Journal of Mathematics and Computing Science, 2021,
Volume 7, No 2, 23-30

Analysis of Identifying Mushroom Species using RapidMiner

Fatnin Hanun Nor Sarizan[1], Muhammad Firdaus Mustapha[2*], Omar Kairan[3], Nurulain Nabilah Mohd Bakhary[4], Nurin Hannani Azmira[5], Siti Haslini Ab Hamid[6]

[1,2,3,4,5]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Kelantan, Bukit Ilmu, 18500 Machang, Kelantan, Malaysia,
[6]FH Training Center, 16800 Pasir Puteh, Kelantan, Malaysia

[*] mdfirdaus@uitm.edu.my

Abstract: Identify analysis basically consists of defining the characteristics and categorizing the class of dataset. This study helps in improving the information of mushroom characteristics by removing the unnecessary information and identify the species of mushroom; edible or poisonous. The process to identify mushroom species data is started by using Tableau to visualize the data. The data is prepared first before undergoing the process of modelling using decision tree as the descriptive analysis. Testing the model and cross validation are applied in this process to get the predictive analysis of mushroom species data. Then, the result is gathered by checking the accuracy of the performance of the data. Overall process is done using RapidMiner in order to get an accurate performance of the data. The process of visualization and modelling is needed to analyse the data and get an accurate performance of the result. Therefore, the mushroom species will be easily classified and categorized based on the characteristics either it is edible or poisonous. The result of the proposed model achieved 99.81% accuracy for predicting the species of the mushrooms; edible or poisonous.

Keywords: Classification, Decision Tree, Edible Mushroom, Poisonous Mushroom, RapidMiner

1 Introduction

Agaricus and Lepiota are genus of gilled mushrooms in the family of Agaricaceae, that both have poisonous and edible species with over 200 until 400 species. Agaricus is the common “button” mushrooms and field mushrooms can be easily found on fields [1]. This species is characterized with having a fleshy cap from the underside of which grow several radiating plates or gills which produce naked spores. They also have chocolate-brown spores. Members of Agaricus also have a stipe or stem, which elevates it above objects, on which the mushrooms grow, or substrate, and a partial veil, which protects the developing gills and later forms a ring or annulus on the stalk. Meanwhile, for Lepiota, they have whitish spores, typically with scaly caps and a ring on the stem [2].

Mushrooms have some benefits such as can increase body immunity, can prevent cancer and useful in weight loss [3]. Maria et al. [4] discussed mushroom effects on human health and treatment for diseases. Mushrooms contain nutrients such as potassium, selenium, riboflavin, vitamin D, niacin, proteins and fiber. Moreover, mushrooms act as antibacterial that immune the body system and some mushroom extracts are used to promote health. However, there are certain mushrooms that are poisonous and not suitable to be eaten. Therefore, it is an important task to distinguish between poisonous and edible mushrooms.

Nowadays, there are many researches that have studied mushroom species. They applied various algorithms to classify the mushrooms such as support vector machine (SVM), random forest, nearest neighbors and artificial neural network [3], [5], [6]. They also identified the attributes for mushrooms that can contribute to high performance accuracy for classification of mushrooms.

Therefore, this study proposes to create a model to identify the species of mushrooms whether they are edible or poisonous. Mushroom dataset from UCI Machine Learning Repository [7] was used in this study that contains 8124 examples with 22 attributes. It includes the descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Lepiota and Agaricus family. Each species will be identified as definitely poisonous, edible, or unknown. The method used in this study was classification where data mining model was used to assign items in a collection to classes or categories that they belong to. The goal of classification is to accurately predict the target class for each item in the data [8]. As for this study, the mushrooms will be divided into cap colours, cap size, stalk roots, etc. The result for this study will be shown by using RapidMiner. This study applied a model in RapidMiner to show the accuracy of the data.

The rest of this paper will be organized as follows. Section 2 reviews about related researches and Section 3 explains about methodology of the study. Then, the result will be shown more accurately and detailed in Section 4. Finally, Section 5 gives the conclusion of the study.

