Mushroom edibility classification using *k-Nearest Neighbors* algorithm

Daianna González Padilla & Phabel Antonio López Delgado

February 27, 2022

Abstract

This work accomplishes a supervised machine learning classification based on vector spaces to classify mushrooms into edible and poisonous according to their physical features. It was achieved by applying the *k-Nearest Neighbors* (kNN) method to a mushroom dataset already classified so that the machine can memorize their spatial locations and predict the class of future unknown data. It was found that for *k* values of 1,3,5 and 7 the model could predict perfectly all the test data considering all the dataset attributes, however, when one of the attributes with missing information was deleted from the data, the model was no longer able to classify correctly all test data for *k=3,5* and *7*.

Introduction

Mushroom poisoning is a significant and increasing form of toxin-induced disease, its consumption can prove fatal, and most of such cases happen out of ignorance. [1,2] Approximately 95% of fatal mushroom poisoning cases worldwide are caused by amatoxins and phallotoxins mostly produced by species such as *Lepiota* and *Amanita*. [3] Particularly, the genus *Lepiota* includes quite a high number of amatoxins-producing species. [3] Nevertheless, other mushrooms are pointed out by their pharmacological and culinary properties, including the genus *Agaricus*. [4] This genus is even recognized for its antioxidant properties, and safety in therapy and human nutrition. [4] However, mushroom classification and further selection based on its physical attributes is far from trivial, since there is no simple rule for determining the edibility of a mushroom; no rule like “leaflets three, let it be" for Poisonous Oak and Ivy.

Nonetheless, *Supervised Machine Learning* (SML), which involves the fitting of a predictive model from training data, can prove advantageous to classify organisms based on their appearance, including mushrooms. [5] Particularly, the *k-Nearest Neighbor* SML algorithm is a classification approach where a data point is classified on the basis of the ground truth of the k most similar points in a training set using a majority voting rule. [5]

Methods and materials

The input dataset consisted of 8124 instances from hypothetical samples corresponding to 23 species of gilled mushrooms in the *Agaricus* and *Lepiota* families. Each species is identified as *definitely edible*, *definitely poisonous*, or of *unknown edibility* and *not recommended* (classified as poisonous as well). Of these particular instances, 4208 (51.8%) were edible and the rest 3916 (48.2%) poisonous. Each instance had 22 categorical attributes (all nominally valued) from which the 11th had 2480 missing values. These attributes describe the cap, gill, stalk, veil and ring features of the mushrooms, as well as the presence of bruises, their odor, their spore print color, their population and habitat. Based on these physical characteristics, they were classified as *edible* or *poisonous* (See details in **agaricus-lepiota (1).names** file). These mushroom records were drawn from The Audubon Society Field Guide to North American Mushrooms (1981). G. H. Lincoff (Pres.), New York: Alfred A. Knopf. And the dataset was obtained from tha UCI Machine Learning Repository as “Mushroom Data Set”. These data were divided into 3 sets: 70% for training, 25% for testing the model and the remaining 5% for prediction. For the training stage, *k* values used were 1,3,5 and 7; for the prediction stage, this 5% of the original data was used and a wider range of *k* values was set for comparison: 1,3,5,7,9,11,21 and 101.

The *k-Nearest Neighbors* classification algorithm was implemented with version 1.0.2 of *Scikit-Learn*, a free machine learning library used in version 3.7.12 of Python in this work. Code is available in **kNN_mushroom_clasification.ipynb** file.

Results

Initially input data was explored and processed to create a matrix containing the 8124 instances and all their 22 attributes and to obtain a vector of each instance class. Once the dataset was divided into data for training and testing (95%) , 70% of it was used to train the model and 30% to test it, so that 5,401 data was used to train the model and the rest 2316 data to validate it, from these last, 1128 belonged to *Poisonous* class and 1188 to *Edible* class.

For all *k* values used, after training and validating the model, it showed a fully precise predictive power with overall accuracy of 1.0 and scores of *precision, recall* and *F-scores* also of 1.0 for both classes, this means that our model was able to classify correctly all the test data (See **Figure 1**) with all *k* values. Accuracy can be taken into account due to having almost 50% of data for each class, so the dataset is balanced.

