Predicting Toxics Release Inventory Carcinogenicity Using A Machine Learning Approach

A Project Presented to

The Faculty of the Department of Industrial and Systems Engineering

San José State University

In Partial Fulfillment of the Requirements

For the Degree Master of Science in Industrial and Systems Engineering

By

Maram Salameh

Spring 2021

SAN JOSÉ STATE UNIVERSITY

The Undersigned Project Committee Approves the Project Titled

Predicting Toxics Release Inventory Carcinogenicity Using a Machine Learning Approach

By

Maram Salameh

APPROVED FOR THE DEPARTMENT OF INDUSTRIAL & SYSTEMS ENGINEERING

Advising Professor: Dr. Hongrui Liu

Department of Industrial and Systems Engineering

Table of contents

Abstract …… 6

1. Introduction ……... 7
2. Literature Review ……. 8
3. Terminology ……. 10
4. Methodology ……..... 11

4.1 Python Libraries ……. 11

4.2 Models ……. 12

5. Analysis and Results ……... 14

5.1 Dataset ……... 14

5.2 Data Preprocessing …….. 14

5.3 Data Analysis …….. 14

5.4 Evaluation Metrics ……... 15

5.5 Results ……. 16

6. Conclusions and Recommendations …… 17

6.1 Conclusions ……. 17

6.2 Recommendations ……... 17

7. References ……... 18

List of Figures

Figure 1: Linear Discriminant Analysis Model …… 12

Figure 2: Random Forest Classifier Model ……... 13

Figure 3: K Nearest Neighbor Model …… 14

List of tables

Table 1: Average Metric Results when applying under-sampling …… 16

Table 2: Average Metric Results when applying over-sampling …….. 16

ABSTRACT

This project applies classification techniques to predict carcinogenicity using public data from the Environmental Protection Agency's (EPA) Toxics Release Inventory (TRI) program. TRI is a database containing annual information on toxic releases and waste management activities gathered from facilities throughout the United States. The TRI database is a self-reported collection of data and some facilities that are required to report their releases do not file any reports. EPA seldom verifies the accuracy of the reported data and due to its light regulation, the inventory database contains estimates, mistakes, and data gaps.

Machine learning methods can help fill those gaps by using the information to identify patterns, potential concerns, and gain a better understanding of possible risks. Three types of machine learning models were developed: Random Forest, K Nearest Neighbor, and Linear Discriminant Analysis. Two sampling methods were used to analyze the data: Synthetic Minority Over-sampling Technique (SMOTE) and Condensed Nearest Neighbor an Under-sampling technique (CNN). The Random Forest Classifier Model outperforms the other models with an overall average accuracy of 99%.

1. INTRODUCTION

Congress passed the 1986 Emergency Planning and Community Right-to-Know Act (EPCRA) in response to the December 1984 industrial disaster, when close to 40 metric tons of methyl isocyanate (CH3NCO) gas was accidentally released at a Union Carbide plant in Bhopal, India [10]. Under the EPCRA, the Environmental Protection Agency (EPA) formed the Toxics Release Inventory (TRI). TRI is a database that provides information to the public about the release and presence of toxic chemicals in their communities. The TRI database includes information that covers the company, its location, industry classification, the chemical characteristics, and the environmental releases of each chemical.[11]

The TRI is a self-reported database, and some facilities that are required to file a report fail to do so. This issue is problematic because the EPA rarely verifies the data's accuracy, facilities often report estimation rather than actually measured releases. Due to light regulations on the industry, the inventory contains mistakes, estimates, and data gaps.

Although the TRI database may not be thorough and complete, it gives the public access to information about different releases in their communities such as carcinogen chemicals. Machine learning can help analyze the data to identify patterns and trends so that potential risks are better understood.

2. LITERATURE REVIEW

This section explores previous work conducted by various authors and researcher. The aim of this chapter is to summarize and critically evaluate the current knowledge in the field of carcinogen classification.