2 Literature Review

Research on mushroom species have been studied by several researches. For an example, Smolskaitė et al. [9] proposed an evaluation on the integrated values of antioxidant scores and antimicrobial agents in mushrooms. They briefly reviewed to apply biorefinery approach in order to valorise some wild mushrooms growing in Midi-Pyrénées region. The antioxidant properties of mushroom extracts were evaluated by methods of ABTS•+, DPPH•, scavenging capacity, ferric reducing antioxidant power (FRAP), oxygen radical absorbance capacity (ORAC) and Folin-Ciocalteu total phenolic content (TPC). Meanwhile, antimicrobial activity used agar diffusion method. The result shows that Phaeolus schweinitzii (9.62 ± 0.03 in DPPH•; 109 ± 3 in FRAP; 164 ± 1 in ABTS•+; 340 ± 3 in ORAC assays) and Inonotus hispidus (9.5 ± 0.04 in DPPH•; 54.27 ± 0.46 in ABTS•+; 88.31 ± 1.96 in FRAP; 290 ± 1 in ORAC assays) produced the highest antioxidant and the more effective mushroom extract against tested microbial species is also Inonotus hispidus. The natural antioxidants and antimicrobial agents can be found in some wild mushrooms. It is safe and edible. Katharina et al. [10] discussed the circumstances of exposure to mushrooms. To determine the clinical relevance of mushroom poisoning for humans in Central Europe, retrospective case study is done by analysing the questionnaire on human exposure to mushrooms gathered from the Swiss Toxicological Information Centre. The report is mostly because of the toxicity from the edible species of mushrooms. Humans who experienced high exposure to amatoxin poisonings in edible mushrooms come with the symptom of ingestion. Safe precaution should be taken before eating various species of mushrooms to avoid from choosing toxic containing mushrooms.

Moreover, Reed et al. [11] described robotic harvesting of high-quality mushrooms. Two stages are involved to develop an automated and emerging mushroom harvester technology which are laboratory rig and pilot harvester. The overall system is control by using a single 468 DX33 MHz personal computer. The successful pick attempts of the mushrooms are over the average of 80%. The new technological development automatically harvests high-quality mushrooms in an easier way. Ayşenur Gürgen et al. [12] discussed the increase in mushroom consumption. Fuzzy analytical hierarchy process (AHP) is applied to determine mushroom preferences. The study is conducted based on the selection criteria of mushroom purchased; place of purchase, packaging, appearance, growing type and quality. Growing type gets the highest percentage according to the criteria weights calculated using Buckley's method. Al-Momany and Saleh [13] identified different species of wild mushrooms. They studied on Agaricus species in North Cyprus by collecting different species of wild mushrooms.

The Agaricus placomyces is poisonous while Agaricus bresadolianus cannot be identified whether it is edible or poisonous. The other seven Agaricus species are edible. The study helps other Agaricus collectors in acknowledging the different species of mushrooms.

Based on the above studies, it is an important task to distinguish between edible and poisonous mushrooms. There are several researchers who used artificial intelligence method to classify or identify mushrooms whether they are edible or poisonous. For instance, Clara Eusebi et al. [14] used data mining technique to analyse mushroom database and to increase the accuracy of machine learning. The data mining tool, Weka, is used in this process and the human-machine interactive application is finally developed. Maurya and Singh [3] developed a method for mushroom classification based on its texture feature using machine learning. The performance of their proposed method is 76.6% by using SVM classifier. Shuhaida et al. [5] applied Principal Component Analysis (PCA) algorithm to choose the best features for mushroom classification. They selected odour as the best feature for classification. Furthermore, Alkronz et al. [6] used multi-layer ANN model to train and classify the mushrooms whether it is poisonous or edible. They identified important attributes and achieved 99.25% of performance accuracy.

