MarijuanaAnalyzer: an Intelligent Visualization and Analysis Web App

Muzi Bi, Yang Lu, Fangxia Tian, Jing Ren, Jian Wang, Yutong Wu
College of Computing, Georgia Institute of Technology, Atlanta, GA

1 INTRODUCTION

1.1 Overview

Since California first legalizing medical marijuana in 1996, most states in U.S. have already legalized medical marijuana. Moreover, 15 states and District of Columbia have fully legalized recreational marijuana. In this study, we made an interactive web app to provide a comprehensive overview of marijuana. Also, we performed statistical analysis and applied various machine learning algorithms to evaluate how marijuana influences an individual's mental health and the use of other substances. The ultimate goal is to provide insights into policymakers' decision making. Also, we hope to offer useful information of marijuana and its effects to the general public.

1.2 Problem Definition

The legalization of marijuana has been studied in various aspects,[1-6] Marijuana can be beneficial to health and stimulate the brain with a euphoric feeling. Nonetheless, it may lead to the use of other illicit drugs and affect mental health.[7-8] An in-depth analysis regarding drug use and mental health risks of marijuana is required to fully elucidate the impact of marijuana legalization. In addition, a detailed study and informative web app can help all stakeholders better understand marijuana and its impact.

1.3 Survey

Despite various studies on marijuana in recent years, there are still some limitations as most of them were too specific and fail to combine the data from other dimensions.[1,7,8,9,10] Consequently, people can not compare the weighted importance of marijuana use with other features. Machine learning, however, has been widely used in various fields to study trends and make predictions.[11,12] The advantage of using machine learning is that people can analyze big data containing multiple features and plenty of algorithms can be used to train models based on the different datasets and aims. [13] Feature importance can be used to compare the significance of target features (e.g. marijuana use) with other input features.[14] Researchers have successfully applied different machine learning models such as artificial neural networks, random forest, linear regression to study drug abuse, drug efficacy, drug toxicity, etc. [15-18] Since we have a large dataset with plenty of features, we would like to employ machine learning models to study the correlation of marijuana use with other features and use the optimized model to make predictions.

In this project, we initially provide a comprehensive overview through visualizing the current situation of marijuana and its impact in terms of other substance use. Then, We utilize different machine learning algorithms to study how marijuana influences an individual's mental health and use of other substances, providing insights for our user in an interactive and visualized way.

2 PROPOSED METHODS

2.1 Intuition

Innovations: To overcome previous limitations, (1) We combined multiple public datasets with our main dataset from the National Survey on Drug Use and Health (NSDUH)[19]. Our main dataset has 67,791 data points. Each point contains 2691 variables, which allow us to explore multiple features when predicting outcomes and give us a more accurate overview. (2) We provided a comprehensive overview of the current situation on marijuana through interactive and dynamic data visualization. Most people do not have a deep understanding of marijuana and its effects, so instead of merely presenting our machine learning models, we provided a general visualized overview to educate our web app users about some basic knowledge about marijuana / substance use / mental health. (3) We used multiple features to train our machine learning models and developed a useful tool to predict substance use and/or individual's mental health for

marijuana users. Not many studies have focused specifically on marijuana's lead to other heavy drug use and its effects on people's mental status. (4) We tested multiple different machine learning models selected the best performing model – random forest. When users input different features of an individual, the deployed models can generate results of that individual's likelihood of using other substances and possible mental state.(5) To our knowledge, there are not many studies that use machine learning in regards to marijuana, but it is a strong and useful method. We were inspired by many non-marijuana related studies[15-18], and incorporated machine learning in our study, which allows us to create an interactive prediction-oriented web app.

2.2 Data Visualization with User Interface

2.2.1 Tools

Our web app was built using HTML/CSS/JS. Bootstrap 4 framework was used to construct responsive layout of the web app. Interactive and dynamic visualization was mainly created by the D3.js library, allowing the users to interact with our graphs. For example, the user can interact with a time slider bar to view the data chronologically and filter the information based on interests. Many features, such as highlight, drop-down, and tooltip, are also provided for better interactivity.

Besides using D3.js as our primary visualization tool here, we used python/numpy/pandas/scikit-learn for data cleaning and organization.

