Butler University
Digital Commons @ Butler University

Undergraduate Honors Thesis Collection | Undergraduate Scholarship

2019

Utilizing Multi-level Classification Techniques to Predict Adverse Drug Effects and Reactions

Victoria Puhl
Butler University

Follow this and additional works at: https://digitalcommons.butler.edu/ugtheses

Part of the Mathematics Commons

Recommended Citation

Puhl, Victoria, "Utilizing Multi-level Classification Techniques to Predict Adverse Drug Effects and Reactions" (2019). *Undergraduate Honors Thesis Collection*. 506.
https://digitalcommons.butler.edu/ugtheses/506

This Thesis is brought to you for free and open access by the Undergraduate Scholarship at Digital Commons @ Butler University. It has been accepted for inclusion in Undergraduate Honors Thesis Collection by an authorized administrator of Digital Commons @ Butler University. For more information, please contact digitalscholarship@butler.edu.

Butler University Honors Program
Honors Thesis Certification

Applicant	Victoria Puhl
Thesis Title	Utilizing Multi-level Classification Techniques to Predict Adverse Drug Effects and Reactions
Intended Date of Commencement	May 11, 2019

Read, Approved, and Signed by:

Thesis Adviser: .... Date 4/16/2019

Reader: William Johnston Date April 10, 2019

Certified by: .... Date ....

For Honors Program use:

Level of Honors conferred: University ____________________

Departmental ____________________

Utilizing Multi-level Classification Techniques to Predict Adverse Drug Effects and Reactions

Victoria Puhl

April 16, 2019

An Undergraduate Thesis

Presented to The Honors Program

of Butler University

Supervised by Dr. Rasitha Jayasekare

In Partial Fulfillment

of the Requirements for Graduation Honors

Acknowledgement of Sources

For all ideas taken from other sources (books, articles, internet), the source of the ideas is mentioned in the main text and fully referenced at the end of the report.

All material which is quoted essentially word-for-word from other sources is given in quotation marks and referenced.

Pictures and diagrams copied from the internet or other sources are labelled with a reference to the web page or book, article etc.

Signed Victoria Puhl Date April 16, 2019

Special thanks to the Center for High Achievement and Scholarly Engagement at Butler University and specifically to Dr. Jason Lantzer for their support regarding this thesis and for funds to attend the 2019 Joint Mathematics Meeting (JMM). Many thanks to Dr. William Johnston for his expertise and assistance throughout the completion of this paper.

Additionally, thanks to the Department of Mathematics, Actuarial Science, and Statistics at Butler University for the resources provided during the 2018 Mathematics Research Camp and for awarding myself with the 2019 Judith Morrel Award, given to a student who achieves academic excellence in an honors project or curriculum. Thank you to the JMM for the opportunity to present this thesis and for the Outstanding Poster Award, which provided support to complete this paper. Heartfelt thanks to Dr. Rasitha Jayasekare, the advisor for this thesis, for her constant encouragement and amazing commitment to the success of this project. It would not be complete without her.

ABSTRACT

UTILIZING MULTI-LEVEL CLASSIFICATION TECHNIQUES TO PREDICT ADVERSE DRUG EFFECTS AND REACTIONS

Victoria Puhl

April 16, 2019

Multi-class classification models are used to predict categorical response variables with more than two possible outcomes. A collection of multi-class classification techniques such as Multinomial Logistic Regression, Naïve Bayes, and Support Vector Machine is used in predicting patients' drug reactions and adverse drug effects based on patients' demographic and drug administration. The newly released 2018 data on drug reactions and adverse drug effects from U.S. Food and Drug Administration are tested with the models. The applicability of model evaluation measures such as sensitivity, specificity and prediction accuracy in multi-class settings, are also discussed.

Contents

Chapter 1

Introduction

Medical professionals work with multitudes of prescription drugs on a daily basis, and drug manufacturers constantly create new treatments. However as these different medications become used in medical practice, one cannot help but wonder about various unintended side effects these drugs may have upon patients. Many prescriptions often are successful when treating a patient's condition, but harmful reactions to treatments are common as well.

Analytics is a multifaceted and complex field that utilizes "mathematics, statistics, predictive modeling and machine-learning techniques to find meaningful patterns and knowledge in recorded data" (Pang-Ning Tan *et al.*, 2017). Through analytical reasoning and statistical methods, the multitudes of harmful reactions and hazardous events associated with certain prescription drugs are more thoroughly understood. This knowledge is helpful to both consumers and medical professionals.

1.1 Literature Review

The United States Food and Drug Administration's (FDA) Adverse Event Reporting System (FAERS) is a publicly available, national database utilized to monitor the safety of various drugs. Other statisticians have worked with the FAERS data as well as other medical databases to make predictions and obtain conclusions involving adverse drug effects. Due to the large amount of records available in the FAERS database, the FAERS data is widely used throughout the fields of statistics and data mining.

Unsupervised learning is a class of models where the data are mined without a predetermined response variable (the variable about which a researcher is attempting to explain). Common algorithms used in unsupervised learning included different methods called clustering, versions of neural networks, and method of moments. Patterns and underlying relationships in the data are then analyzed and uncovered after the unsupervised learning algorithm has been run. On the other hand, supervised learning algorithms analyze the training or input data and produce an inferred function that maps new examples to different values of the response variable. Classification is a supervised learning technique, which means it estimates a value (or class) from various input variables. It can also be defined as "a task of learning a target function f that maps each attribute x to one of the predefined class labels y" (Pang-Ning Tan *et al.*, 2017). Various statistical algorithms create this so-called 'target function,' which is also more commonly known as a classification model. Statisticians use both supervised and unsupervised learning to predict and understand FAERS or related healthcare data.

A group of researchers (Liu *et al.*, 2012) used support vector machine (SVM), a classification algorithm, to build a statistical model that would predict adverse drug reactions. Five machine learning algorithms - logistic regression, naïve Bayes, K-nearest neighbor, random forest, and SVM - were used in the study. The data used in this research came from the SIDER database, which "presents an aggregate of dispersed public information on side effects and indications" (Liu *et al.*, 2012). The group utilized multiple evaluation methods to determine the highest performing algorithm, and determined that the random forest and SVM algorithms created the most useful models. The researchers concluded that "different models that combine chemical, biological or phenotypic information can be built from approved drugs" and that the models have the ability to detect important adverse drug reactions (Liu *et al.*, 2012). This type of analysis is similar to the data investigation in this report; however, this study utilizes recent 2018 FAERS data and explores a different statistical model.

Harpaz *et al.*, (2012) have used unsupervised learning techniques (including association rules mining, biclustering, and logistic regression based techniques) in an attempt to identify new adverse drug events. The data sources in this study comprise spontaneous reporting systems (such as FAERS), healthcare databases, administrative claims, and chemical information sources. The group also analyzed harmful interactions between certain drugs.

Additionally, another data mining technique called Bayesian data mining, Dumouchel (1999) used to uncover patterns in the FAERS database. His goal was to identify frequent combinations of drugs and their adverse effects on patients. He utilized an empirical Bayes model to handle that large amount of data records.

A different research team looked at the FAERS database with a more specific objective in mind. They focused on data resulting in a certain diagnosis called torsades de pointes (TdP). TdP is described as "an uncommon type of ventricular tachycardia, or disturbance of the heart's rhythm" (Roland, 2017). It can lead to sudden cardiac arrest and is considered quite dangerous. The researchers wondered which drug interactions were leading to a higher frequency of drug-induced TdP diagnoses. They found that specifically antibacterial, antidepressant, and antipsychotic drugs were associated with the largest number of TdP event reports in the FAERS database (Poluzzi *et al.*, 2009). They utilized data mining and classification techniques to uncover this pattern between the above mentioned types of drugs and their adverse effects.

Researchers also used clustering (another unsupervised learning technique) with the goal of "identifying clinically relevant multimorbidity group" and uncovering relationships between chronic diseases and prescription drug treatments.(Cornell *et al.*, 2007). The data were provided from a single health care system rather than a large public database, but the study focused on recognizing patients with multiple chronic illnesses and prescription drugs that may have brought on those harmful diseases.