A study by Moorthy et al. [2] investigates machine learning classification models for the carcinogenicity and mutagenicity. The authors analyzed data extracted from the Carcinogenic Potency Database (CPDB). The dataset consists of carcinogenic and mutagenic information of 1481 chemically diverse molecules in various species. The authors applied Random Forest algorithm using physicochemical descriptors and structural fingerprints. Their analysis found that Random Forest models correctly classify more than 70%.

Another study by Zhang et al. [3] aimed to develop three ensemble classification models, (Ensemble XGBoost Ensemble, and SVM Ensemble Random Forest), to predict carcinogenicity of chemicals. The authors analyzed seven types of molecular fingerprints based on a dataset containing 1003 various compounds with rat carcinogenicity. They concluded that ensemble models surpassed the basic classifiers in both overall accuracy and Area Under the Curve (AUC). The top ensemble model (Ensemble XGBoost) achieved an average accuracy of 70.1%.

Zhang et al. [4] created a novel prediction model of carcinogenicity of chemicals by constructing a naïve Bayes classifier. The authors utilized data from several databases such as such as the Carcinogenic Potency Database, and the US National Toxicology Program (NTP) database. The proposed prediction model was validated through an internal 5-fold cross validation and an external test set. The naïve Bayes classifier returned an average prediction accuracy score of 90%.

A study conducted by Tan et al. [5] compares support vector machine (SVM) and artificial neural network (ANN) applications in predicting chemical carcinogenicity. The accuracies of the models were evaluated using a set of 844 compounds, including 600 carcinogenic and 244 noncarcinogenic molecules. Their study concluded that on average the SVM model returned an overall prediction accuracy of 88.1% while the ANN model gave an average overall accuracy of 84.2%.

Luan et al. [6] conducted a study to establish an accurate classification model for the prediction of the carcinogenic property of N-Nitroso compounds (NOC). The authors implemented linear discriminant analysis (LDA) and support vector machine to 148 N-nitroso compounds data extracted from which 116 are carcinogenic compounds, and 32 are non-carcinogenic compounds. The molecular structure of the compound was used to compute seven descriptors that were used as input to the classification models. The obtained results show that SVM model outperformed LDA with an overall accuracy of 95.2%.

3. TERMINOLOGY AND CONCEPTS

The following section lists the extended definition of the concepts mentioned throughout the report.

Machine Learning: A subset of artificial intelligence where computers or systems are able to learn from past experience without being programmed to do so. Machine learning applies a variety of techniques and statistical methods to handle and ingest large amounts of data. Machine learning aims to draw and improve upon knowledge using past historical data. Machine learning aids in identifying patterns to make recommendations for decisions, as well as predictions about the future. [7]

Supervised learning: A machine learning method where past labeled data is used to train the algorithm to make future predictions. The algorithm is able to make prediction on new data based on the trained model. The training set requirements for supervised learning include inputs (i.e., features) and outputs (i.e., targets). The targets in the training set are labeled with specific values [7]. Popular supervised learning algorithms include regression, and decision trees.

Classification: The prediction task when the outcome is comprised of two or more different classes. The class of each testing event is determined by aggregating the features and uncovering shared patterns among each class using the training data. This process is achieved in two parts, first the classification algorithm is applied on the training data and then the best fitting model is validated using a labeled test data to measure the model's accuracy. [8]

4. METHODOLOGY

This section describes the tools used to conduct the analysis. How the tools are applied on this particular dataset.

4.1 Python Libraries

Pandas: A python language open-source library, intended to make working with structured and time series data uncomplicated and effortless. This library package is designed for data manipulation and analysis. Pandas enables for importing data from numerous file formats such as SQL, JSON, Microsoft Excel, and comma-separated values. Pandas makes it easy to perform various data manipulation tasks such as merging and joining, as well as data cleaning and wrangling.

NumPy: Short for numerical python, a numerical computing package in Python. NumPy provides a high-level multidimensional array object and other derived objects such as metrices. NumPy also provides procedures for swift operations on arrays including mathematical, logical, linear algebra, and Fourier transform. NumPy's main functionality is the "ndarray", it encapsulates multidimensional arrays of the same data types and size. NumPy arrays makes possible for advanced mathematical operations to execute on large number of data efficiently and with less code than Python's built-in sequence.