2.3 Machine Learning

We used machine learning to understand the impact of marijuana use on other drug use. We trained six machine learning models with marijuana use status, along with demographic information including employment status, education, sex identification, general health, as well as general mental status. Models output predictions on the individual's risk on alcohol and tobacco use, as well as other heavy drugs such as heroine and methamphetamine. Similarly, two models were trained to predict the mental health status, such as difficulty dealing with daily work, and difficulty dealing with social activities.

2.3.1 Data Wrangling

We cleaned the data by eliminating duplicates, dropping invalid records, and imputing the missing data. Input features were re-categorized so that the survey data labeled as "unknown", "refused to answer", "bad record" are considered a distinct category. We then chose nine features including the marijuana use from the same dataset as used in the last section. Training data contains 55882 records with 10 features. The data was then normalized by standard scaling function to minimize co-variance and improve model robustness. We used a pair-plot to visualize the correlation between input features. For the output feature, we had two main areas: other drug use prediction and mental health prediction. All the predicted features in these two areas were either binary or multi-classification.

2.3.2 Model Selection

We tried different machine learning models in different problems, and we found the random forest had the optimal performance. Random forest is an ensemble bagging algorithm. It consists of multiple decision trees, each tree is trained with a random sampling of training data. It reduces the variance of a single tree and gives an overall better accuracy. There are two major challenges in our model training. First, the data has categorical values that require mapping/encoding. Second, the classes in y are severely unbalanced in some of the models, such as heroin and methamphetamine uses. This will lead to low accuracy on the minority class. The random forest model performs well with unbalanced data, as well as categorical data. We also used up-sampling to balance the classes and achieving optimal results.

2.3.3 Model Training

For the prediction value with well balanced output, we used the random forest classification model– Some parameters were chosen after hyperparameter tuning. For the prediction with unbalanced classes, we first up-sampled the minority class, then used the same random forest classification model parameters as before. We also tried Penalized-SVM algorithms by using class-weight = "balanced" to penalize mistakes on the minority class.

2.3.4 Model Evaluation

We evaluated our machine learning model in several aspects. First, we calculated and visualized our model accuracy, precision, f1, and recalled values using a chart which includes how we solve the imbalanced class problem, how we compare multi-class labels and finally showing our finalized model

parameters. Second, we used the confusion matrices to visualize the predicted results, which is a common method to compare the real and predicted data and evaluate the model. Third, we could directly see how important is marijuana compared to other input features by showing the feature importance of our prediction. Finally, we used the ROC curve as well as the AUC value to further evaluate our prediction results. The ROC curve gives information on both the true positive and false positive rate on the trained model. The more the ROC curve bends towards the left upper corner, the better the prediction is. The integrated area under the ROC curve (AUC) can also be used to show the likelihood of a model to distinguish observations from two classes. All of our evaluations will be elaborated in Section 3.

2.3.5 Prediction & Website User Interface

Multiple machine learning models have been deployed on AWS S3. Our web app will interact with models via AWS API Gateway and Lambda. The results gathered from various models will be visualized by an interactive d3.js tree diagram.

3 EVALUATION & EXPERIMENTS

3.1 Data Visualization

3.1.1 Overview

We aim to give users an overview of marijuana and the purpose of our study via interactive and dynamic visualization with descriptions.

3.1.2 Legalization Timeline

We gathered information about the status of marijuana legalization for each state from state laws (1996 - 2020) and created a timeline map to visualize the progress over these years (Figure 1). The map used a single-hue color scale to represent the status of the state's marijuana legalization, specifying medical and recreational use. We provided a slide bar for users to drag through the years and see the changes over the years. The process can be automatically demonstrated using the "play" button. If the cursor hovers over the state, specific information of marijuana legal status will show up on the tooltip.