Many research studies have been completed revolving around the issue of adverse outcomes of drugs that uses the FAERS database or other publicly or privately obtained data. Classification and other data mining techniques were heavily used throughout the statistics field to uncover helpful information and create prediction models. However, the statistical research reports that involve prescription drugs vary in a numerous amount of ways. Some studies focus on specific diagnoses and reactions, while others attempt to discover certain patterns regarding the harmful side effects.

1.2 Various Multi-level Classification Models

In order to create a predictive model, various classification algorithms have to be created using a training dataset. The classification algorithms used in this project will be discussed below.

1.2.1 Logistic and Multinomial Regression

A generalized linear model (GLM) is a generalized form of the classical linear model, where the response variable Y takes one of the distributions of the exponential family (Pang-Ning Tan *et al.*, 2017). If the goal is to estimate Y based on k explanatory variables, then the population regression model is

$$Y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \epsilon, \tag{1.1}$$

and the estimated regression model is

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_k x_k \tag{1.2}$$

A GLM requires three important components. First (as stated above) it requires that that the response variable follows a distribution that is part of the 'Exponential Family' such as Normal, Poisson, Geometric, Gamma, and various others. Additionally, a GLM must contain a linear combination of predictor variables modeled as

$$n = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k \tag{1.3}$$

A GLM must also have a link function, $g(.)$, which defines the relationship between the linear predictor n and the mean value parameter (Pang-Ning Tan *et al.*, 2017):

$$g(p) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k,\ where\ E(Y) = p. \tag{1.4}$$

When performing logistic regression, the most frequently used link function is the "logit" or "log-odds" function is used to calculate a probability P:

$$logit(p) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k; \tag{1.5}$$

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \epsilon; and \tag{1.6}$$

$$P = \frac{e^{\beta_0 + \beta_1 x_1 + ... + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + ... + \beta_k x_k}} + \epsilon. \tag{1.7}$$

With respect to this dataset, the response variables (hazardous events and harmful reactions) are not considered binary as they both take multiple values that are not just 0 or 1. If records need to be classified into more than two classes, multiclass classification applies. Instead of categorizing observations into two classes, multi-level classification algorithms classify records into a finite number of known values.

Multinomial logistic regression is an extension of binary logistic regression that allows for more than two categories of the dependent or outcome variable. In multinomial logistic regression, "the log odds of the outcomes are model as a linear combination of the predictor variable" (Pang-Ning Tan *et al.*, 2017). When working with multiclass data, it is easiest to designate one of the response classes as a baseline, or the jth class. The random variable Y_i can be one of the predefined discrete values and the probability of an observation in the jth class is expressed as:

$$P_{ij} = P(Y_i = j) \tag{1.8}$$

The multinomial logit model assumes that the log-odds of each response follow a linear model

$$log\left(\frac{P_{ij}}{P_{iJ}}\right) = \beta_{0j} + \beta_{1j}x_1 + ... + \beta_{kj}x_k + \epsilon. \tag{1.9}$$

1.2.2 Naïve-Bayesian Classification

The next classification method discussed is based on the Bayes' theorem, which describes the probability of an event, based on prior knowledge related to another event. Bayes' in stated mathematically below, where *A* and *B* are events and the *P(B)* is not equal to 0.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1.10}$$

With a few assumptions, Bayes' theorem expands into the "Naïve Bayesian algorithm". The most important assumption is to presume that each attribute is conditionally independent of other attributes, given a class label (Pang-Ning Tan *et al.*, 2017).

Naive Bayesian classification is a fast algorithm to implement, and works well with large amounts of data. It is powerful and useful when the dataset has a large amount of noise points as these points are balanced out when calculating conditional probabilities. However, a large disadvantage of this classification technique is that any correlated predictor variables can reduce the performance, due to an assumption of conditional independence. This classification method is commonly used when in e-mail spam detection, facial recognition software, medical diagnoses, and weather predictions.

1.2.3 Support Vector Machine

Support Vector Machine (SVM) classifies via a separating hyperplane. It is "one of the most widely used classification algorithms" (Pang-Ning Tan *et al.*, 2017). SVM constructs a hyperplane or set of hyperplanes in a high dimensional space. Given labeled training data, the algorithm outputs an optimal hyperplane, which categorizes new inputs. It divides the observations of the separate categories with a clear gap – known as the large-margin decision boundary. Then it, maps and predicts new observations to belong to a category depending on which side of the gap they fall into. The SVM algorithm can be used for multiclass classification. The single multiclass problem reduces into multiple binary classification problems.

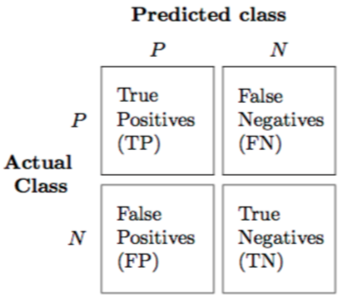
The SVM algorithm has many advantages, which contribute to its popularity. It is based on sound mathematical theory and works well with fewer training samples, since the number of support vectors does not matter too much. This classification technique can also be applied to categorial data through the introduction of dummy variables for each categorical value present in the dataset (Pang-Ning Tan *et al.*, 2017). Additionally, the prediction accuracy of SVM is generally high, and the algorithm usually avoids overfitting, which is the production of a model that corresponds too exactly to a certain dataset and may fail to predict future records correctly. One large disadvantage of SVM is its long training time.

1.3 Evaluation of Classification Models

After fitting the classifications models via the training dataset, the accuracy of the models needs to be evaluated. There are multiple ways to assess the performance of each model and to see if it has the ability to correctly classify new records. These evaluation methods will be discussed below.

1.3.1 Prediction Accuracy

The performance of a model is centered around the number of testing records that are accurately and inaccurately classified by the predictive model (Pang-Ning Tan *et al.*, 2017). The different counts are grouped into a table, which is more commonly as a confusion matrix. Each row of the matrix represents instances in a predicted class, and each column represents the records in the actual class.



		Predicted class	
		P	*N*
Actual Class	*P*	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

Figure 2: Confusion Matrix
Source. Raschka, 2018.

A confusion matrix (as visually represented above) consists of four different measures that can be identified from a table. True Positive Rate (TPR) represents the proportion of actual positives that are correctly identified as such. TPR is also known as the "sensitivity" of a model. False Negative Rate (FNR) measures the proportion of positives which are wrongly identified. False Positive Rate (FPR) represents the proportion of negatives that are wrong identified. True Negative Rate measures the proportion of actual negatives that are correctly identified as such. TNR is also known as the "specificity" of a model. These calculations are very important when analyzing the accuracy of a classification model.

Additionally, there are multiple other measures that can be computed using the calculations discussed above. Accuracy is known as the percentage of correct predictions by a statistical model and is commonly used to compare model performance. A prediction accuracy of above 80% can be interpreted that the model is particularly useful in prediction the response variable. The calculation for prediction accuracy is shown below.

$$Accuracy = \frac{TPR + TNR}{SampleSize} * 100 \tag{1.11}$$

1.3.2 Receiver Operating Characteristics Curve (ROC) and Area Under the Curve (AUC) Value

Another method of evaluating and comparing overall model performance is through observing a model's Receiver Operating Characteristics (ROC) curve. A model's ROC curve is a very useful way to view how predictive models distinguish between true positive and negative classifications. This graphical approach to comparing classification models is created by plotting *sensitivity* also known as TPR against *1-specificity* also known as FPR. Both of these measures in Section 1.3.1. In a visual display, the FPR is typically plotted along the x-axis while the TPR is shown on the y-axis (Pang-Ning Tan *et al.*, 2017).

The most useful classifier will have an ROC curve that displays a high measure of *sensitivity* and a low measure of *1-specificity.* This produces a curve that is close to the upper left corner of the ROC curve graph. lf a classifier produces a ROC curve that looks like a straight line across the diagonal of the graph, it is making random guesses as its predictions and is considered not useful. Because ROC curves are graphical displays, they are very useful for comparing the performance of different classification models (Pang-Ning Tan *et al.*, 2017).

Another performance measure associated with the ROC curve is known as the area under the the ROC curve or AUC value. The AUC value measures the two-dimensional area underneath the entire ROC curve. A model that predicts the response variable perfectly would have an AUC value equal to 1. A model that randomly predicts the response variable would have an AUC value of 0.5. When comparing model performance, the model with the highest AUC value is the most accurate and useful in predicting the response variable while the model with the lowest AUC value is the least accurate predictive model (Pang-Ning Tan *et al.*, 2017).