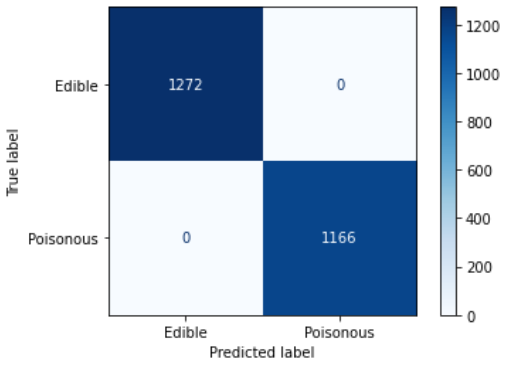

Figure 1: Composition matrix for *k=1,3,5,7*. This figure shows the number of predicted data for each class with respect to their real classes.

Then, test data was graphically represented in a two-dimensional space where data from the same class tended to form groups, though not completely contiguous (See **Figure 2A**). The same was done with the predicted test data (See **Figure 2B**) and unsurprisingly, both graphs show the same groups for each class with all *k* values (See *Supplementary Figure S1*).

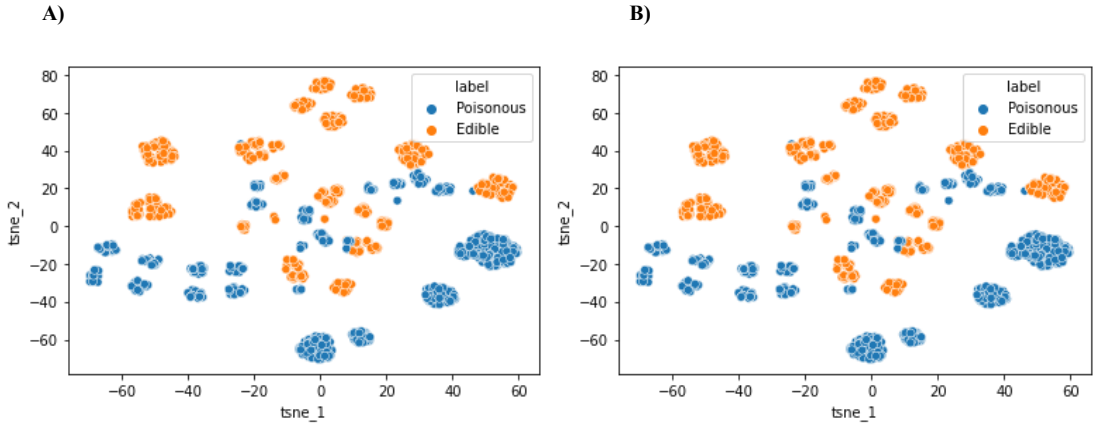

Figure 2: Spatial distribution of test data for *k*=1. **A)** Scatterplot of test data grouped by their real classes. **B)** Scatterplot of test data grouped by their predicted classes with kNN.

Importantly, when the 11th attribute (absent in ~30% of the instances) was not considered, the same scores were obtained with *k=1*, probably because '?' character that accounts for the missing values was also converted to a number when encoding the data to analyze it numerically. However, the scatterplot changes with respect to the last since there are now only 21 attributes (See *Supplementary Figure S2A*). With *k=3, 5* and *7*, all overall scores changed to 0.99 because 3, 6 and 6 poisonous mushrooms were incorrectly predicted as edible, respectively (See *Supplementary Figure S3*). Obviously, the test data distributions in the two-dimensional space changed too (See *Supplementary Figure S2*).

When new data (5% of original data) was tested for the first time, the model predicted all with strong reliability since the only probabilities associated to the predicted classes were {0, 1}, meaning that the model had no problem in predicting the categories to which the elements of the new data belonged. This first prediction was made with the best training value k = 1, but greater k values were used as well, yielding similar results for small *k* values: *1, 3*. When *k = 5*, the probabilities decreased to {0.2, 0.8} and {0, 1}, meaning the model had a tougher time at classifying the new data; presumably due to overfitting. As *k* value got larger, the probabilities decreased; and when *k = 101* for prediction stage, there were probabilities ranging {0,1} and {0.45544554, 54455446}.

Discussion

The machine learning algorithm kNN used had quite an outstanding performance in certain situations. It was seen that with $k = 1$ and 66.5% of the original dataset implemented as the training set was enough to obtain outstanding results in the validation stage with 28.5% of the original data, with an equally good performance in the class prediction of 5 % of the original data.