Scikit-learn: An open-source Python library that provides support for supervised and unsupervised machine learning. It also provides a variety of packages for building regression models, tree-based models, and clustering models. Scikit-learn allows for model selection by validating, comparing, as well as tuning parameters and models. Model section helps improve accuracy through the use of parameter tuning algorithms such as cross validation.

Imbalanced-learn: An open-source python library that relies on Scikit-learn. It provides a number of re-sampling methods usually used when dealing with classification with imbalanced datasets. These methods include under-sampling, over-sampling, combination of over- and under-sampling, and ensemble learning methods.

4.2 Models

Linear Discriminant Analysis: A method that finds a linear combination of features that characterizes or separates two or more classes. By assuming equal covariance between K classes, a linear function in x can be created, denoting that the decision boundary between any pair of classes is also a linear function in x. This mechanism is illustrated in figure 1. The resulting combination is used as a linear classifier, or for dimensionality reduction. [8]

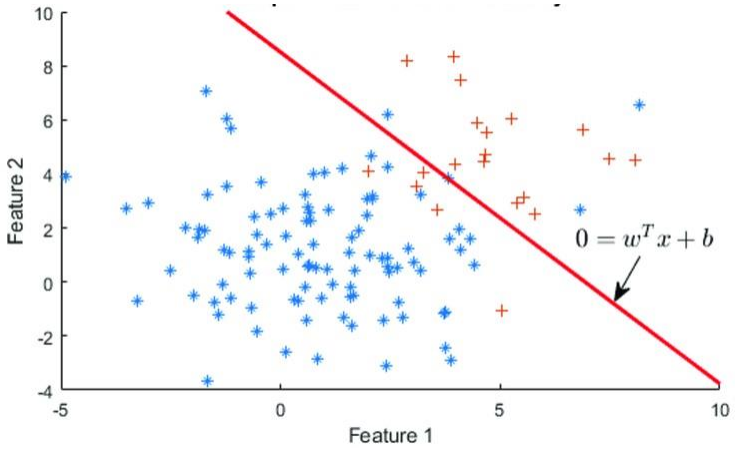

Figure 1: Working Algorithm of Linear Discriminant Analysis [9]

Random Forest Classifier: A supervised learning technique that can be applied to classification or regression tasks. Random forest is derived from ensemble learning and utilizes bagging or bootstrap aggregation; a process that incorporates multiple classifiers to make prediction and improve the performance of the model. Random forests work by aggregating a number of decision trees on different subsets of a dataset and then averaging the output of each tree to improve the predictive accuracy of that dataset [1]. The basic architecture for this

algorithm is depicted in figure 2. A large number of trees in the forest reduces the variance of the model and prevents the overfitting issue.

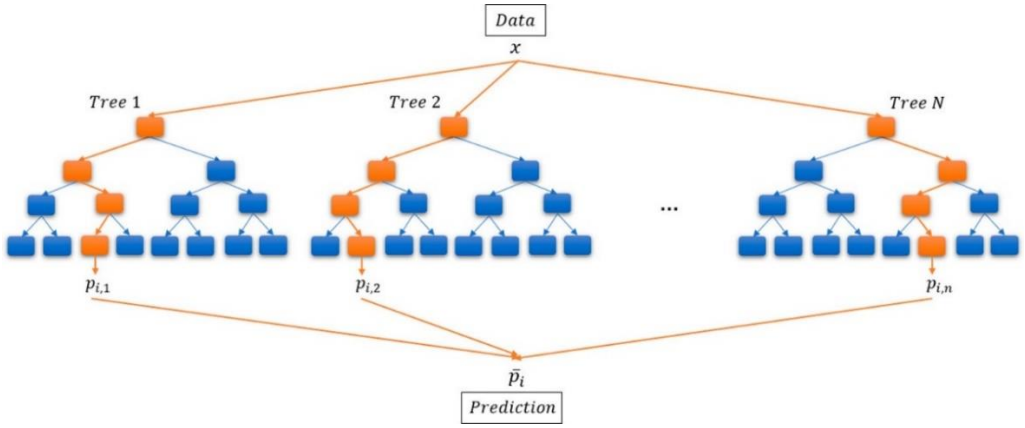

Figure 2: Working algorithm of Random Forest Model [7]

K Neighbors Classifier (KNN): A non-parametric classification method that works by assigning unlabeled observations the class of the nearest of previously labeled observations. This method assigns classes independently of the joint distribution of the observations and their classification. There is no primary way to choose ‘k’ except through cross validation. Performance is affected by noise and depends on the size of the data [8]. An illustration of KNN method is shown in figure 3.