1.4 Objective

The FAERS database consists of adverse event reports, product quality complaints, and medication error reports that result in harmful reactions (AERS, 2018). Lawyers, physicians, drug manufacturers, and even patients are able to submit reports concerning irregular side effects experienced during pharmaceutical drug treatments. Healthcare professionals also may report a harmful event directly to the individual drug manufacturer. If a manufacturer receives a serious report, the company is mandated to send the report directly to the FDA.

While monitoring recurrent patient safety concerns regarding medical drugs, the FAERS database also aids the FDA in determining whether or not to take certain regulatory actions to protect public health, such as restricting the use of a certain drug or updating a product's label information. The FDA also utilizes the FAERS to evaluate certain manufacturers' compliance with reporting serious submissions.

This thesis further explores the FAERS database and applies various multi-level classification methods with the goal of answering the following questions:

i. Will multilevel classification techniques create a statistical model that predicts hazardous events caused by certain prescription drugs?

ii. Will multilevel classification techniques create a statistical model that predicts harmful reactions to certain prescription drugs?

iii. If so, which multilevel classification model is the most accurate, and why?

As discussed earlier, various other statisticians have worked with and analyzed the data provided from the FAERS and other similar reporting sites. These researchers used both supervised and unsupervised learning techniques to predict certain side effects of prescription drugs. The goal of these studies was either to predict a certain response variable or to uncover important relationships between prescription drugs and mortality rates or unwanted reactions. A number of research projects also focused on patterns between a specific type of prescription drug and a specific disease.

This project solely uses supervised learning, specifically multi-level classification, to predict two different response variables of interest: hazardous events and harmful reactions. Both of these are responses to prescription drugs. The data used in this study is very recent data from the second quarter of 2018, while other similar projects have used much older data. Medical records change over time as medications are always being updated. Additionally, different classification techniques are compared using multiple performances measures in order to analyze which predictive model is the most useful. No published studies have utilized the multinomial logistic regression to predict the respective response variable. This thesis determines which classification model is the most effective in predicting hazardous events and reactions that occurs due to prescription drug usage.

1.5 Chapter Outline

As mentioned, this paper will employ three statistical classification techniques to create prediction models. They are Multinomial Logistic Regression, Naïve Bayes, and Support Vector Machine. The paper then compares and contrasts their effectiveness at each step of their use.

The paper will further describe in detail in regards to its complexity. Multiple visual representations of the data will be provided. Additionally, the paper will discuss the process of data cleaning, sampling, dimension reduction, and data visualization. It will also compare the results of the various classification techniques along with the accuracy of those statistical models. Finally, the paper will present the overall findings will be presented and thoroughly compare the results.

Chapter 2

Data Analysis

The dataset used in this project is the 2018 Second Quarterly Files from the Federal Adverse Event Reporting System which is also known as FAERS. The data are very complex and consists of many different variables. Due to the complexity of the data, pre-processing, which is the act of converting the original, complex input data into a useable form for later examination, is essential (Pang-Ning Tan *et al.*, 2017). All computer programming was done in R, a software package for statistical computing.

2.1 Data Cleaning

The FDA provides different data files that relate to each quarter of the year. The seven various data files for this paper are comprised of information grouped around a specific topic including the following:

- Therapy - contains drug therapy start dates and end dates for the reported drug
- Indication - contains all Medical Dictionary for Regulatory Activities (MedDRA) terms coded for the diagnoses for the reported drugs
- Drug - contains drug information for as many medications as were reported for the vent
- Demo - contains patient demographic and administrative information
- Reaction - contains all MedRa terms coded for the adverse reaction
- Outcome - contains patient outcomes for the event
- Report Sources - contains report sources for the event

Each data file contains different sophisticated information and all of the various files are linked together through a primary key of "primaryid". This relationship between the seven data files and their respective variables can be seen below in the data diagram.

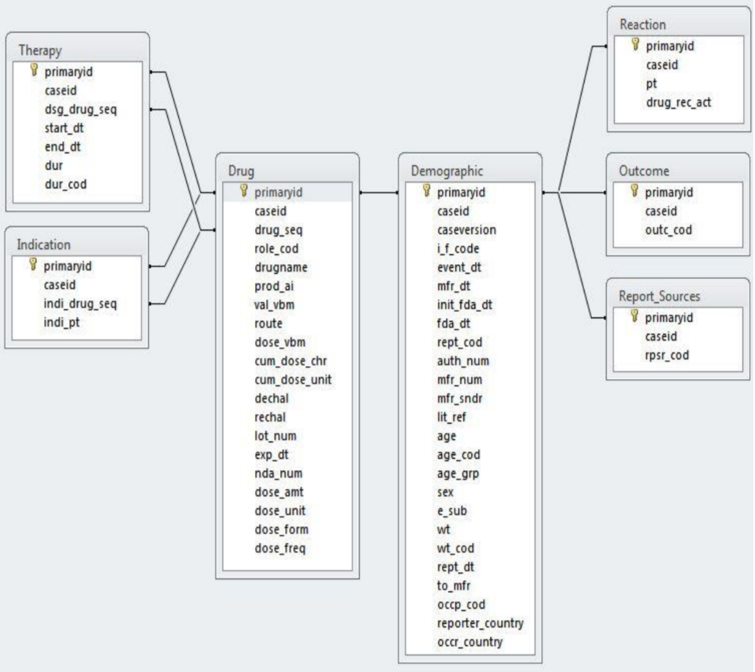


Figure 1: Data Diagram
Source. AERS, 2018.

In this report, the data files consisting of drug information, demographic information, reaction information and outcome information are of interest. These four different data files were merged together using a primary key of "primaryid" and a secondary key of "caseid" since there may have been multiple reports filed under the same primary identification code. After merging the data files together, the new dataset contained 38 variables and 1,913,806 records.

The following variables were then removed due to duplicity, added complexity, large amount of unknown values, or insignificance with respect to the statistical goal of predicting hazardous events or reactions.

- I_F_cod - code for initial or follow-up status of report
- event_dt - date the adverse event occurred or began

- mfr_dt - date manufacturer received initial information
- init_FDA_dt - date FDA received first version of the case
- fda_dt - the latest manufacturer received date
- rept_dt - date report was sent
- rept_cod - code for the type of report submitted
- auth_num - regulatory authority's case report number
- mfr_num - manufacturer's unique report identifier
- lit_ref - literature reference information
- age_cod - unit abbreviation for patient's age
- age_grp - patient's age group code
- e_sub - whether the report was submitted electronically
- wt - numeric value of patient's weight
- wt_code - unit abbreviation for patient's weight
- to_mfr - whether the voluntary reporter also notified manufacturer
- reporter_country - the country of the reporter in the latest version of the case
- occp_cod - abbreviation for the reporter's type of occupation
- drug_seq - the unique number for identifying a drug for a case
- val_vbm - code for source of drug name
- dose_vbm - verbatim text for dose, frequency, and route, exactly as entered on report
- cum_dose_chr - cumulative dose to first reaction
- NDA_num - NDA number
- dechal - Dechallenge code, indicating if reaction abated when drug therapy was stopped
- dose_form - form of dose reported
- drug_rec_act - if the event reappears upon re-administration of the drug

When examining the variables, two in particular were chosen as response variables - “outcome” and “reaction.” Overall, variables with date formats were hard to incorporate into the statistical models, as many records had incomplete dates or the date formats varied. Due to this, all variables with date formats were removed, as shown above. Additionally, age group was removed as the variable “age” was subsetted so only ages in terms of years were present. Another variable – “route” – is described as the route of drug administration, which is similar to “dose form,” and therefore “dose form” was removed due to duplicity. Weight was a variable that would have been important to include in the statistical models; however, there were too many missing values to confidently include it in the training models. Dosage amount was subsetted to avoid complex conversions and only included drugs measured in milligrams.

After this variable reduction, twelve potentially useful variables (other than the response variables of “outcome” and “reaction”) remained. Further techniques of dimension reduction will be applied in an attempt to identify the most informative variables to create the classification models used to predict hazardous events and reactions to certain prescription drugs .