It is clear that the k parameter is of utter importance, since its alteration had relevant changes in the model reliability and prediction. Given the amount of training data, a $k = 1$ value was enough to yield excellent results; an *accuracy*, *precision*, *recall* and *F-score* of 1. Nevertheless, when the k value increased, the prediction evaluation tended to become less accurate, as an obvious result of the overfitting phenomena. A possible interpretation for this is that, the more *k nearest common neighbors* the model needed, the less specifically it could classify the data, probably because the data from the training set was not diverse enough to have that k number of neighbors at a favorable distance between classes and it did not follow the *contiguity hypothesis*.

In the best-model scenario, the classification was equally well along all the classes; the same happened in the worst-model scenario when the k value was inconsistently high. This might be caused by the good quality of data and the fact that it was properly fractionated into the different training, validation and prediction datasets, respectively.

It must not be forgotten that the evaluation measures need to be interpreted with different parameters, in order to reach the best of these, formulating the best model and, consequently, accomplishing the best results when validating and predicting. Every evaluation measure has its own meaning and can help to the overall interpretation. In the present case, all of them were favorable, reassuring the consistency of the found model.

Conclusion and future prospects

Machine learning classification methods such as *k-nearest neighbors* are useful tools to predict interest variable outcomes; in this case, to predict if mushrooms are harmless or harmful for human consumption. This analysis could be done under the assumption that physical characteristics of mushrooms inform about their biochemical properties, specifically, of their toxicity. Also, this analysis assumes the *contiguity hypothesis* in which mushrooms from the same class tend to group into separated regions, which was not completely the case in this study. In general, these classification supervised techniques represent powerful predictive methods to evaluate not only the classes the instances belong to, but also their most weighty attributes. This, in turn, allows to automate many subjective human tasks into quantitatively objective processes such as object recognition, discrimination and categorization.

Even if these results were promising, this same dataset must be analyzed with other vector space classification ML methods such as SVC and Rocchio, each with its own assumptions and approaches. There is no rule to choose which model to use, the best model fit will ultimately depend on the data and the purposes of each study. Finally, data from other mushroom families must be tested as well since this work only considered mushrooms from *Agaricus* and *Lepiota* families and the model could be able to predict well only mushrooms from those.

References

1. White J, Weinstein SA, De Haro L, Bédry R, Schaper A, Rumack BH, Zilker T. Mushroom poisoning: A proposed new clinical classification. Toxicon. 2019 Jan;157:53-65. doi: 10.1016/j.toxicon.2018.11.007. Epub 2018 Nov 12. PMID: 30439442.
2. Wennig R, Eyer F, Schaper A, Zilker T, Andresen-Streichert H. Mushroom Poisoning. Dtsch Arztebl Int. 2020 Oct 16;117(42):701-708. doi: 10.3238/arztebl.2020.0701. PMID: 33559585; PMCID: PMC7868946.
3. Sarawi S, Shi YN, Lotz-Winter H, Reschke K, Bode HB, Piepenbring M. Occurrence and chemotaxonomical analysis of amatoxins in Lepiota spp. (Agaricales). Phytochemistry. 2022 Mar;195:113069. doi: 10.1016/j.phytochem.2021.113069. Epub 2021 Dec 26. PMID: 34965486.
4. Vinhal Costa Orsine J, Vinhal da Costa R, Carvalho Garbi Novaes MR. Mushrooms of the genus Agaricus as functional foods. Nutr Hosp. 2012 Jul-Aug;27(4):1017-24. doi: 10.3305/nh.2012.27.4.5841. PMID: 23165537.
5. Greener JG, Kandathil SM, Moffat L, Jones DT. A guide to machine learning for biologists. Nat Rev Mol Cell Biol. 2022 Jan;23(1):40-55. doi: 10.1038/s41580-021-00407-0. Epub 2021 Sep 13. PMID: 34518686.

