DATA 240 Data Mining and Analytics

Final Project: Pre-report

“Mushroom Classification”

Group: 4

Greeshma Venkatesh
Neelima Jagtap
Rishikumar Ravichandran
Rushikesh Jagtap
Sougandhika Ayathamaraju

Introduction

Motivation and Dataset Description

In this technological era, it is significant for mankind to get adequate nutrients in order to live a healthy life. Nature has a plethora of edible resources that are highly rich in nutrients. One such important natural food is mushrooms. Mushrooms are formed from macrofungi that can be both hypogeous and epigeous [1]. Mushrooms are considered to have an inclined medicinal value and some types are even considered as an agent to cure one of the Chronic disease cancers [2]. According to research conducted in 2020, the first mushroom formed on the earth was around 810 million years ago [3]. In contrast to having numerous positive sides, the fact is not all mushrooms are edible, which means there are many poisonous mushrooms that cause internal organ failure to end up taking human lives [4].

On taking count of the death caused by fungi worldwide, it is recorded that poisonous mushrooms are the reason for 9 out of 10 deaths [5]. This paves the way for many researchers to study about mushrooms and classify them into edible which is non-poisonous and non-edible which is poisonous based on the features of the mushroom. Biologists have classified the list of poisonous mushrooms based on their morphology and the scientific family of the mushroom. But, with the development of the data science field, it is much easier to classify the non-edible mushrooms that will be useful for mankind. The goal of this project is to identify the poisonous and non-poisonous mushrooms by collecting the mushroom data and classify them based on important features by implementing data mining concepts.

In order to achieve the goal of the project, the most prominent step is to collect the appropriate dataset that includes the detailed features of mushrooms. In this project, the data is collected from the UCI machine learning repository that was sourced by the Audubon Society Field Guide to North American Mushrooms. The dataset holds a total of 8124 rows with 22 attributes that include cap-shape, cap-color, odor, gill-size, gill-color, stalk-root, ring-type, and habitat of the mushroom [6]. All the 22 attributes in the dataset contain the perfect feature of mushrooms and none of the attributes are named irrelevant or changed using PCA. In this dataset, a total of 23 different species of mushrooms are taken in count from the Agaricus and Lepiota family.

Literature Survey

Over the years researchers have tried various methods and techniques to classify different types of mushrooms but haven't obtained results that can be highly relied upon. People rely on their intuition and experience to identify and classify mushrooms as poisonous and not poisonous. They have performed this classification considering the shape, color, odor, and secretion skins. The accuracy for using this method has proven to provide very less accurate results [7].

As mushrooms are one of the best sources of nutrients and in high demand for research in the medical science field, finding out a highly reliable classification approach is essential. Zahan et al. in their recent work have proposed deep learning approachs to classify mushrooms as edible, inedible or poisonous. They have used algorithms such as InceptionV3, Resnet50, and VGG16 to acquire the best accuracy. As extracting the features from an image is an essential component for classification. To evaluate the performance metrics they extracted features from images both with backgrounds and without background. Authors also performed the comparison on raw image and contrast images in order to obtain the best test effectiveness. Upon comparing the performance metrics of CNN architures authors concluded that the InceptionV3 outperforms well in comparison with other algorithms and provides the best accuracy of 88.40% when used against contrasted images [8].

Wibowo et al. in their research work have used and compared two classification algorithms in data mining, they are Decision Tree (C4.5) and Support Vector Machine (SVM). The study method that they have used is an experiment with the assisted tool of WEKA that has been tested in the comparison of the two algorithms. For the classification purpose they have considered mushroom dataset from UCI machine learning repository mainly belonging to two families Agaricus and Lepiota. They performed Kappa statistics on both the algorithms and also evaluated the performance of the algorithms on various parameters such as Mean absolute error (MEA), Root mean squared error (RMSE), Relative absolute error (RAE), Root relative squared error (RRSE). Based on these results and analyzing the classifiers outputs the authors concluded that the Decision Tree i.e C4.5 algorithm has the highest accuracy compared to the Support Vector Machine. It provided highest accuracy and also high processing speed [9].

