns are observed utilizing the Support Vectors Machine (SVM) model. The model's recall (53.9%) is the lowest compared to Naïve Bayes and KNN. However, this model achieved the highest accuracies in specificity at 78.3% and precision at 70.4% between the models.

The overall accuracy of the model was 66.4%, with an F1 score of 61.1%.

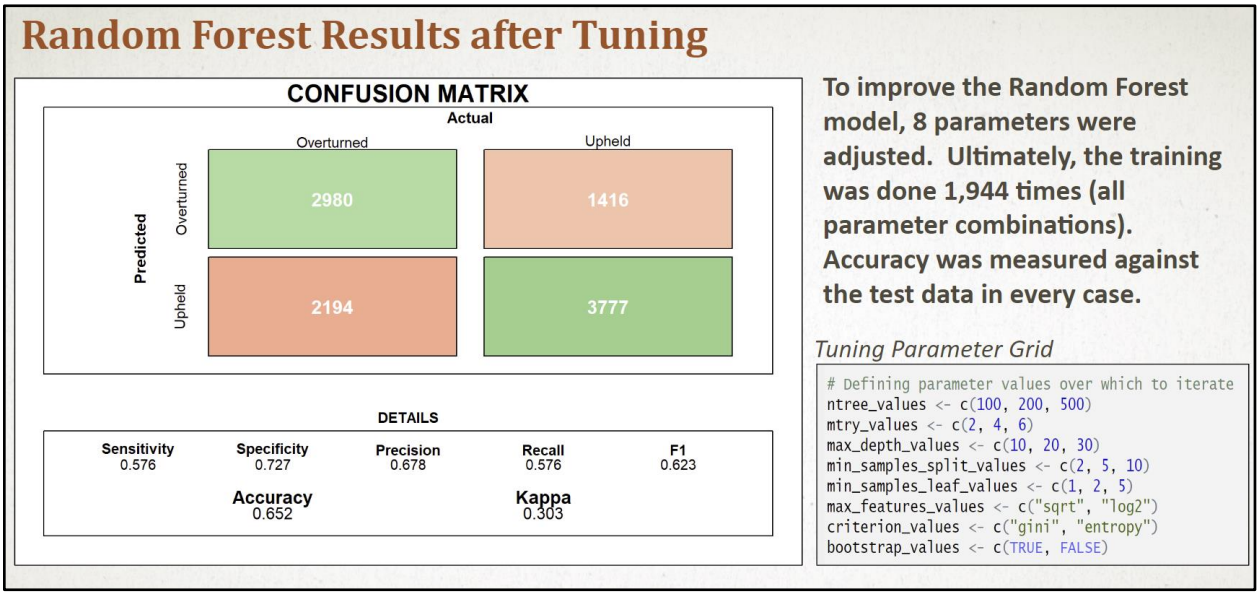


Again, the same curve is observed for SVM, indicating poor performance in classifying cases as Overturned.



3.4: Random Forest

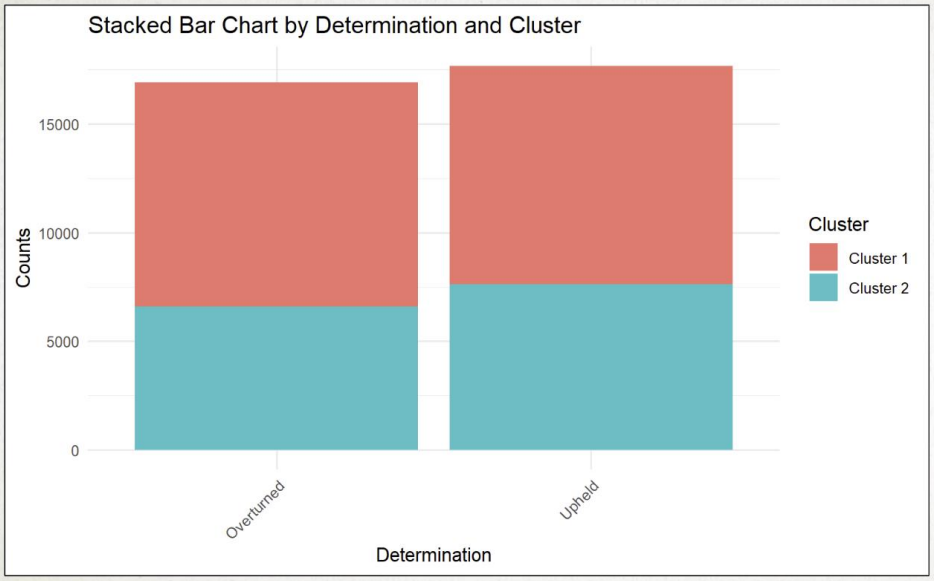
Random Forest modeling was turned via 1944 iterations given a parameter grid. The results were reasonable at 65.2% accuracy and a 62.3% F1, but not quite as significant as the KNN model. Noteworthy, running these 1944 iterations was computationally intensive, taking more than 6 hours to complete.



3.5: K-Means

K-means clustering did develop nice clusters as previously discussed, however, there was no significant relationship between the IMR Determination and the unsupervised clusters.

Clusters compared to IMR Determination



While the cluster analysis did reveal 2 distinct clusters, there is no relationship between these clusters and the IMR Determination.