Supplementary Material

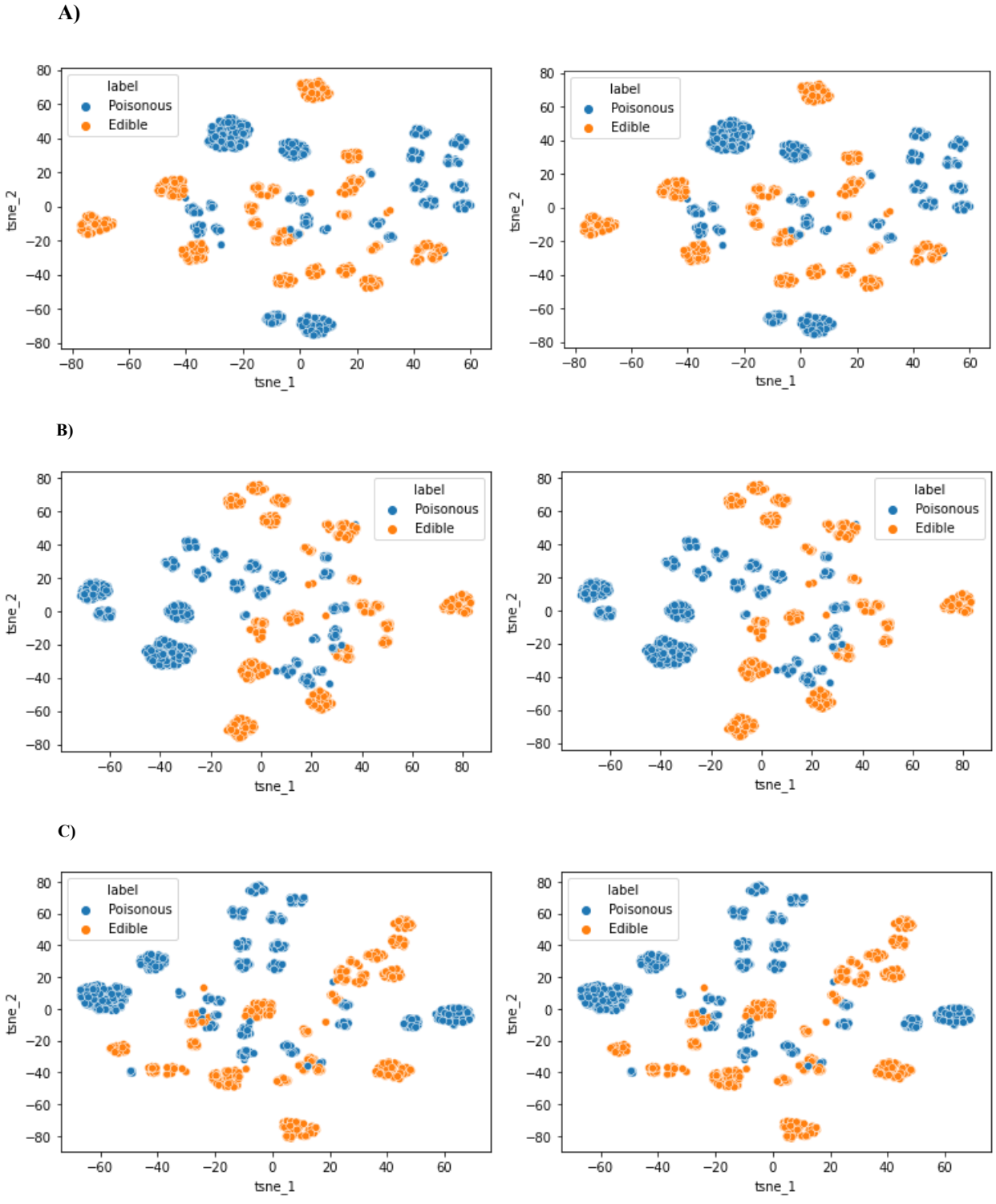

Figure S1: Spatial distribution of test data with all 22 attributes **A)** Scatterplots of test data grouped by their real classes and their predicted classes with *k*=3. **B)** Scatterplots of test data grouped by their real classes and their predicted classes with *k*=5. **C)** Scatterplots of test data grouped by their real classes and their predicted classes with *k*=7.