2.2 Creating Train and Test Datasets

Before decreasing the amount of variables in the dataset, the overall size of the dataset must be reduced in order to perform statistical algorithms in a timely manner. Due to computer and time restrictions, the entire dataset of around 2 million records could not be used. The number of records in the dataset had to be reduced before continuing on with creating a model.

When using prediction models, a dataset needs to be split into a “training” dataset and a “testing” dataset. The training dataset is the sample of data used to fit or train the model. The testing dataset is the sample of data used to provide an evaluation of the model fitted on the training dataset. It is used to analyze how well the model performs. The training dataset usually consists of around 80% of the records in the dataset before it is split while the testing data sets consists of the other 20% of the records.

To accommodate for the available computational power, the dataset was reduced to 50,000 records. The 50,000 records were randomly sampled from the original merged records. A sample of 50,000 records was obtained for the model predicting hazardous events. The dataset consisting of those 50,000 records was then split into separate training and testing datasets. The training dataset for fitting the hazardous events model consisted of 40,000 records while the testing dataset consists of 9,596 records. The testing dataset did not consist of exactly 10,000, as there were some variable factors that existed in the testing dataset and not the training dataset. These variable factors had to be removed since the model developed from the training data would not know how to predict a factor not present in fitting the modell.

When creating training and testing datasets for the model that predict harmful reactions, the process was slightly different than simply randomly sampling 50,000 records. The response variable (reaction) took around 2,000 different values. When testing the statistical model, it would have a difficult time accurately predicting so many similar reactions (i.e. headache vs. migraine). To take into account this problem, the most frequently occurring 15 reactions were identified and a reduced dataset was created from only selecting those top 15 reactions. These various reactions will be listed further along in this section. This subsetted dataset consisted of 51,316 records. This data was then randomly sampled to consist of 50,000 records and the split to contain around 80% of records in the training dataset and 20% of records in the testing dataset. The training dataset for fitting the harmful reactions model consisted of 40,000 records while the testing dataset consisted of 8,656 records. As stated above, this testing dataset also did not contain exactly 10,000 records, as some variable factors existed in the testing dataset but not in the training dataset.

2.3 Dimension Reduction

When working with large and sophisticated datasets, it is important to identify the most useful variables and utilize them to create the various statistical models. This FAERS dataset is multifaceted, and even after removing trivial and redundant variables, twelve variables still remained. Through the use of dimension reduction, which decreases the high dimensionality of a dataset through various processes, one can pinpoint the most informative variables in the data.

A specific type of dimension reduction technique known as Principle Component Analysis (PCA) is one of the more common algorithms to make high dimensional data less complex. It is known as an unsupervised algorithm. In addition to being used for dimension reduction, PCA is also helpful to recognize the strongest patterns in a dataset and to eliminate noise points in the data. PCA measures data in terms of their principle components – uncorrelated variables that summarize the underlying structure of the data. PCA then generates new variables that are linear combinations of the original variables. The objective of PCA is "to explain the maximum amount of variance [in the data] with the fewest number of principal components" (Pang-Ning Tan *et al.*, 2017).

Usually applied to datasets with numerical values only, a modification of PCA can perform on datasets with both quantitative and qualitative variables, such as the FAERS dataset. This process is known as "mixed PCA," as the variables are both categorical and numerical. The PCA algorithm is also only performed on the training dataset, since that dataset fits the model. The unimportant variables identified from PCA are then removed from the corresponding testing dataset.

2.3.1 Dimension Reduction of Hazardous Events Dataset

To identify the most informative variables for creating the model, PCA must determine the number of principle components. PCA generates the proportion of variance that the components explain with respect to the number of principle components that will be used. When running PCA on the training dataset used to fit the hazardous events model, the proportion of variance that 12 principal components would explain was 76.90%. When determining the number of principle components used, a standard has at least 70% of the variance explained. Since the dataset only has 12 variables, at most 12 principal components are possible.

The PCA algorithm also generates eigenvalues associated with the number of principal components used. Eigenvalues are described as "variances of the principal components" and principal components should always have eigenvalues above 1.00 (Pang-Ning Tan *et al.*, 2017). The eigenvalue corresponding with 12 principle components was 4.01, showing that twelve principal components are sufficient; the eigenvalue is above 1.00.

PCA then produces a correlation matrix with squared correlation coefficients associated with variable and number of principal components. Generally, variables selected should have a squared loading value (SLV) above a certain threshold, to ensure that the variables chosen are the most useful when creating the model. The SLV chosen in this analysis is greater than or equal to

Below is the matrix with the chosen variables highlighted in yellow and the removed variables shown in red.

Training Data - Outcome

Number of observations	40000	
Number of variables:	12	
	Number of numerical variables:	2
	Number of categorial variables:	10

Squared Loadings:

	dim 1	dim 2	dim 3	dim 4	dim 5	dim 6	dim 7	dim 8	dim 9	dim 10	dim 11	dim 12
caseversion	0.24	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
age	0.00	0.03	0.01	0.01	0.02	0.02	0.03	0.02	0.03	0.02	0.00	0.00
role_cod	0.13	0.02	0.01	0.01	0.09	0.02	0.01	0.01	0.00	0.01	0.00	0.00
drugname	0.96	0.96	0.97	0.94	0.95	0.96	0.96	0.96	0.97	0.98	1.00	0.99
prod_ai	0.94	0.95	0.96	0.91	0.93	0.93	0.93	0.94	0.94	0.97	1.00	0.98
route	0.73	0.41	0.21	0.50	0.20	0.09	0.04	0.11	0.04	0.03	0.00	0.02
dose_amt	0.50	0.72	0.84	0.58	0.73	0.84	0.87	0.78	0.69	0.85	0.99	0.93
dose_freq	0.76	0.35	0.20	0.33	0.10	0.07	0.05	0.01	0.04	0.02	0.00	0.01
mfr_sndr	0.31	0.65	0.81	0.40	0.62	0.66	0.69	0.78	0.80	0.53	0.03	0.54
sex	0.02	0.00	0.00	0.02	0.03	0.02	0.01	0.01	0.00	0.00	0.00	0.00
occur_country	0.56	0.05	0.03	0.24	0.18	0.03	0.03	0.05	0.04	0.03	0.01	0.01
pt	0.39	0.42	0.52	0.38	0.37	0.44	0.46	0.39	0.47	0.58	0.97	0.53

The eight variables used to fit the model to predict hazardous events related to prescription drugs are as follows:

i. drugname - name of medicinal product

ii. prod_ai - product active ingredient

iii. route - route of drug administration (i.e. oral, intravenous, etc.)

iv. dose _amt - amount of dosage in milligrams

v. dose _freq - frequency of dosage amount (i.e. daily, weekly, etc).

vi. mfr _sndr - name of drug manufacturer sending report

vii. occr _country - country where event occurred

viii. pt - "preferred term" medical terminology describing the reaction

The response variable is coded as "outc _cod" and is the code for a hazardous event resulting from prescription drug usage. It can take the seven categorical values described below:

i. CA - Congenital Anomaly

ii. DE - Death

iii. DS - Disability

iv. HO - Hospitalization (Initial or Prolonged)

v. LT - Life-Threatening

vi. OT - Other Serious Medical Event

vii. RI - Required Intervention to Prevent Permanent Impairment

2.3.2 Dimension Reduction of Reaction Dataset

The same process to determine the number of principal components and informative variables must be repeated for the training dataset involving predicting harmful reactions with respect to prescription drug usage. The proportion of variance explained by 12 principal components is 89.23%, which is above 70%, which means these 12 principal components are acceptable. The eigenvalue relating to 12 principal components is 3.53, which is above 1.00. This also reinforces the decision to use 12 principal components to determine the informative variables to fit the model. There are only 12 variables currently in the pre-dimension reduction training dataset and one should not use more than the number of variables for the principal components. The SLV chosen in this analysis is again, greater than or equal to

Below is the matrix with the chosen variables highlighted in yellow and the removed variables shown in red.