3.1.3 Marijuana, Substance Use Disorders, and Mental Disorders

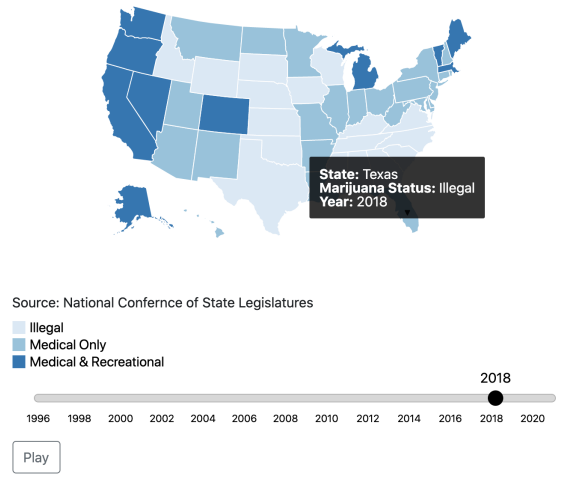

Figure 1: Timeline of marijuana legalization in the United States from 1996 to 2020

The Global Burden of Disease Study[20] dataset is utilized to visualize the prevalence of substance use disorder in a choropleth world map (figure 2a). The choropleth showed the percentage of people with substance use disorder in each country from 1990 to 2016: The corresponding drug use disorder prevalence data will show up when hovering over a specific country, and years can be selected by using the drop-down feature. At least 5 % population in the United States suffer from drug use disorder since 1990. There is also an interactive bar chart (figure 2b) comparing the percentage of people having different mental disorders between the US and the World average values in 2017. People in the United States have a significantly higher risk of "anxiety disorders" (68 %), "schizophrenia" (104 %), and "eating disorders" (57 %) compared to the world average.

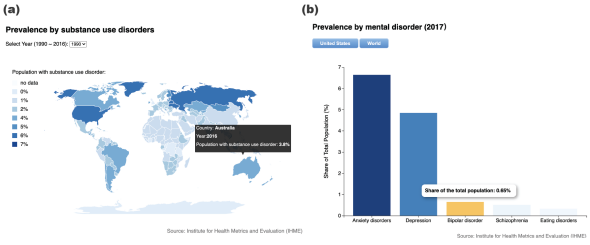

Figure 2: Share of population with (a) substance use disorders and (b) mental disorders

3.2 Correlation Matrices Between Input and Output Features

3.2.1 Overview

Besides the descriptive statistics, we also performed statistical analyses of Pearson's correlations between marijuana and substance use and mental health.

3.2.2 Drug Abuse

We draw a correlation matrix for all substances, such as marijuana, alcohol, tobacco to visualize the relationship among all variables (Figure 3a). We then created an interactive matrix in that the correlation between two specific variables is highlighted with a mouse cursor hovering over. Meanwhile, correlation coefficient (r) will show up in tooltip as well. The darker the grid is, the stronger the relationship between two attributes is. Marijuana has a strong correlation with tobacco, alcohol, cocaine, and hallucinogens ($r > 0.4$).

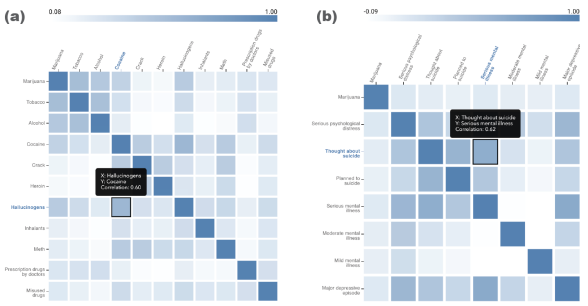

Figure 3: Interactive correlation matrices for (a) drug substances, and (b) mental health

3.2.3 Mental Health

Similar processes were performed on correlations between marijuana use and mental health (Figure 3b). We only used the adult dataset for mental health analysis because the survey for youth (12-17 years old) is insufficient. The results show relatively weak correlations between different types of mental issues with marijuana (r : $0.06 \sim 0.13$).

3.3 Machine Learning

3.3.1 Data Wrangling

We followed the method section to prepare our machine learning data. Figure 4 shows the pairlot of the selected input features. Pariplot shows the relationship between two variables. From the graph, we can observe that there is no strong correlation between our input features which is good since when choosing the input features people prefer independent features to avoid multicollinearity issues.