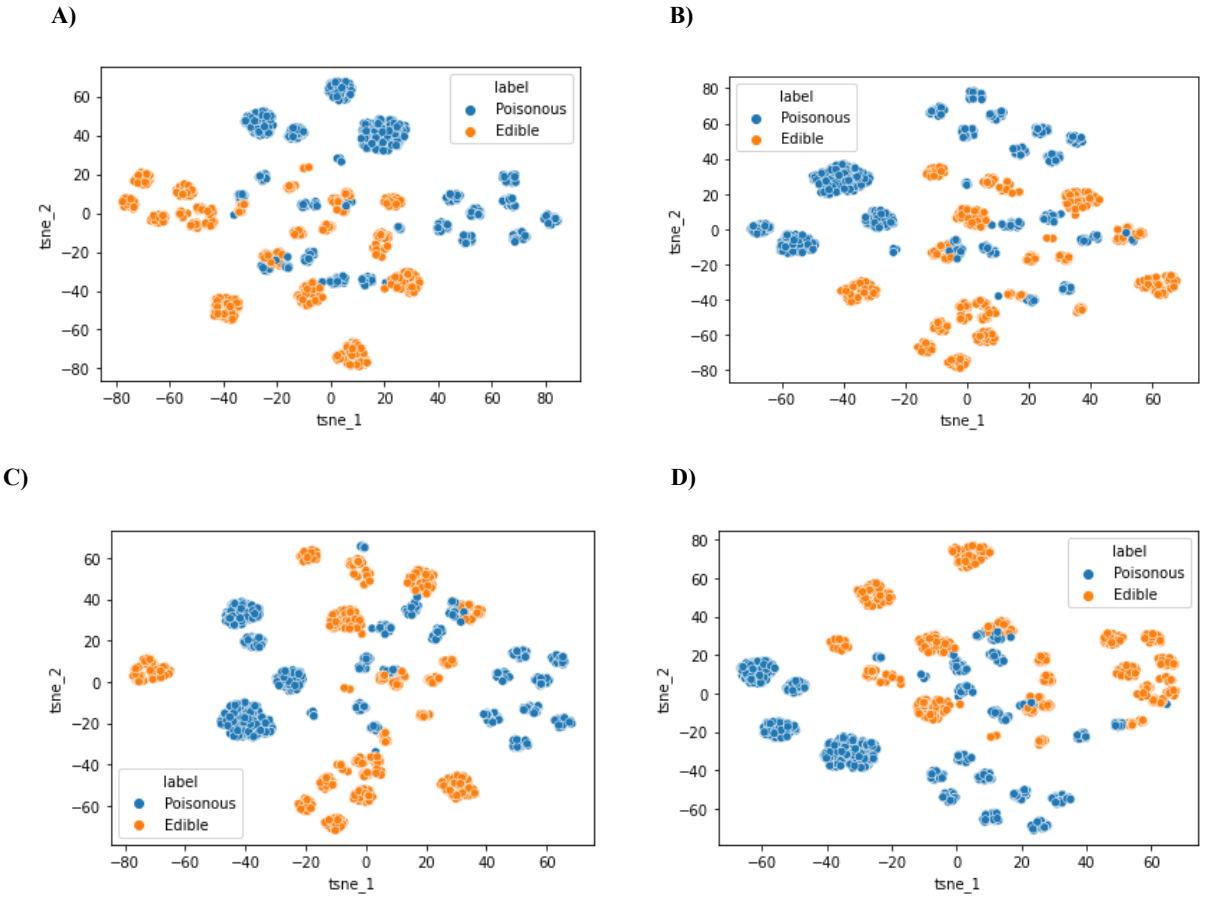

Figure S2: Spatial distribution of test data without 11th attribute **A)** Scatterplot of test data grouped by their real classes with *k*=1. **B)** Scatterplot of test data grouped by their real classes with *k*=3. **C)** Scatterplot of test data grouped by their real classes with *k*=5. **D)** Scatterplot of test data grouped by their real classes with *k*=7.

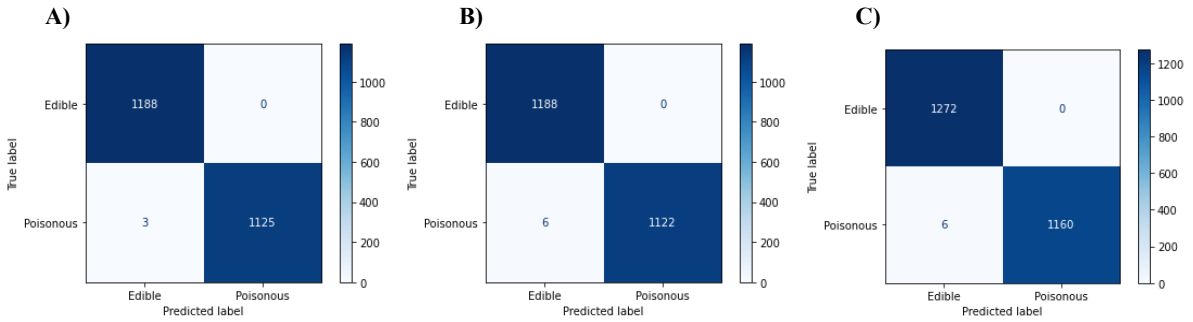

Figure S3: Confusion matrices for data without 11th attribute. **A)** kNN for *k*=3 predicted 3 poisonous mushroom as edible. **B)** for *k*=5 predicted 6 and **C)** for *k*=7 predicted 6.