Training Data - Reaction

Number of observations	40000		
Number of variables:	12		
	Number of numerical variables:	2	
	Number of categorial variables:	10	

Squared Loadings:

	dim 1	dim 2	dim 3	dim 4	dim 5	dim 6	dim 7	dim 8	dim 9	dim 10	dim 11	dim 12
caseversion	0.23	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.01	0.00	0.02
age	0.00	0.05	0.02	0.00	0.01	0.02	0.03	0.00	0.00	0.17	0.00	0.01
role_cod	0.12	0.03	0.01	0.00	0.06	0.03	0.02	0.00	0.00	0.02	0.10	0.02
out_cod	0.03	0.01	0.00	0.00	0.01	0.00	0.02	0.00	0.01	0.01	0.01	0.00
drugname	0.97	0.96	0.96	0.99	0.96	0.97	0.98	0.99	0.99	0.94	0.96	0.96
prod_ai	0.95	0.94	0.94	0.99	0.93	0.94	0.96	0.98	0.98	0.89	0.92	0.93
route	0.74	0.63	0.56	0.08	0.15	0.17	0.05	0.22	0.04	0.28	0.11	0.24
dose_amt	0.50	0.61	0.69	0.96	0.73	0.83	0.87	0.65	0.89	0.56	0.57	0.61
dose_freq	0.78	0.56	0.33	0.03	0.12	0.13	0.02	0.03	0.01	0.12	0.24	0.30
mfr_sndr	0.29	0.46	0.45	0.93	0.61	0.65	0.84	0.89	0.87	0.43	0.51	0.35
sex	0.02	0.01	0.00	0.00	0.04	0.02	0.01	0.00	0.00	0.02	0.03	0.00
occur_country	0.55	0.07	0.08	0.02	0.31	0.07	0.04	0.03	0.01	0.20	0.13	0.10

The seven variables used to fit the model to predict harmful reactions related to prescription drugs. They are as follows:

i. drugname - name of medicinal product

ii. prod_ai - product active ingredient

iii. route - route of drug administration (i.e. oral, intravenous, etc.)

iv. dose _amt - amount of dosage in milligrams

v. dose _freq - frequency of dosage amount (i.e. daily, weekly, etc).

vi. mfr _sndr - name of drug manufacturer sending report

vii. occr _country - country where event occurred

The response variable is coded as “pt” and is the “preferred term” medical terminology describing the reaction. As mentioned earlier, in the initial dataset, this “pt” variable could take around 2,000 different values. To improve the accuracy of the model, it used only the most frequently occurring 15 reactions. These different reactions, coded numerically in alphabetical order, are described below:

i. Anemia (1) - condition in which the blood does not have enough healthy red blood cells or hemoglobin

ii. Asthenia (2) - unusual weakness or lack of energy

iii. Cough (3) - the act of expelling air from lungs suddenly

iv. Diarrhea (4) - condition in which loose / liquid stools are passed frequently

v. Dyspnea (5) - labored breathing

vi. Fall (6) - the act of collapsing or losing one’s balance

vii. Fatigue (7) - extreme tiredness

viii. Headache (8) - continuous head pain

ix. Kidney Injury (9)- sudden episode of kidney failure or kidney damage

x. Malaise (10) - feeling of illness or uneasiness

xi. Nausea (11) - an inclination to vomit

xii. Pain (12) - physical suffering caused by injury or illness

xiii. Pneumonia (13) - condition in which lungs are inflamed due to an infection

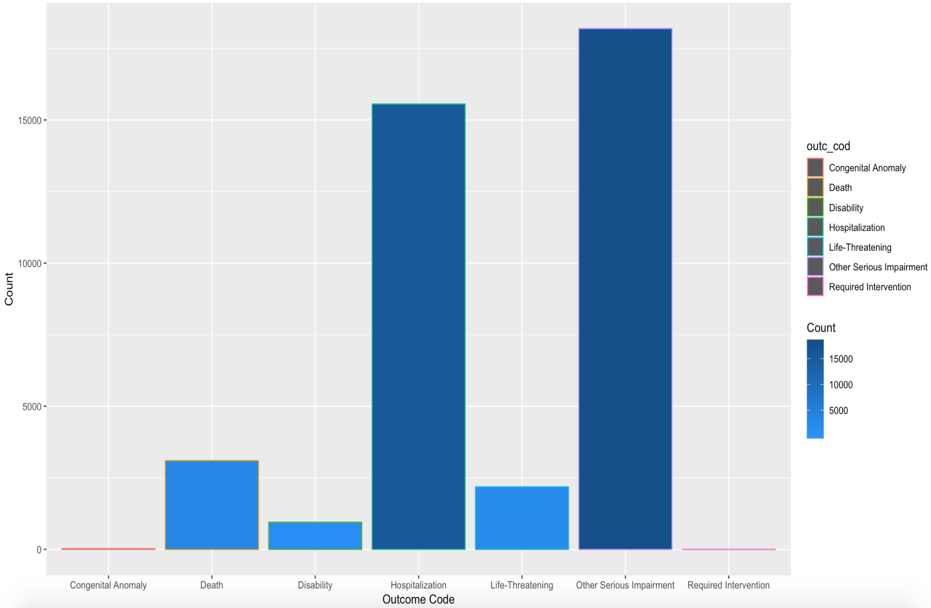
xiv. Pyrexia (14) - also known as fever, increased body temperature

xv. Vomiting (15) - the act of expelling matter from the stomach through the mouth

2.4 Visualization

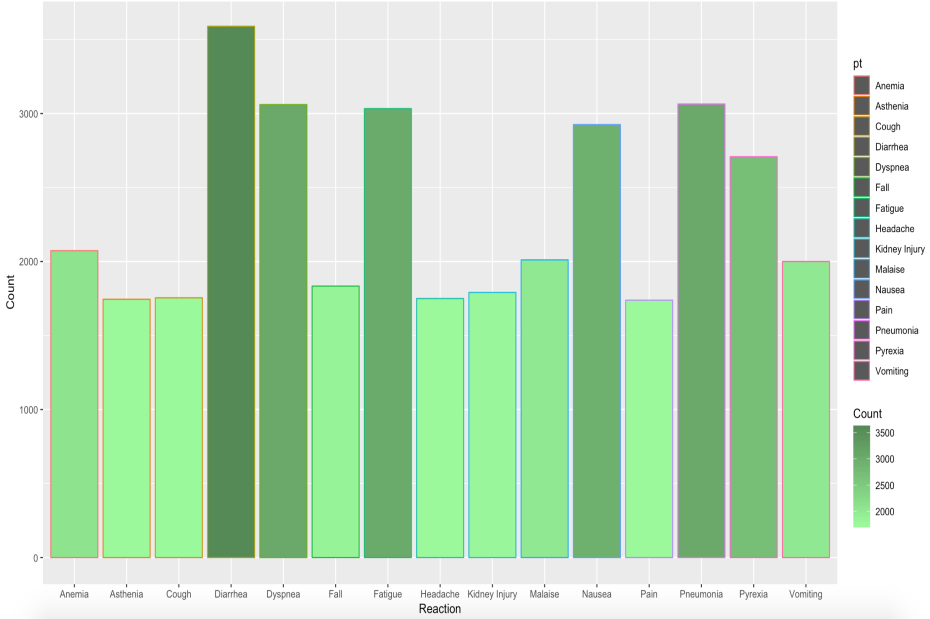
After determining the most informative and useful variables to create the predictive models, it is helpful to visually analyze the dataset. Visual aids can help to understand the dataset's distribution and any existing relationships in the data. A certain package in R, known as ggplot2, is widely used and creates elegant and complex plots that aid in summarizing data. The graphing package is based on Leland Wilkinson's "Grammar of Graphics" which is a "method of independently specifying plot building blocks and combining them to create just about any kind of graphical display" (Pang-Ning Tan *et al.*, 2017). Ggplot2 is an effective way to produce colorful, informative, and quality visual aids.

A bar-chart displays the frequency counts of different categories of hazardous events in the respective training dataset. Many hazardous events resulted in other serious impairments that probably could not be placed into any of the other categories. This may be why this category has the highest frequency of records, while the other categories are more specific. In comparison, the hazardous events of congenital anomaly and required intervention have very low occurrences while death, disability, and life-threatening impairment all havesimilar counts ranging from around 1,000 to 3000 reported occurrences.



Another bar-chart displays the frequency counts of different categories of reactions in the respective training dataset. These occurrences are more evenly distributed than the different outcomes shown above. The most common reaction is diarrhea with around 3,600 occurrences. Diarrhea is a very common side effect. In fact, almost all prescription drugs may have diarrhea as a drug-induced reaction (MedlinePlus, 2019). This may be why it is the most commonly reported reaction in this FAERS dataset.The next most frequent reactions appear to be dyspnea, pneumonia, and fatigue, all having around 3,000 occurrences. The least common reactions are asthenia, cough, headache, and general pain, all having around 1,700 occurrences.