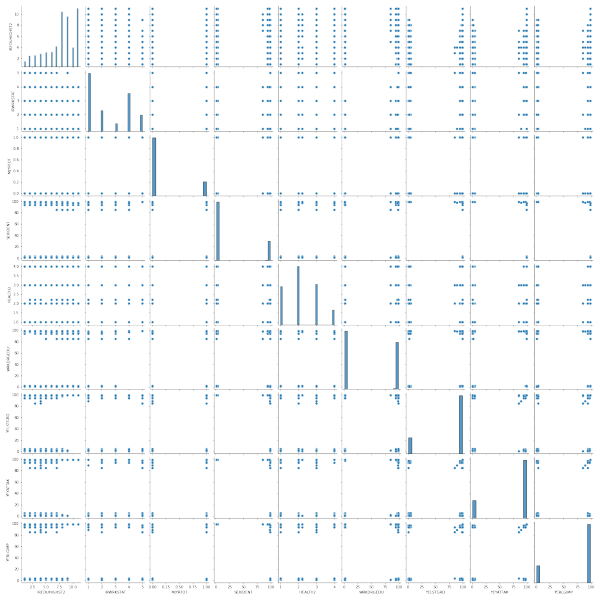

Figure 4: Pairplot of selected machine learning input features

3.3.2 Model Selection

To predict the other drug abuse, we classify the predicted results to binary classes with 0 for not using the drug and 1 for using the drug. We found that other than tobacco and alcohol, all the other predicted features have imbalanced classes with nearly 90 percent of people fall in category 0. If we directly train the model with this imbalanced dataset, we could get very high accuracy, however, the model will be meaningless for real application since it ignores the minority class. To solve this, we used the up-sampling method by increasing the number of minority classes. This process was achieved by resampling the minority class with replacement to match the majority class. The resampled data was then combined with the majority to form a new dataset for training. We then train a random forest model with the resampled data as well as pristine tobacco and alcohol data. We choose to use the random forest as the training model because it uses the bagging ensemble method and has the best performance.

We also tried Penalized-SVM to solve the class imbalance issue and compared it to the up-sampling method. Table 1 compares model performance between Support Vector Machine (SVM) and Random

Table 1: Model Comparison

model	report	precision	recall	f1-score
SVM	1	0.41	0.94	0.57
	2	0	0	0
	3	0	0	0
	4	0	0	0
	5	0.88	0.6	0.71
Random Forest	1	0.42	0.77	0.54
	2	0.3	0.08	0.13
	3	0.2	0.03	0.05
	4	0.14	0.01	0.03
	5	0.77	0.67	0.71

Table 2: Machine learning model parameters comparison

class	feature	accuracy	precision	recall	f1
0	Tobacco	0.69	0.76	0.59	0.67
0	Alcohol	0.84	0.80	0.59	0.68
0	Heroin	0.98	0.98	1.00	0.99
0	Heroin(US)	0.79	0.82	0.74	0.78
0	Heroin (PA)	0.83	0.99	0.84	0.91
0	Cracker(US)	0.78	0.81	0.73	0.77
0	Cocaine(US)	0.73	0.72	0.76	0.74
0	Methamphetamine(US)	0.75	0.77	0.70	0.73
1	Tobacco	0.69	0.63	0.79	0.70
1	Alcohol	0.84	0.85	0.94	0.90
1	Heroin	0.98	0.08	0.01	0.01
1	Heroin(US)	0.79	0.77	0.84	0.80
1	Heroin (PA)	0.83	0.06	0.56	0.11
1	Cracker(US)	0.76	0.83	0.79	0.67
1	Cocaine(US)	0.73	0.75	0.71	0.73
1	Methamphetamine(US)	0.75	0.73	0.80	0.76

Forest on predicting individuals' difficulty in dealing with daily work. We can see that SVM performs poorly on 3 classes (outputs are all 0s), indicating the result is severely biased. Random Forest is able to provide more meaningful predictions, although the results on class = 2, 3, 4 are still not optimal.

For future work, we recommend increasing model complexity, along with adding more features to seek a better bias-variance tradeoff and improve predicting accuracy.