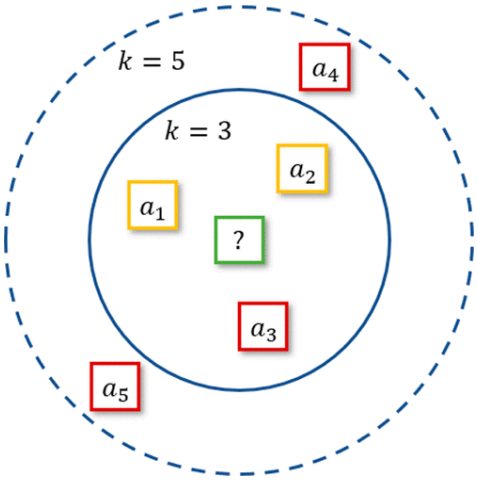

Figure 3: Working algorithm of K Neighbor Model [7]

5. ANALYSIS AND RESULTS

This section explains in detail the analysis approach and calculations performed.

5.1 Dataset

The compiled dataset from the reporting years 2018 and 2019 contains information reported by facilities that manufacture or process more than a threshold quantity of a listed toxic chemical in the state of California. The data includes releases of each chemical, facility characteristics such as location and industry sector, as well as the medium of release (i.e., air, water, or land). The dataset also includes chemical characteristics such as mass, carcinogen, and metal category.

5.2 Data Preprocessing

The compiled data contains geospatial data such as address, city, ZIP code, as well as longitude and latitude. Agglomerative Clustering from the Scikit-learn package is implemented to draw relevant information from geospatial data. The longitude and latitude features represent a three-dimensional space. In their raw format they create a lot of noise when modeling the data. In order to input useful information in the model it makes sense to group the geospatial observations into clusters. Applying feature engineering aids in extracting useful information from the existing data so that model performance can be enhanced.

5.3 Data Analysis

After data preprocessing and cleaning the final dataset is composed of 5808 observations distributed among 38 different types of chemical compounds that are either labeled as Carcinogen or noncarcinogen. The data is highly imbalanced and there is an unequal distribution of observations in the dataset. To address this issue techniques from the Scikit-learn and Imbalanced-learn were applied to re-sample the dataset and create a more balanced class

distribution in the training and testing datasets. Synthetic Minority Over-sampling Technique (SMOTE), and Condensed Nearest Neighbor an Under-sampling technique (CNN) were implemented on the dataset so that there is a balanced representation between the majority and minority class.

5.4 Evaluation Metrics

Having a highly imbalanced dataset requires thorough model evaluation, this is important in order to determine how well the models perform for each individual class. Using "StratifiedKFold" and "cross_val_score" methods from Scikit-learn each classification model is evaluated using cross validation.

Classification metrics are based on the Confusion Matrix, a cross table that reports the number of events between two observations, the true observation and the predicted observation.

Accuracy: is the percentage of correctly predicted observations with respect to the total number of observations.

$$Accuracy = \frac{True\ Positive + True\ Negative}{Total\ Observation}$$

Recall: is the percentage of correctly predicted observations with respect to the total number of observations in that class.

$$Recall\ = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Precision: is the percentage of correctly predicted observations with respect to the total predicted positive observations.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

F1: is the weighted mean of recall and precision, the closer the f1 score to 1 the better the model.

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

5.5 Results

The tables below summarize the mean metric results of each model. Table 1 shows the classification metrics scores when implementing an under-sampling technique, while Table 2 shows the classification metrics scores when applying an over-sampling technique.

