Prediction of EPA Pesticide Tolerance Using Machine Learning and Publicly Available Data

Sahej Singh

November 2021

Abstract

The EPA Tolerance Level for pesticide/commodity pair (*Tol*) is an important indicator in the environmental risk assessment of common pesticides. This metric is used to tell how much residue in parts-per-million (ppm) is tolerated on food. Pesticides must go through rigorous and costly testing to be approved for public use. For this reason, it is necessary to accurately estimate the *Tol* of a pesticide. This study aims to use publicly available pesticide data, along with collected values of physiochemical properties and molecular descriptors of chemical structures, to develop a reproducible model capable of predicting whether a pesticide can be tolerated. More specifically, the accuracies of models based on a Support Vector Machine, Decision Tree, Logistic Regression, and K-Nearest Neighbors algorithms were compared and evaluated. The experimental results suggest that it is possible to reach a relatively high accuracy using molecular descriptors and specific values from publicly available data. Compared to previous models, these models are more transparent in their methodology and input. Therefore, while not as accurate, the generalizable and modular workflow can be used in the preliminary evaluation of pesticides and reproduced in more data-intensive studies.

1 Introduction

The far-reaching impact of pollution in the Earth's soil and water by chemical runoff is one of the most difficult problems to address in the sphere of climate change. The United States uses more than one billion pounds of pesticides yearly, which are directly applied to the soil to advance crop growth (Taylor, 2021). However, this makes it relatively easy to soak into the soil and pollute the surrounding area. For this reason, there are strict regulations in place, including the EPA's process by which they vet out pesticides based on the parts-per-million that would be tolerated on food (*About Pesticide Registration*, n.d.). This process eliminates or regulates the most toxic pesticides and ensures no ecologically disastrous pesticides will be used on fruits and vegetables. However, this process can take almost a decade to complete, and costs millions of dollars to pesticide companies (Fishel, n.d.). While this process is necessary to vet out intolerable pesticides, it can impede the rollout of further advanced and potentially less toxic pesticides. For this reason, there is

a need for a quicker system that can preliminarily vet out pesticides to save development companies time and money for more research.

While there is some merit in the idea of speeding up the governmental process itself, such sweeping reforms would take years, and there is no indication that the process itself is flawed. This means that there is a need for a change in the chemical evaluation process itself. However, while there is a need, there have been very few proposed solutions to this issue. Thus far, there has been a single study on the prediction of the soil adsorption coefficient of different pesticides, which is a metric that represents the distribution ratio of chemicals between the sediment and aqueous phases of certain pesticides. Using machine learning, this study was able to predict the coefficient to a high accuracy (Kobayashi, Uchida, & Yoshida, 2020). However, the paper did suffer from a lack of cheminformatics knowledge, in the explanation of use cases and the number of descriptors used. For this reason, there is a need for a more reproducible pesticide tolerance model to aid future research. This is because of the usefulness of machine learning in this case.

For this issue, the main draw of machine learning is the data efficiency, availability, and cheminformatics tools available for data creation. For this reason, a more reproducible method using machine learning is needed for the furthering of the prediction of the EPA tolerance levels of pesticides to cut down times and costs of use.

Herein, a method of developing such a prediction model was outlined and tested using multiple models and publicly available data.

2 Methodology

2.1 Materials

To create the models, a data source was needed to train on and describe the pesticides in the selected dataset using physicochemical descriptors. As such, the USDA's PDP database was used as preliminary data (*PDP Database Search*, n.d.). Due to its free availability, amount of data, and the fact that it was created by a government agency, this dataset was chosen to maximize reproducibility and reliability. Initially, the dataset contained over 500,000 rows of data for different pesticide-commodity pairings, with 437 unique pesticides (Table 1). From here, rows with NaN values, repeat pesticides, and pesticides that did not fall under "Tolerated" or "Not Tolerated" tolerances were dropped, making the data come down to 369 rows. When testing the data, it was found that generally, there was an even spread of data, with more "Not Tolerated" than "Tolerated" pesticides.