According to Vanitha et al., Mushrooms have been discovered to be one of the most nutritious foods, including high levels of proteins, vitamins, and minerals. It's high in antioxidants, which help individuals avoid heart disease and cancer. But customers are unaware about the category of mushroom. Some mushrooms are good for eating and others are poisonous to health. Around 45000 mushroom species have been discovered around the world [10].

Only a few mushroom species have been discovered to be edible. Some of them are extremely dangerous to consume. To distinguish between edible and dangerous mushrooms, some data mining techniques are applied to the mushroom dataset obtained from the UCI Machine Learning Repository. Weka is a machine learning platform for pre-processing, analyzing, categorizing, displaying, and forecasting data.

As a result, the Wrapper and Filter methods in Weka are used to find the best attributes for classification in order to select the traits that aid in better mushroom classification. For enhancing the classification of edible and poisonous mushrooms, the metrics 'odor' and 'spore print color' were chosen as the best. Following the identification of the critical attributes, a classification is performed, and a decision tree is constructed based on the features chosen, with Precision, Recall, and F-Measure values examined [11].

Nadya Chitayae and Andi Sunyoto Because they don't realize the mushrooms are harmful, many people acquire food poisoning. Poisonous mushrooms have been linked to incidences of poisoning in several nations. Due to the enormous number of mushrooms and their comparable qualities, distinguishing between edible and dangerous mushrooms is difficult. Data mining science, namely classification, can be used to identify the type of mushroom by helping to detect essential patterns from millions or even billions of data records. The K-Nearest Neighbor and Decision Tree approaches were used to classify the different species of mushrooms. The results of the two approaches were compared to evaluate which method performed better in classifying the type of fungus. In this paper the author compares SVM, KNN, Random forest algorithms. They found that K near neighbor has the best accuracy compared to the other algorithms.

The proposed method is effective and the most acceptable method for the categorization of mushroom varieties, according to an experimental analysis done on the UCI Mushroom dataset. The results show that the Decision Tree technique outperforms the others, with an

accuracy of 0.9193 (91.93%), precision of 0.9227, recall of 0.9193, and an F1 score of 0.9210 [12].

Classification Methods

Classification is implemented to categorize the input given data into different labels or classes. This is often performed on datasets that feature many attributes and would want to categorize them based on these attributes. This project deals with different types of mushrooms and attributes of these mushrooms like gill size, color etc. The goal of this project is to categorize these wide varieties of mushroom into poisonous and non-poisonous mushrooms. The final target determines if the mushrooms are edible or not. The best method to categorize is by implementing classifiers and training them to assign the target labels. In this project, three different classifiers will be trained and compared to understand the best fit model. The classifiers are selected based on the literature and applicability.

Random Forest Classifier

Random Forest is an ensemble model that combines many individual decision trees. The classifier generates many different trees with varying sample size and different numbers of features selected with replacement randomly. Each tree predicts the target class and the majority in the model is the final decision. The importance of random forest relies on the low correlation of the trees. The low correlation helps in minimizing the error among the trees. This is a powerful classifier that outperforms many classifiers in minimizing the error and predicting accurately. In this project the Random Forest algorithm will also be used to analyze the importance of different features and classify the mushrooms accurately [13].

Naive Bayes Classifier

Naive Bayes is derived from Bayesian Theorem and is a supervised learning algorithm. We get ‘Naive’ from the algorithm's assumptions that there is conditional independence between features when there is a known value for the class variable [14]. Validity of mathematical mapping between vectors of features is to be determined and this calculation can get computationally expensive and complex. To reduce this complexity of calculations it assumes that two features have no dependence on each other, that is one does not affect the other. This assumption of the features being independent may not always be correct but works fine with

most of the real world problems. The ability to calculate necessary parameters with minimum training data faster than compared to similar algorithms makes Naive Bayes a powerful classification algorithm. It grounds the problematic situations arising from the curse of dimensions. They have outstanding performance when it comes to spam filtering and document classification [15].

K Nearest Neighbor

K Nearest Neighbor is a supervised learning algorithm which is used for regression as well as classification and is heavily used for pattern recognition and data mining problems. It is non-parametric in nature, that is it does not make presumptions of the prior data [16]. Labeled data is provided after which it forms clusters of nearly lying data points. When test data is to be predicted the algorithm calculates the difference between the test data and the previously provided training data then labels the test data by the minimum distance between the data point and the clusters. The mechanism is highly sensitive to the value of k and euclidean distance. These values are computationally expensive as they are calculated at the time of training data classification. There are advanced versions of KNN which solves this problem by determining the optimal value for k. Works well when there is no domain knowledge of the data [17].