It is interesting that these occurrences of these 15 different reactions are somewhat evenly distributed, most ranging from around 1,700 to 3,600 occurrences, a much closer spread than seen in the frequency distribution of outcomes which range from around 10 to 18,000 occurrences. A possible explanation could hypothesize that these reactions are considered more similar to each other than the outcomes codes. For example, nausea and vomiting are quite similar along with dyspnea and fatigue, which are often confused with each other.



Chapter 3

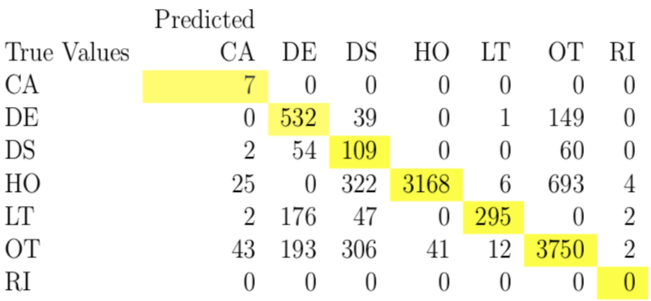
Multiclass Classification

Various classification algorithms need to be utilized to train the predictive multiclass model. The models used in this project to predict hazardous events and reactions observed in patients as a response to prescription drug usage include: Multinomial Logistic Regression, Naïve Bayesian, and Support Vector Machine algorithms. Chapter 1 describes these algorithms. The models all predict the given response variable at a varying level of accurateness. This is due to the underlying assumptions and nature of the algorithms. This chapter compares the models with respect to three measures – prediction accuracy, receiving operator characteristics (ROC) curve, and area under the curve (AUC) value, which Chapter 1 also defines.

3.1 Classification of Hazardous Events

The Naïve Bayesian, Support Vector Machine, and Multinomial Logistic Regression algorithms each create different classification models to predict the hazardous event, the response variable. They each use using the following predictor variables: drug name, drug active ingredient, route of drug administration, dosage amount, dosage frequency, drug manufacturer, occurrence country, and patient reaction. These different classification models can be evaluated and compared using prediction accuracy, their respective ROC curve, and corresponding AUC value.

Pictured below is the confusion matrix for the Naïve Bayesian model, which predicts patient outcome codes. The display highlights the correctly classified records in yellow.



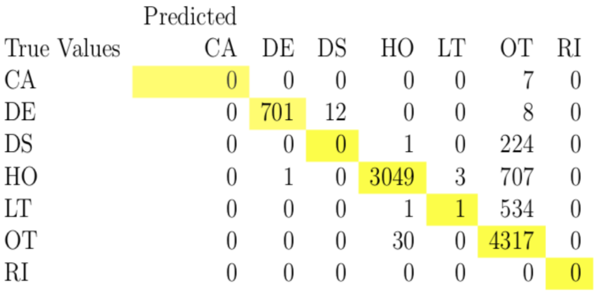
	Predicted						
True Values	CA	DE	DS	HO	LT	OT	RI
CA	7	0	0	0	0	0	0
DE	0	532	39	0	1	149	0
DS	2	54	109	0	0	60	0
HO	25	0	322	3168	6	693	4
LT	2	176	47	0	295	0	2
OT	43	193	306	41	12	3750	2
RI	0	0	0	0	0	0	0

To determine the prediction accuracy of the Naïve Bayesian model, simply perform this calculation:

$$Accuracy = \frac{7 + 532 + 109 + 3168 + 295 + 3750 + 0}{9596} * 100 = 81.92\% \quad (3.1)$$

This prediction accuracy means the Naïve Bayesian model predicts patient outcomes correctly 81.92% of the time. This percentage is above 80%, showing that this model is useful to predict the given response variable.

Pictured below is the confusion matrix for the Support Vector Machine model that predicts patient outcome codes. The display highlights correctly classified records in yellow.



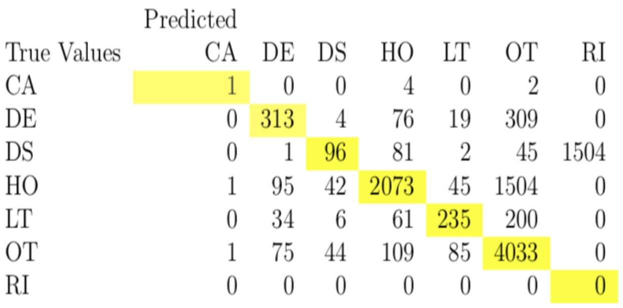
	Predicted						
True Values	CA	DE	DS	HO	LT	OT	RI
CA	0	0	0	0	0	7	0
DE	0	701	12	0	0	8	0
DS	0	0	0	1	0	224	0
HO	0	1	0	3049	3	707	0
LT	0	0	0	1	1	534	0
OT	0	0	0	30	0	4317	0
RI	0	0	0	0	0	0	0

To determine the prediction accuracy of the SVM model, calculate:

$$Accuracy = \frac{0 + 701 + 0 + 3049 + 1 + 4317 + 0}{9596} * 100 = 84.07\% \quad (3.2)$$

This prediction accuracy means the SVM model predicts patient outcomes correctly 84.07% of the time. This percentage is above 80%, also showing this model is useful to predict the given response variable. However, the SVM prediction accuracy is higher than that of the Naïve Bayesian prediction accuracy, indicating SVM as a more accurate and useful classification model than Naïve Bayesian.

Pictured below is the confusion matrix for the Multinomial Regression model that predicts patient outcome codes. It highlights correctly classified records yellow.



	Predicted						
True Values	CA	DE	DS	HO	LT	OT	RI
CA	1	0	0	4	0	2	0
DE	0	313	4	76	19	309	0
DS	0	1	96	81	2	45	1504
HO	1	95	42	2073	45	1504	0
LT	0	34	6	61	235	200	0
OT	1	75	44	109	85	4033	0
RI	0	0	0	0	0	0	0

To determine the prediction accuracy of the Multinomial Regression model, calculate:

$$Accuracy = \frac{1 + 313 + 96 + 2073 + 235 + 4033 + 0}{9596} * 100 = 70.35\% \qquad (3.3)$$

This prediction accuracy means that the Multinomial Regression model predicts patient outcomes correctly 70.35% of the time. This percentage is below 80%, which says this model is not particularly useful to predict the given response variable. Additionally, the Multinomial Regression prediction accuracy is lower than that of the Naïve Bayesian and SVM prediction accuracy, making Multinomial Regression the least accurate and useful classification model for predicting patient outcomes discussed so far.

Shown below are the ROC curves associated with the models predicting patient outcomes in the following order: Support Vector Machine, Naïve Bayesian, and Multinomial Regression.

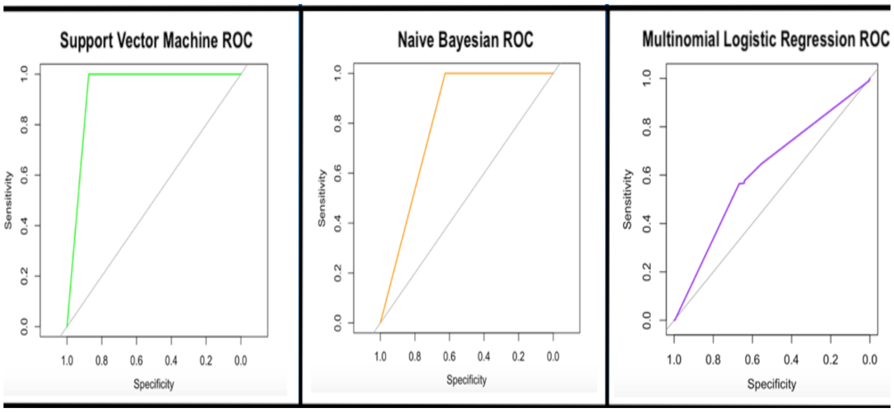


Figure 3: Left to right: Best to worst performance of multiclass classification.