2.2 Preprocessing: Potential generation of 3D descriptors for higher accuracy

To enrich the feature set, the Chemical Identifier Resolver API (*NCI/CADD Chemical Identifier Resolver*, n.d.) was used to convert the names of each chemical to SMILES, a chemical identifier to describe molecules using strings, to use other Python chemistry packages requiring SMILES input. Two such packages, RDkit and PubChemPy, were used to generate 53 chemical descriptors for each pesticide (*PubChemPy documentation*, n.d.; Landrum, Sforna, De Winter, & Deric, n.d.). Of these descriptors, there were 3D (volume, x, y, and z steric quadrupoles, etc.) and 2D descriptors (molecular weight, XLogP, etc.) generated to get the most complete set of data. In comparison to other papers in the field of chemistry, 3D descriptors are often overlooked. However, they are important to get the clearest picture of an atom since these atoms exist in 3D. After dropping the chemicals that could not be converted to SMILES, and the rows that had no data for the given pesticide, there were 240 rows of data, with two distinct classes (Table 2).

Due to the number of descriptors, Principal Component Analysis (PCA), a method of reducing data-dimensionality while preserving most of the data variation, was used to increase the interpretability of the data for modeling by creating new features from the generated ones. Alongside PCA, the data was scaled to reduce bias, and then split into training and testing data for modelling.

Table 1: Raw data table before cleaning

	Sample ID	Commod	Pesticide Code	Pesticide Name	Test Class	Concentration	LOD	pp_	Confirm 1	Confirm 2	Annotate	Quantitate	Mean	Extract	Determ	EPA Tolerance (ppm)
0	CA9407190123BNCA1	BN	157	Thiabendazole	B	0.1000	0.030	M	MO	NaN	NaN	H	O	NaN	NaN	0.4
1	CA9407190151BNCA1	BN	157	Thiabendazole	B	0.1400	0.030	M	MO	NaN	NaN	H	O	NaN	NaN	0.4
2	CA9409260152BNCA1	BN	157	Thiabendazole	B	0.1600	0.030	M	M	NaN	NaN	H	O	NaN	NaN	0.4
3	CA9409260278BNCA1	BN	157	Thiabendazole	B	0.1700	0.030	M	M	NaN	NaN	H	O	NaN	NaN	0.4
4	CA9410040233BNCA1	BN	157	Thiabendazole	B	0.0500	0.030	M	MO	NaN	Q	H	O	NaN	NaN	0.4
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
524943	CA1906100004SZWA1P	SZ	AKG	Fluopyram	A	0.0850	0.005	M	LU	NaN	NaN	NaN	O	805.0	52.0	2.0
524944	CA1907080547SZWA1	SZ	AKG	Fluopyram	A	0.1000	0.005	M	LU	NaN	NaN	NaN	O	805.0	64.0	2.0
524945	MD1703150001CUFL1	CU	AGX	Mandipropamid	N	0.0093	0.005	M	LU	NaN	NaN	NaN	O	805.0	52.0	0.6
524946	NY1705230071CUFL1	CU	AGX	Mandipropamid	N	0.0053	0.005	M	LU	NaN	NaN	NaN	O	805.0	52.0	0.6
524947	TX1711060201CATX1	CA	AGW	Chlorantraniliprole	I	0.0054	0.005	M	LU	NaN	NaN	NaN	O	805.0	52.0	2.5

524948 rows × 16 columns

2.3 Modeling

For a diverse set of models, multiple modelling methods were tested and optimized on the cleaned EPA data to determine which would be best for predicting the tolerance level. To that end, after the cleaning and feature creation, the data was put through Hyper Optimization using multiple specific models. Hyper Optimization is a method of optimizing every parameter of a model to get the best accuracy possible. Using Hyper Optimization, over 500 evaluations, the first model developed was a Support Vector Machine (SVM) based model, which is a classification model that functions by finding a "boundary" between different