3.3.3 Model Training

For other drug abuse prediction, after up-sampling the imbalanced data, we input the data to random forest classification model. We use 0.3 as the test size. After hyperparameters tuning, some of our model parameters are: n-estimators =50, criterion = 'gini', max-Depth =100, max-features= 0.5. We run a 5-fold cross-validation on the model. To compare with this upsampling method, we also run the Penalized-SVM model directly on the imbalanced data. Results are listed in table 2. US in the parenthesis stands for after up-sampling and PA in the parenthesis stands for using Penalize-SVM model (if not labeled PA, the models are random forest). We can see that tobacco and alcohol shows reasonable results. However, for heroin and other heavy drugs, random forest without up-sampling has accuracy up to 0.98 but with nearly zero precision, recall and f1 score for class 1. After up-sampling, we have seen a significant improvement in accuracy, precision, recall and f1 score for both classes. Compared to up-sampling method, the Penalized-SVM model increases the recall score, however, precision and f1 score are not improved, indicating the model does not handle imbalanced classes well.

3.3.4 Model Evaluation

Confusion Matrix Figure 5 compares the confusion matrices of Heroine Use predicted by random forest without (5a) and with up-sampling (5c) and by Penalized-SVM model (5b). We can clearly see the improvement from 5a to 5c by balancing the data.

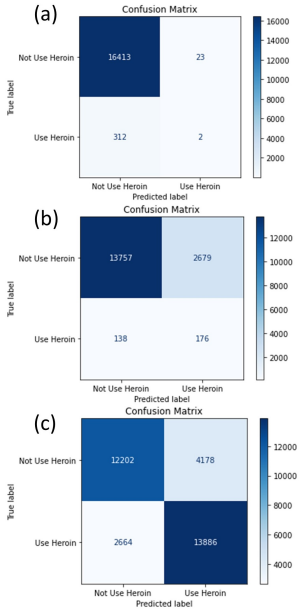

Figure 5: Confusion matrices of (a) Heroin, (b) Heroin(PA), (c) Heroine (US) prediction

Feature Importance Figure 6 shows the feature importance of tobacco-use prediction (a) and heroin-use prediction (b). The marijuana feature is highlighted in red. We can see that marijuana has an obvious impact on the predicting results compared to other input features, which shows that our prediction is legitimate based on marijuana use.

ROC Analysis Figure 7a & b show the ROC analysis of heroin-use prediction using random forest classification model without and with up-sampling. The ROC curve can give an overall accuracy of the machine learning model. We can clearly see that the ROC curve bends more towards to the upper

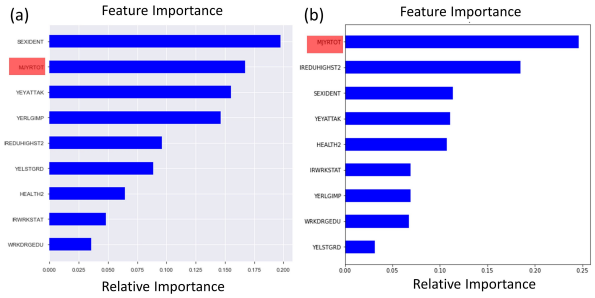

Figure 6: Feature importance of (a) tobacco use prediction (b) heroin use prediction

left corner which indicates a higher true positive rate and a lower false positive rate. The AUC value indicates how well a model distinguishing positive and negative classes. It ranges from 0 to 1 with higher value indicating better performance. Compared with the imbalanced dataset in Figure 7a, the AUC value of up-sampled dataset increases to 0.88 from 0.72, which is a huge improvement.

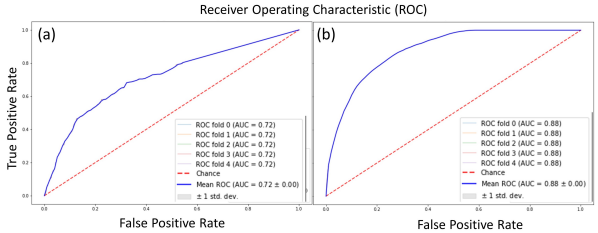

Figure 7: ROC analysis of (a) heroin and (b) heroine (US) predictions

Machine Learning Predictor 10 trained machine learning models were deployed on AWS. Users can input parameters from drop-down lists as shown in Figure 8. When "Predict" button is clicked, the web app will interact with models endpoints through AWS API Gateway and AWS Lambda. The predicted values are displayed in a collapsible tree. Figure 8 shows an representative results using the specified parameters.