The ROC curve produced by the Support Vector Machine model is well above the diagonal line in the center of the graph. Therefore, this classifier is not randomly predicting all patient outcomes. The AUC value associated with this model is 0.9015, which is much closer to 1.00 than 0.50. This means that the model is rarely performing random predictions. The model has a high TPR and a low FPR, resulting in the high AUC value and an ROC curve close to the top left corner of the graph.

The ROC curve produced by the Naïve Bayesian model is above the diagonal line in the center of the graph, similar to the SVM ROC curve. The AUC value associated with this model is 0.7332, which is closer to 1.00 than 0.50, meaning that the model performs accurate classifications more than inaccurate classifications. This AUC value is lower than the AUC value associated with the Support Vector Machine, making the latter model a more accurate classification model for predicting patient outcomes.

The ROC curve produced by the Multinomial Regression model is slightly above the diagonal line in the center of the graph. The AUC value associated with this model is 0.6045, which is closer to 0.50 than 1.00. This means that more often than not, the model is performing random predictions rather than using the predictor variables to accurately predict the patient's outcome.

Overall, Support Vector Machine's AUC value is the closest out of the three model's AUC values to 1.00, making it the most accurate model when predicting patient outcomes. Additionally, the peak of the SVM's ROC curve is closest to the top left-hand corner of the graph out of the three ROC curves.

A table summarizes the prediction accuracy and AUC values of the various classification models that predicts patient outcomes. The algorithms are listed in the order of highest performing model to lowest performing model.

Algorithm	**Prediction Accuracy**	**AUC Value**
Support Vector Machine	84.08%	0.9015
Naïve Bayesian	81.92%	0.7332
Multinomial Regression	70.35%	0.6045

3.2 Classification of Harmful Reactions

The Naïve Bayesian, Support Vector Machine, and Multinomial Logistic Regression algorithms are all useful to create different classification models to predict the harmful reaction (the response variable) using the following predictor variables – drug name, drug active ingredient, route of drug administration, dosage amount, dosage frequency, drug manufacturer, and occurrence country. These different classification models can be evaluated and compared in terms of their prediction accuracies, their respective ROC curves, and their corresponding AUC values.

After discussing the classification models estimating patient outcomes, one can now analyze the confusion matrices produced by testing the models predicting patient reactions. Pictured below is the confusion matrix for the Naïve Bayesian model that predicts patient reactions. The display highlights correctly classified records in yellow.

	Predicted														
True Values	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	297	15	10	14	12	13	11	9	12	0	14	11	10	8	13
2	8	324	10	7	12	14	0	19	11	7	19	17	8	10	12
3	10	5	282	12	8	10	12	9	13	11	4	12	0	7	10
4	5	0	10	321	11	4	10	12	8	12	6	0	8	12	9
5	9	27	22	15	597	45	28	18	19	14	16	17	13	42	12
6	10	19	22	15	17	518	18	15	20	20	17	34	19	12	10
7	8	17	12	8	11	9	376	9	10	17	13	4	7	0	10
8	11	6	0	18	10	21	17	562	20	13	9	21	11	19	12
9	2	0	8	19	12	11	20	10	294	8	12	6	0	10	19
10	5	11	14	3	19	0	22	7	10	393	3	9	12	0	4
11	8	13	9	12	11	16	6	15	11	17	517	10	19	7	15
12	7	4	0	10	16	12	8	15	0	10	9	302	11	12	7
13	0	9	10	12	5	17	34	11	21	29	18	16	539	16	0
14	18	22	17	0	8	11	18	23	26	18	4	12	15	423	14
15	7	19	20	11	15	16	12	15	10	19	18	12	9	13	342

To determine the prediction accuracy of the Naïve Bayesian model, calculate:

$$Accuracy = \frac{297+324+282+321+597+518+376+562+294+393+517+302+539+423+342}{8656} * 100 = 70.32\% \qquad (3.4)$$

The prediction accuracy shown above means that the Naïve Bayesian model predicts patient reactions correctly 70.32% of the time. This percentage is below 80%, making this model not particularly useful in predicting the given response variable. This measure will be compared with the other models' prediction accuracy further below.

Pictured below is the confusion matrix for the Multinomial Regression model predicting patient reactions. The correctly classified records are highlighted in yellow.

	Predicted														
True Values	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	305	14	6	9	10	15	22	2	5	5	6	11	12	12	15
2	9	317	11	10	8	6	19	7	12	11	14	11	15	6	12
3	23	10	184	14	15	7	24	17	22	8	18	16	14	10	13
4	19	11	7	265	12	12	11	6	12	17	10	15	12	11	8
5	12	7	10	7	771	13	10	10	10	8	7	11	7	12	11
6	6	8	13	1	2	625	12	11	15	12	10	12	8	9	13
7	6	12	15	19	8	10	333	16	15	18	13	8	9	13	8
8	18	9	12	16	15	6	10	592	4	6	7	8	17	16	13
9	9	11	16	11	8	10	21	14	280	16	5	10	9	3	8
10	18	16	13	6	11	5	15	7	5	340	13	19	10	9	15
11	8	14	1	3	13	9	7	16	10	18	553	10	3	1	10
12	17	13	21	27	11	13	9	7	16	10	18	216	12	8	17
13	12	8	7	12	9	11	4	10	16	13	8	15	594	5	9
14	12	8	7	16	12	3	13	5	6	3	16	13	11	507	10
15	11	10	10	6	16	14	10	15	12	12	14	7	3	9	388

To calculate the prediction accuracy of the Multinomial Regression model, find:

$$Accuracy = \frac{305+317+184+265+771+625+333+592+280+340+553+216+594+507+388}{8656} * 100 = 72.43\% \tag{3.5}$$

This prediction accuracy means the Multinomial Regression model predicts patient reactions correctly 72.43% of the time. This percentage is below 80%, which indicates that model is not particularly useful in predicting the given response variable. However, this prediction accuracy is higher than the prediction accuracy of the Naïve Bayesian model, which shows the Multinomial Regression model more useful in predicting various patient reactions. This is very interesting, as the Multinomial Regression model was the least accurate algorithm in classifying different hazardous patient outcomes. When predicting harmful patient reactions, it is the most accurate.

The confusion matrix for the Support Vector Machine model to predict patient reactions highlights correctly classified in yellow.

	Predicted														
True Values	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	232	25	18	24	22	6	13	15	6	22	11	5	3	29	18
2	22	256	24	15		2	13	29	28	12	12	13	17	17	12
3	13	10	197	21	3	18	10	12	27	9	27	30	9	4	15
4	27	15	28	221	7	12	4	3	9	30	16	16	1	29	10
5	19	26	35	42	645	21	17	2	26	20	1	15	8	14	5
6	12	31	18	31	5	503	18	28	32	33	6	11	12	11	25
7	15	29	11	17	15	16	294	4	2	33	5	17	13	28	12
8	16	30	24	30	20	12	13	537	6	15	24	7	3	5	8
9	23	23	10	19	10	10	19	4	228	8	7	9	16	22	23
10	14	10	17	29	23	14	8	9	23	293	21	14	11	19	7
11	20	16	9	16	7	15	18	19	21	30	470	6	17	12	10
12	13	18	12	15	9	13	6	27	12	27	4	219	21	15	12
13	16	18	30	27	14	12	26	11	9	1	6	15	544	16	8
14	18	12	13	16	8	6	10	11	29	17	11	28	9	423	19
15	30	9	29	19	11	9	10	28	18	7	9	14	18	10	317

To determine the prediction accuracy of the Support Vector model, calculate:

$$Accuracy = \frac{232 + 256 + 197 + 221 + 645 + 503 + 294 + 537 + 228 + 293 + 470 + 219 + 544 + 423 + 317}{8656} * 100 = 62.14\% \tag{3.6}$$

This prediction accuracy means the Support Vector model predicts patient reactions correctly 62.14% of the time. This percentage is below 80%, which indicates this model not particularly useful in predicting the given response variable. The Support Vector machine models has the lowest prediction accuracy out of the three models discussed. Therefore, it is the least useful model in predicting patient reactions.

Shown below are the ROC curves and corresponding AUC values associated with the models predicting patient reactions in the following order: Multinomial Regression, Naïve Bayesian, and Support Vector Machine.