Table 2: Cleaned data with relevant descriptors. NT: Not Tolerated, T: Tolerated

	Pesticide Name	LOD	EPA Tolerance (ppm)	Smiles	ExactMolWt	HeavyAtomMolWt	NumRadicalElectrons	NumValenceElectrons	HeavyAtomCount	NHOHCount	NOCount
0	Procymidone	0.014	NT	CC12CC1(C)C(=O)N(C2=O)c3cc(Cl)cc(Cl)c3	283.016684	273.054	0	94	18	0	3
1	Fenamiphos	0.004	NT	CCO[P](=O)(NC(C)C)Oc1ccc(SC)c(C)c1	303.105801	281.188	0	108	19	1	4
2	Metribuzin	0.030	NT	CSC1=NN=C(C(=O)N1N)C(C)(C)C	214.088832	200.182	0	78	14	2	5
3	Fenamiphos sulfone	0.008	NT	CCO[P](=O)(NC(C)C)Oc1ccc(c(C)c1)[S](C)(=O)=O	335.095630	313.186	0	120	21	1	6
4	Methiocarb	0.043	NT	CNC(=O)Oc1cc(C)c(SC)c(C)c1	225.082350	210.193	0	82	15	1	3
...	...	...	...	...	...	...	...	...	...	...	...
235	Oxamyl oxime	0.001	T	CSC(=N/O)/C(=O)N(C)C	162.046299	152.134	0	58	10	1	4
236	Pyriproxyfen	0.014	T	CC(COc1ccc(Oc2ccccc2)cc1)Oc3ccccn3	321.136493	302.224	0	122	24	0	4
237	Boscalid	0.001	T	Clc1ccc(cc1)c2ccccc2NC(=O)c3ccccc3Cl	342.032668	331.117	0	114	23	1	3
238	Fenpropathrin	0.026	T	CC1(C)C(C(=O)OC(C#N)c2cccc(Oc3ccccc3)c2)C1(C)C	349.167794	326.246	0	134	26	0	4
239	Carfentrazone ethyl	0.003	T	CCOC(=O)C(Cl)Cc1cc(N2N=C(C)N(C(F)F)C2=O)c(F)cc1Cl	411.036431	398.083	0	142	26	0	6

240 rows × 61 columns

classes of data on a plane using different "kernels," or mathematical functions that map the data into higher dimensions. SVM is most useful in high-dimensional planes, which is why it was used for this specific dataset, since over 50 features need to be taken into consideration. To get the best model, multiple different kernels (*linear*, *poly*, *rbf*) were tested on the data, to varying success. While SVM is very useful in this case, it is prone to overfitting, which is why different modelling methods needed to be tested.

Next, the data was used in a hyper-optimized Decision Tree (DT) based classification model. This type of model infers simple if-then statements from the data to make "decisions" on the classification of data among other functions. Because a DT can handle both categorical and numerical data, it was a good fit for the dataset, which involved both types of data. However, DTs can get very complex and time-expensive, which meant that even more models needed to be tested.

To remedy this, the data was then used to make a hyper-optimized Logistic Regression (LR) model as a base test case. This type of model uses a sigmoid function of probabilities to classify data. This model's ubiquity in the realm of classification makes it suited to act as a base for comparison to other models.

Finally, the data was used to create a hyper-optimized K-Nearest Neighbors model as an extra case. This type of model groups data points together based on proximity in a plane. KNN is most well known for being quick, efficient, and highly accurate, which is valuable with such high feature counts.

3 Results and Discussion

The individual model performances were analyzed separately and then compared together (Table 3). The PCA performed on the data yielded a 95.7% information retention when confined to 18 PCA features from 53 overall, which is what the models were trained on.

3.1 Support Vector Machine

In this study, SVM linear, polynomial, radial basis function, and sigmoid kernels were compared and evaluated. Using grid search hyper-optimization of parameters on the C (regularization parameter), gamma (the weight of the regression error), and sigma (kernel parameters), it was found that the SVM with a sigmoid kernel (C=1478) achieved the optimal score.

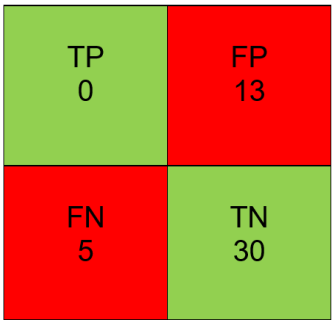

Figure 1: Confusion Matrix for the SVM model.

3.2 Decision Tree

The decision tree was trained with parameter hyper-optimization. The parameter optimization showed that the Decision Tree needed a Gini criterion, random splitter, a minimum of 6 samples to split, a minimum of 4 samples to become a leaf node, and a maximum of log2 features to get the optimal accuracy.

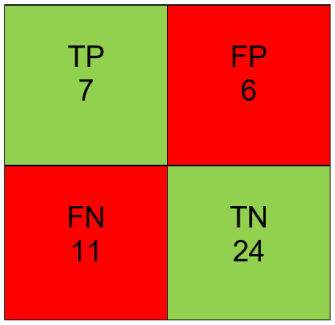

Figure 2: Confusion Matrix for the DT model.