Model	Accuracy	F1	Recall	Precision
Random Forest	99.25 %	99.25 %	99.25%	99.28 %
K Nearest Neighbor	42.21 %	42.70 %	42.21 %	44.40 %
Linear Discriminant Analysis	79.43 %	76.47 %	79.43 %	84.36 %

Table 1: Average Metric Results when applying under-sampling.

Model	Accuracy	F1	Recall	Precision
Random Forest	99.60 %	99.60 %	99.60 %	99.61 %
K Nearest Neighbor	83.44 %	83.37 %	83.44 %	83.94 %
Linear Discriminant Analysis	70.85 %	68.04 %	70.85 %	81.63 %

Table 2: Average Metric Results when applying over-sampling.

6. CONCLUSIONS AND RECOMMENDATIONS

6.1 Conclusions

Machine learning models can be implemented to tackle data gaps and discrepancies in the TRI database. Machine learning models implemented aimed to classify the chemical releases as carcinogen or noncarcinogen. Carcinogenicity analysis is made possible using Python libraries and classification techniques. Classification metrics were compared using cross validation to find the best fitting model with the highest classification accuracy.

6.2 Recommendations

The obtained results show that Random Forest model outperforms the K nearest neighbor and Linear Discriminant Analysis models. Random Forest model is able to accurately classify chemical with a 99% accuracy score, this is due to the mechanism of the algorithm. It is recommended to implement Random forest model in classifying chemicals in the TRI database.

REFRENCES

[1] Hastie T., Tibshirani R., Friedman J. (2009) “Random Forests. In: The Elements of Statistical Learning”. Springer Series in Statistics. Springer, New York, NY. Pp. 587-604, doi:10.1007/978-0-387-84858-7_15

[2] Moorthy, N. H., Kumar, S., & Poongavanam, V. (2017). Classification of carcinogenic and mutagenic properties using machine learning method. Computational Toxicology, 3, 33-43. doi:10.1016/j.comtox.2017.07.002

[3] Zhang, L., Ai, H., Chen, W. et al. CarcinoPred-EL: Novel models for predicting the carcinogenicity of chemicals using molecular fingerprints and ensemble learning methods. Sci Rep 7, 2118 (2017). doi:10.1038/s41598-017-02365-0

[4] Zhang, H., Cao, Z., Li, M., Li, Y., & Peng, C. (2016). Novel naïve Bayes classification models for predicting the carcinogenicity of chemicals. Food and Chemical Toxicology, 97, 141-149. doi:10.1016/j.fct.2016.09.005

[5] Tan NX, Rao HB, Li ZR, Li XY. Prediction of chemical carcinogenicity by machine learning approaches. SAR QSAR Environ Res. 2009;20(1-2):27-75. doi: 10.1080/10629360902724085. PMID: 19343583.

[6] Luan F, Zhang R, Zhao C, Yao X, Liu M, Hu Z, Fan B. Classification of the carcinogenicity of N-nitroso compounds based on support vector machines and linear discriminant analysis. Chem Res Toxicol. 2005 Feb;18(2):198-203. doi: 10.1021/tx049782q. PMID: 15720123.

[7] Jiao, Z., Hu, P., Xu, H., & Wang, Q. (2020). Machine Learning and Deep Learning in Chemical Health and Safety: A Systematic Review of Techniques and Applications. ACS Chemical Health & Safety, 27(6), 316-334. doi:10.1021/acs.chas.0c00075

[8] A. Singh, N. Thakur and A. Sharma, "A review of supervised machine learning algorithms," 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), 2016, pp. 1310-1315.

[9] Staat, M., Erni, D., YRA - Young Researchers Academy MedTech in NRW (Eds.), 2019. 3rd YRA MedTech Symposium 2019: May 24 / 2019 / FH Aachen. YRA MedTech Symposium. doi:/10.17185/duepublico/48750

[10] Koehler, Dinah A., and John D. Spengler. "The Toxic Release Inventory: Fact or Fiction? A Case Study of the Primary Aluminum Industry." Journal of Environmental Management 85 (2007): 296.

[11] U.S. Environmental Protection Agency (EPA). 1987–2019. TRI basic data files: calendar years 1987–2017. www.epa.gov/toxics-release-inventory-tri-program/tri-basic-data-files-calendar-years-1987-2019