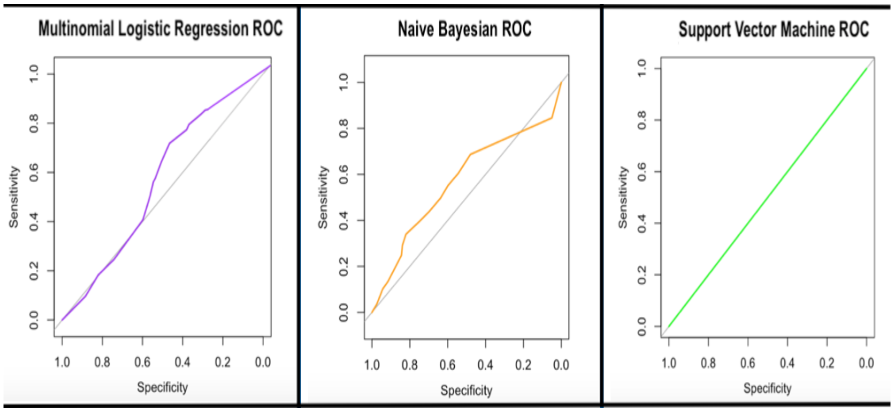


Figure 4: Left to right: Best to worst performance of multiclass classification.

The ROC curve produced by the Multinomial Regression model meets the main diagonal for some values in the graph but eventually moves above the diagonal. Therefore, this classifier sometimes randomly predicts patient reactions. The AUC value associated with this model is 0.6818, which is closer to 0.50 than 1.00. This means that more often that not, the model performs random predictions rather than accurately predicting the patient's outcome with combinations of the predictor variables.

The ROC curve produced by the Naïve Bayesian model is mostly slightly above the main diagonal of the graph. However it does dip below the diagonal at one point in the graph. When a model's ROC curve is below this diagonal, it means the model makes inaccurate prediction more often that accurate ones. This is not a desirable performance. The AUC value associated with this model is 0.6015, which is closer to 0.50 than to 1.00. This means the model performs a majority of random predictions. The ROC curve is not close to the top left corner of the graph. This model is not useful model in classifying patient reactions and would hopefully not be utilized in a real-world setting. This AUC value is lower than the AUC value associated with the Multinomial Regression model, which indicates the former model a more accurate classification model for predicting patient reactions.

The ROC curve produced by the Support Vector Machine model exactly meets the diagonal line in the center of the graph. Therefore, this classifier is randomly predicting all patient reactions. The AUC value associated with this model is exactly 0.5000, again demonstrating the randomness of the prediction model. Support Vector Machine's AUC value is the lowest out of the three model's AUC values, making it the least accurate model when predicting patient reactions. This classification model performs no differently than an individual who randomly selects a different patient reaction as the prediction.

Overall, Multinomial Logistic Regression's AUC value is the closest out of the three model's AUC values to 1.00, making it the most accurate model when predicting patient reactions. Additionally, the peak of the Multinomial Regression ROC curve is closest to the top left-hand corner of the graph out of the three ROC curves.

A table summarizes the prediction accuracy and AUC values of the various classification models that predicts patient reactions. The algorithms are listed in the order of highest performing model to lowest performing model.

Algorithm	**Prediction Accuracy**	**AUC Value**
Multinomial Regression	72.43%	0.6818
Naïve Bayesian	70.32%	0.6015
Support Vector Machine	62.14%	0.5000

Chapter 4

Discussion and Conclusion

4.1 Overview of Results for Hazardous Events

After analyzing the results of this statistical analysis, multiple conclusions can be drawn. The best model for predicting hazardous patient outcomes is clearly the Support Vector Machine model. Its prediction accuracy, ROC curve, and AUC value outperform the other two models. The SVM algorithm may produce a more useful model, as this particular algorithm works well with high-dimensional and complex data such as the FAERS dataset used in the project (Pang-Ning Tan *et al.*, 2017).

When analyzing specific risks of prescribing certain drugs to patients, medical professionals could run this classification model. This way, the medical staff could analyze which specific dangerous outcome would occur when giving a patient a certain prescription drug. Many medical staff could also avoid malpractice lawsuits from prescribing patients with a potentially very dangerous drug treatment. However, if the model predicted that the patient would suffer a minor hospitalization as a result of taking the drug, the medical team could weigh the benefits and disadvantages of inducing a certain patient with that prescription, instead of making a blind decision.

4.2 Overview of Results for Reactions

When comparing the classification models that predict patient reactions, it is clear that the most useful model for classifying harmful reactions is the Multinomial Regression model. Its prediction accuracy, ROC curve, and AUC value suggest that it performs better that the other two models tested. The Multinomial Regression algorithm is considered an attractive analysis because it does not assume normality or linearity of the dataset, meaning that these assumptions do not have to be true to create a high-performing, accurate model.

A classification model that accurately predicts the various reactions associated with a certain drug given particular input variables could be very useful to the medical professional as well as consumers. If a patient feels unsure about experiencing a side-effect to a specific drug, he or she can input certain information into the model. As a result, the patient could expect a likely reaction to the prescription drug, based on his or her inputs. Likewise, a medical professional could enter a patient's drug, dosage, and demographic information and become knowledgable about certain reactions an individual may have to a medication. The medical staff then could warn the individual in advance about likely reactions, such as a headache or fever, that the patient may experience. Comprehensively, the classification models could prove very useful in a real-word setting.

4.3 Conclusion

Overall, a variety of statistical classification models can be applied to reported medical data and utilized to predict hazardous events and harmful reactions with respect to patient usage of prescription drugs. This project aims to identify the most accurate predictive model. The model uses drug name, drug active ingredient, route of drug administration, dosage amount, dosage frequency, drug manufacturer, occurrence country, and patient reaction to classify hazardous events. To predict harmful reactions, the model utilizes the same predictor variables listed above with the exception of patient reaction, as that is the designated response variable.

Specifically, the Support Vector Machine model is the most useful in classifying possible serious patient outcomes while the Multinomial Regression model is the most accurate in predicting a variety of unfavorable patient reactions. The usage of these models could ease consumer anxiety about dangerous and abnormal side-effects associated with medicinal drugs. Medical staff could also become more knowledgable about these particular side-effect and events and pass on this information more confidently to patients. The use of these statistical models in the medical field could prove very helpful to both medical professionals and public consumers.

Bibliography

AERS. (2018), *FDA Adverse Event Reporting System (faers) Quarterly Data Extract Files* [online], https://fis.fda.gov/extensions/FPD-QDE-FAERS/FPD-QDE-FAERS.html.

Cornell, John E.,*et al.* (2009), *Multimorbidity Clusters: Clustering Binary Data From Multimorbidity Clusters: Clustering Binary Data From a Large Administrative Medical Database* Applied Multivariate Research, amr.uwindsor.ca/index.php/AMR/article/view/658.

Dumouchel, William. (1999), *Bayesian Data Mining in Large Frequency Tables, with an Application to the FDA Spontaneous Reporting System*, The American Statistician vol. 53, https://amstat.tandfonline.com/doi/abs/10.1080/00031305.1999.10474456.

Harpaz, R. *et.al.* (2012) *Novel Data-Mining Methodologies for Adverse Drug Event Discovery and Analysis.* Clinical Pharmacology and Therapeutics, https://doi.org/10.1038/clpt.2012.50.

Liu, Mei. *et al.* (2012), *Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs.* Journal of the American Medical Informatics Association, Volume 19, Issue e1, Pages e28–e35, https://doi.org/10.1136/amiajnl-2011-000699.

Medline Plus. (2019). *Diarrhea* U.S. National Library of Medicine, https://medlineplus.gov/diarrhea.html.

Poluzzi, E. *et al.* (2009), *Drug-induced torsades de pointes: data mining of the public version of the FDA Adverse Event Reporting System (AERS).* Pharmacoepidem. Drug Safe., https://doi.org/10.1002/pds.1746.

Raschka, Sebastian. (2018). *Confusion Matrix - Gradient Descent and Stochastic Gradient Descent.* rasbt.github.io/mlxtend/user_guide/evaluate/confusion_matrix.

Roland, James. (2017), *What is Torsades de Pointes?* Healthline. https://www.healthline.com/health/torsades-de-pointes

Tan, Pang-Ning. *et al.* (2017), *Introduction to Data Mining*, Pearson.