3.6: Apriori

Findings:

Patients Aged 11-20 with Mental Disorder undergoing a treatment of Reconstruction/Plastic Procedure:

- Among patients with this characteristic, 97% had cases overturned.
- This suggests that patients with these characteristics (demographics) are more likely to cause self-harm /does not attain the age of employment and therefore should have their treatment plan covered by insurance.

```
##      lhs                                          rhs                                                    suppo
rt confidence    coverage     lift count
## [1]  {DiagnosisCategory=Mental Disorder,
##       TreatmentCategory=Reconstr/Plast Proc,
##       AgeRange=11 to 20}                      => {Determination=Overturned Decision of Health Plan} 0.0012133
47  0.9767442 0.001242236 1.996575    42
## [2]  {DiagnosisCategory=Mental Disorder,
##       TreatmentCategory=Reconstr/Plast Proc,
##       AgeRange=21 to 30}                      => {Determination=Overturned Decision of Health Plan} 0.0015022
39  0.8965517 0.001675574 1.832653    52
## [3]  {DiagnosisCategory=Mental Disorder,
##       TreatmentCategory=Reconstr/Plast Proc,
##       PatientGender=Male}                     => {Determination=Overturned Decision of Health Plan} 0.0011266
79  0.8863636 0.001271125 1.811827    39
## [4]  {DiagnosisCategory=Mental Disorder,
##       TreatmentCategory=Reconstr/Plast Proc,
##       AgeRange=21 to 30,
##       PatientGender=Female}                   => {Determination=Overturned Decision of Health Plan} 0.0012422
36  0.8775510 0.001415571 1.793813    43
## [5]  {DiagnosisCategory=Mental Disorder,
##       TreatmentCategory=Reconstr/Plast Proc,
##       AgeRange=31 to 40,
##       PatientGender=Female}                   => {Determination=Overturned Decision of Health Plan} 0.0012422
36  0.8431373 0.001473350 1.723467    43
```

Findings:

Female Patients Aged 51-64 conducting Mammography:

- Among patients with this characteristic, 97% had cases overturned.
- This suggests that patients with these characteristics (demographics) are more prone to breast cancer and hence to diagnose them at an early stage (if any) or rule out there is none , mammograms have to be conducted as a preventive measure and should have their treatment plan covered by insurance.

```
##       lhs                                                rhs                                                    suppo
rt confidence    coverage     lift count
## [1]   {DiagnosisSubCategory=Breast,
##        TreatmentSubCategory=Mammography,
##        AgeRange=51 to 64}                             => {Determination=Overturned Decision of Health Plan} 0.0011266
79  0.9750000 0.001155568 1.993010    39
## [2]   {DiagnosisSubCategory=Breast,
##        TreatmentSubCategory=Mammography,
##        AgeRange=51 to 64,
##        PatientGender=Female}                          => {Determination=Overturned Decision of Health Plan} 0.0010977
90  0.9743590 0.001126679 1.991699    38
## [3]   {TreatmentSubCategory=Facial Feminization Surgery,
##        PatientGender=Female}                          => {Determination=Overturned Decision of Health Plan} 0.0010400
12  0.9729730 0.001068901 1.988866    36
## [4]   {DiagnosisSubCategory=Gender Dysphoria,
##        TreatmentSubCategory=Facial Feminization Surgery,
##        PatientGender=Female}                          => {Determination=Overturned Decision of Health Plan} 0.0010400
12  0.9729730 0.001068901 1.988866    36
## [5]   {DiagnosisSubCategory=Obesity,
##        TreatmentSubCategory=Weight Control,
##        PatientGender=Female}                          => {Determination=Overturned Decision of Health Plan} 0.0010111
22  0.9722222 0.001040012 1.987332    35
```

Conclusions

The goal of this analysis was to evaluate the performance of several machine learning algorithms for the task of predicting the outcome of an Independent Medical Review (IMR) of a patient's denied, delayed, or modified health plan. Four different models were trained and tested: Naive Bayes, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest (RF). Additionally, K-Means Clustering and Association Rule mining was performed to identify patient characteristics that are more likely to have a case overturned.

The confusion matrix metrics suggest that our predictive models may not be performing optimally. The low recall could result in overlooking legitimate overturned cases, leading to patients or physicians not pursuing a review if the models were to predict probabilities. The moderate average precision of 68.9% between the models, indicates room for improvement in reducing false positives, which could lead to inefficiencies or unnecessary case reviews.

The Association Rules Mining suggest that when applied with a support metric of 0.001 and confidence metric of 0.8 there is a 97% chance of getting the decision getting overturned with the patients having the following attributes.This percentage of decisions getting overturned for the given patient attributes is supported by the fact that lift (strength of the association rule) is greater than one (1.99) suggesting that there is a strong positive correlation of patients with that characteristics would have the decisions overturned. This is a significant finding given that even though the volume of the records with those characteristics in the dataset is comparatively low the percentage of the number of cases where the decision is overturned is very high.

Patient Characteristics:

A.)

DiagnosisCategory=Mental Disorder,
TreatmentCategory=Reconstr/Plast Proc,
AgeRange=11 to 20

B)

DiagnosisSubCategory=Breast,
TreatmentSubCategory=Mammography
AgeRange=51 to 64
gender=Female

Final Summary Conclusions:

Conclusions

- Unsupervised Clustering developed reasonably "tight" clusters, however when married with the IMR decision (overturned/upheld), there was no correlation.
- Predictive modeling was an improvement! Overturned/upheld IMR decisions are 50/50 by chance. Best modeling has improved this by 18% (68% accuracy vs random).
  - The best modeling accuracy/confidence was the KNN model with an accuracy of 68.2% and F1 of 66.3%.
- The strong association rules accounted for 2221 (~7%) identified by Association Rule Mining offer a very high degree of confidence (80-90%) that the IMR will over turn the Insurance Company:
  - Confidence level (e.g. 80%), support level (e.g. .01)