4 Conclusions and Discussion

In this project, we developed a web app, in which we used data visualization techniques and illustrates multiple aspects of marijuana, including overview of marijuana legalization, correlations between marijuana, and drug disorders and mental disorders. In addition, the web app utilized machine learning models to help users predict the influence of marijuana on drug use and mental health. To visualize

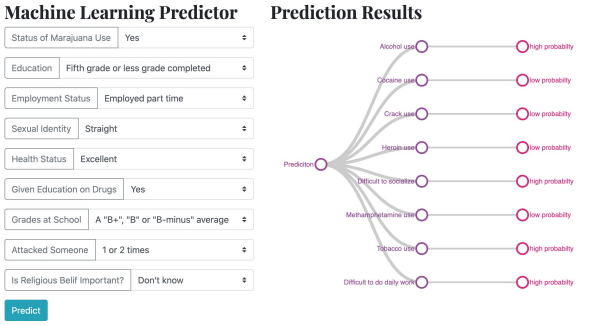

Figure 8: Machine learning predictor tool: Input feature selection and model prediction results

the overview of marijuana legalization and its effects, we combined several big datesets and used D3.js to visualize the legalization of marijuana with both geographic information and time information. The users can view the information of interest in an interactive manner. Also, to predict the impact of marijuana on other drug uses and mental health, we use marijuana use with other demographic variables as input features to train multiple machine learning models. We have used different sampling methods, different classification models, and determined the best performed models for each feature. After training and evaluating the machine learning models, we built an API and deployed them on AWS. We also designed an interactive user-interface to let them choose their input features and view the results at the web app frontend.

5 Distribution of team member effort

All team members have contributed similar amount of effort. Specifically, everyone contributed in the data cleaning process. Yang, Jian, Fangxia did the overview visualization; Yutong and Jing worked on the machine learning model training; Muzi, Yang, and Jian did front-end logic implementation; Fangxia implemented the back-end logic. Muzi and Yang worked on front-end visualisation; Everyone contributed in the report writing and poster design. Jian was in charge of delegating tasks and managing overall activities.

References

[1] K. Ally Memedovich et al., "The Adverse Health Effects and Harms Related to Marijuana Use: An Overview Review," CMAJ Open 6, no. 3 (2018): E339–46, https://doi.org/10.9778/cmajo.20180023.

[2] Bin Yu et al., "Marijuana Legalization and Historical Trends in Marijuana Use among US Residents Aged 12-25: Results from the 1979-2016 National Survey on Drug Use and Health," BMC Public Health 20, no. 1 (2020): 1–10, https://doi.org/10.1186/s12889-020-8253-4.

[3] Robert G. Morris et al., "The Effect of Medical Marijuana Laws on Crime: Evidence from State Panel Data, 1990-2006," PLoS ONE 9, no. 3 (2014), https://doi.org/10.1371/journal.pone.0092816.

[4] Eric L. Sevigny, Rosalie Liccardo Pacula, and Paul Heaton, "The Effects of Medical Marijuana Laws on Potency," International Journal of Drug Policy 25, no. 2 (2014): 308–19, https://doi.org/10.1016/j.drugpo.2014.01.003.

[5] Davide Dragone et al., "Crime and the Legalization of Recreational Marijuana," Journal of Economic Behavior and Organization 159 (2019): 488–501, https://doi.org/10.1016/j.jebo.2018.02.005.

[6] Jeffrey Brinkman and David Mok-Lamme, "Not in My Backyard? Not so Fast. The Effect of Marijuana Legalization on Neighborhood Crime," Regional Science and Urban Economics 78, no. July (2019): 103460, https://doi.org/10.1016/j.regsciurbeco.2019.103460.

[7] Mohammad Hajizadeh, "Legalizing and Regulating Marijuana in Canada: Review of Potential Economic, Social, and Health Impacts," International Journal of Health Policy and Management 5, no. 8 (2016): 453–56, https://doi.org/10.15171/ijhpm.2016.63.