References

[1] Wikipedia Contributors, "Edible mushroom," *Wikipedia*, Sep. 27, 2019. https://en.wikipedia.org/wiki/Edible_mushroom

[2] S. Ismail, A. R. Zainal, and A. Mustapha, "Behavioural features for mushroom classification," *IEEE Xplore*, Apr. 01, 2018. https://ieeexplore.ieee.org/abstract/document/8405508

[3] "First mushrooms appeared earlier than previously thought," *phys.org*. https://phys.org/news/2020-01-mushrooms-earlier-previously-thought.html

[4] "Heads Up: Those Wild Mushrooms Growing in Your Backyard Could Be Toxic," *Cleveland Clinic*, Sep. 30, 2018. https://health.clevelandclinic.org/heads-up-those-wild-mushrooms-growing-in-your-backyard-might-be-toxic

[5] "Mushroom poisoning | betterhealth.vic.gov.au," *www.betterhealth.vic.gov.au*. https://www.betterhealth.vic.gov.au/health/healthyliving/fungi-poisoning

[6] "UCI Machine Learning Repository: Mushroom Data Set," *Uci.edu*, 2019. https://archive.ics.uci.edu/ml/datasets/Mushroom

[7] T. Fukuwatari, E. Sugimoto, K. Yokoyama, and K. Shibata, "Establishment of animal model for elucidating the mechanism of intoxication by the poisonous mushroom Clitocybe acromelalga," Journal of the Food Hygienic Society of Japan, vol.42, no.3, pp.185– 189, 2001.

[8] Zahan, Nusrat, et al. "A Deep Learning-Based Approach for Edible, Inedible and Poisonous Mushroom Classification." *IEEE Xplore*, International Conference on Information and Communication Technology for Sustainable Development, 2021, https://ieeexplore.ieee.org/abstract/document/9396845/figures.

[9] Wibowo, Agung, et al. "Classification Algorithm for Edible Mushroom Identification." *IEEE* , International Conference on Information and Communications Technology, 2018, https://www.researchgate.net/publication/324955852_Classification_algorithm_for_edible_mushroom_identification.

[10] E. S. Alkronz, K. A. Moghayer, M. Meimeh, M. Gazzaz, B. S. Abu-nasser and S. S. Abu-naser, "Prediction of Whether Mushroom is Edible or Poisonous Using Back-propagation Neural Network", *Int. J. Acad. Appl. Res*, vol. 3, no. 2, pp. 1-8, 2019, [online] Available: http://www.ijeais.org/ijaar.

[11] M. Husaini, A Data Mining Based On Ensemble Classifier Classification Approach for Edible Mushroom Identification, pp. 1962-1966, July 2018.

[12] H. Liu and S. Zhang, "Noisy data elimination using mutual k-nearest neighbor for classification mining", *J. Syst. Softw*, vol. 85, no. 5, pp. 1067-1074, 2012..

[13] Breiman, Leo. "Random Forests." *SpringerLink*, Kluwer Academic Publishers, 2001, https://link.springer.com/article/10.1023/A:1010933404324.

[14] "1.9. Naive Bayes — scikit-learn 0.21.3 documentation," *Scikit-learn.org*, 2019. https://scikit-learn.org/stable/modules/naive_bayes.html

[15] F. -J. Yang, "An Implementation of Naive Bayes Classifier," 2018 International Conference on Computational Science and Computational Intelligence (CSCI), 2018, pp. 301-306, doi: 10.1109/CSCI46756.2018.00065.

[16] K. Taunk, S. De, S. Verma and A. Swetapadma, "A Brief Review of Nearest Neighbor Algorithm for Learning and Classification," 2019 International Conference on Intelligent Computing and Control Systems (ICCS), 2019, pp. 1255-1260, doi: 10.1109/ICCS45141.2019.9065747.

[17] "sklearn.neighbors.KNeighborsClassifier — scikit-learn 0.22.1 documentation," 2019. https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html