3.3 Logistic Regression

Next, the LR model was trained using hyper-optimization as well. The parameter optimization did not reveal any parameters due to package limitations, however, the accuracy results can be seen in Table 3, along with all the other model results.

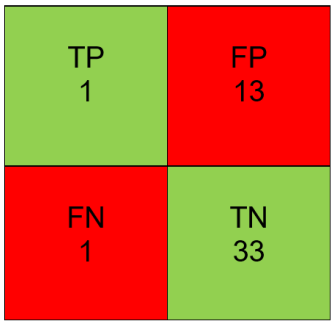

Figure 3: Confusion Matrix for the LR model.

3.4 K-Nearest Neighbors

Finally, a KNN model was developed with parameter hyper-optimization. The optimization exhibited the optimal performance when the 9 nearest neighbors were used with a *Manhattan* distance metric. The number of neighbors suggest that the model is not overfitting due to the small sample size.

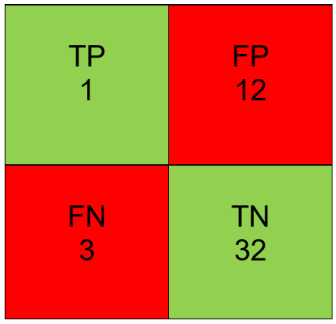

Figure 4: Confusion Matrix for the KNN model.

3.5 Comparison of Models

To confirm the performance of each model, the statistical values of each model in Table 3 were compared. The Logistic Regression model achieved the highest overall accuracy and recall. This is most likely due to LR's suitability and efficiency in classification, and its overall optimization due to its commonality in the realm of classification. However, the Decision Tree achieved a higher precision level, meaning the values were more closely grouped

Table 3: Scores for each model

Model	Accuracy	Precision	Recall
SVM	0.625	0.510	0.620
DT	0.646	0.690	0.650
LR	0.708	0.650	0.708
KNN	0.688	0.600	0.690

together when testing. Looking at the confusion matrix for each individual model, we can see that Decision Tree has the highest accuracy in terms of identifying the true positives without introducing bias in the data due to the small dataset size and the complexity of the problem. Regardless, with easily available cheminformatics packages as RDKit and PubChemPy, and publicly available data, we could achieve above 70% accuracy in our predictions using various models.

4 Conclusion

This study did reach its intended goal of outlining a procedure for the creation and testing of models for the prediction of EPA pesticide tolerance. From cleaning the data, to generating descriptors with rounds of cleaning in between each generation, to the Hyper Optimization of the models trained on the split final dataset, the procedure was outlined throughout this study. The results of this study suggest that this procedure can be used to create a method of preliminarily testing pesticides to save time and money for corporations and the EPA.

However, this study could have benefitted from far more data on the pesticides. Although there was an excess of 50 features, there were only 240 unique pesticides available after cleaning. More data could have contributed to far better model training and accuracies. Furthermore, a deep-learning method would have also resulted in a more complete model. By looking deeper into the features, with more powerful models and equipment, a better model could have been developed.

References

About Pesticide Registration. (n.d.). Environmental Protection Agency. Retrieved from https://www.epa.gov/pesticide-registration/about-pesticide-registration

Fishel, F. M. (n.d.). *EPA Approval of Pesticide Labeling.* Retrieved from https://edis.ifas.ufl.edu/publication/PI203

Kobayashi, Y., Uchida, T., & Yoshida, K. (2020). Prediction of soil adsorption coefficient in pesticides using physicochemical properties and molecular descriptors by machine learning models. *Environmental toxicology and chemistry*, *39*(7), 1451–1459.

Landrum, G., Sforna, G., De Winter, H., & Deric. (n.d.). *RDKit.* Retrieved from https://www.rdkit.org/

NCI/CADD Chemical Identifier Resolver. (n.d.). U.S. Department of Health and Human Services. Retrieved from https://cactus.nci.nih.gov/chemical/structure

PDP Database Search. (n.d.). Retrieved from https://apps.ams.usda.gov/pdp

PubChemPy documentation. (n.d.). Retrieved from https://pubchempy.readthedocs.io/en/latest/

Taylor, J. (2021, Mar). *New Federal Study: Extremely Toxic Pesticide Breakdown Products Found in 90% of Streams Sampled Across U.S.* Retrieved from https://biologicaldiversity.org/w/news/press-releases/new-federal-study-extremely-toxic-pesticide-breakdown-products-found-in-90-of-streams-sampled-across-us-2021-03-26/