[8] Andrew A. Monte, Richard D. Zane, and Kennon J. Heard, "The Implications of Marijuana Legalization in Colorado," JAMA - Journal of the American Medical Association 313, no. 3 (2015): 241–42, https://doi.org/10.1001/jama.2014.17057.

[9] Janessa M. Graves et al., "Employment and Marijuana Use Among Washington State Adolescents Before and After Legalization of Retail Marijuana," Journal of Adolescent Health 65, no. 1 (2019): 39–45, https://doi.org/10.1016/j.jadohealth.2018.12.027.

[10] Jeffrey Brinkman and David Mok-Lamme, "Not in My Backyard? Not so Fast. The Effect of Marijuana Legalization on Neighborhood Crime," Regional Science and Urban Economics 78, no. July (2019): 103460, https://doi.org/10.1016/j.regsciurbeco.2019.103460.

[11] Ngiam, Kee Yuan, and Ing Wei Khor. "Big Data and Machine Learning Algorithms for Health-Care Delivery." The Lancet Oncology 20, no. 5 (2019): e262–73. https://doi.org/10.1016/S1470-2045(19)30149-4.

[12] Luechtefeld, Thomas, Dan Marsh, Craig Rowlands, and Thomas Hartung. "Machine Learning of Toxicological Big Data Enables Read-across Structure Activity Relationships (RASAR) Outperforming Animal Test Reproducibility." Toxicological Sciences 165, no. 1 (2018): 198–212. https://doi.org/10.1093/toxsci/kfy152.

[13] Ganapathi, Archana, Harumi Kuno, Umeshwar Dayal, Janet L. Wiener, Armando Fox, Michael Jordan, and David Patterson. "Predicting Multiple Metrics for Queries: Better Decisions Enabled by Machine Learning." Proceedings - International Conference on Data Engineering, 2009, 592–603. https://doi.org/10.1109/ICDE.2009.130.

[14] Weinstein, J N, K W Kohn, M R Grever, V N Viswanadhan, L V Rubinstein, A P Monks, D A Scudiero, et al. "Neural Computing in Cancer Drug Development: Predicting Mechanism of Action." Science 258, no. 5081 (October 16, 1992): 447 LP – 451. https://doi.org/10.1126/science.1411538.

[15] Sahker, Ethan, Laura Acion, and Stephan Arndt. "National Analysis of Differences among Substance Abuse Treatment Outcomes: College Student and Nonstudent Emerging Adults." Journal of American College Health 63, no. 2 (2015): 118–24. https://doi.org/10.1080/07448481.2014.990970.

[16] Hammann, F., H. Gutmann, N. Vogt, C. Helma, and J. Drewe. "Prediction of Adverse Drug Reactions Using Decision Tree Modeling." Clinical Pharmacology and Therapeutics 88, no. 1 (2010): 52–59. https://doi.org/10.1038/clpt.2009.248.

[17] Chen, Yixin, Shilpa S. Thosar, Reba A. Forbess, Mark S. Kemper, Ronald L. Rubinovitz, and Atul J. Shukla. "Prediction of Drug Content and Hardness of Intact Tablets Using Artificial Neural Network and Near-Infrared Spectroscopy." Drug Development and Industrial Pharmacy 27, no. 7 (2001): 623–31. https://doi.org/10.1081/DDC-100107318.

[18] Byvatov, Evgeny, Uli Fechner, Jens Sadowski, and Gisbert Schneider. "Comparison of Support Vector Machine and Artificial Neural Network Systems for Drug/Nondrug Classification." Journal of Chemical Information and Computer Sciences 43, no. 6 (2003): 1882–89. https://doi.org/10.1021/ci0341161.

[19] "National Survey on Drug Use and Health (NSDUH)." National Survey on Drug Use and Health (NSDUH) | SAMHDA. Accessed October 31, 2020. https://www.datafiles.samhsa.gov/study-series/national-survey-drug-use-and-health-nsduh-nid13517.

[20] Institute for Health Metrics and Evaluation.Global Health Data Exchange. http://ghdx.healthdata.org/gbd-results-tool. Accessed July 30, 2020.