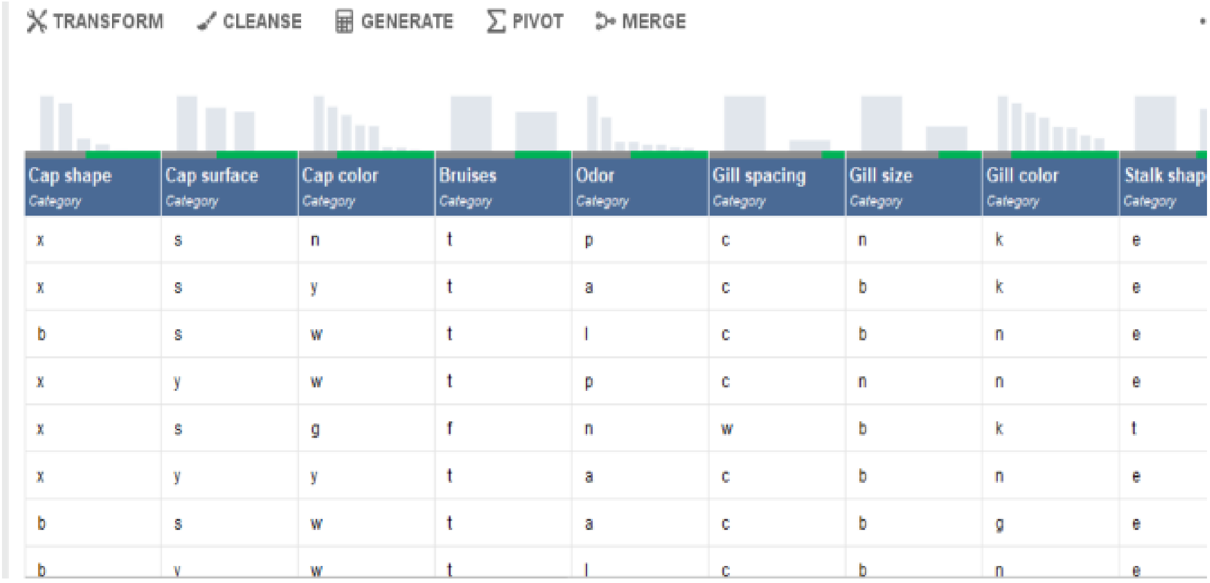
Therefore, this study proposes to create a classification model to identify the species of mushrooms whether they are edible or poisonous by using artificial intelligence method.

3 Method

In this study, RapidMiner Studio was used to analyze the data. RapidMiner is a software package that allows text mining, data mining and predictive analytics [15]. RapidMiner Studio helps to access, load, and analyze any type of data such as structured and unstructured data. Below are the steps involved in this study that include data preprocessing, decision tree model, applying a model, testing and validation.

A Data preprocessing

After getting the Mushroom dataset from UCI Machine Learning Repository, a process to remove the unwanted and unimportant data was performed. The less necessary attribute was removed using data cleaning to get a better result in the next process. Figure 1 shows the result of data cleaning at the turbo prep.



Cap shape Category	Cap surface Category	Cap color Category	Bruises Category	Odor Category	Gill spacing Category	Gill size Category	Gill color Category	Stalk shap Category
x	s	n	t	p	c	n	k	e
x	s	y	t	a	c	b	k	e
b	s	w	t	l	c	b	n	e
x	y	w	t	p	c	n	n	e
x	s	g	f	n	w	b	k	t
x	y	y	t	a	c	b	n	e
b	s	w	t	a	c	b	g	e
b	y	w	t	l	c	b	n	e

Figure 1: Processing the Data

B Model

i. Decision Tree Model

Data obtained from previous operations will be used to create a decision tree. Decision tree is a predictive model or tool that supports decisions [16]. It is known to deliver accurate inferences by using designs, design models, or representations that follow a tree-like structure. The decision tree is created using Filter Examples and Decision Tree operators. The filter examples filter out all the data from the dataset that contains exclamation mark ('?'). Next, it will generate the decision tree of the above dataset. Figure 2 and Figure 3 consist of the decision tree obtained from RapidMiner. From the figure, we can see which mushrooms are edible(e) or poisonous(p) depending on the odor of the mushrooms.

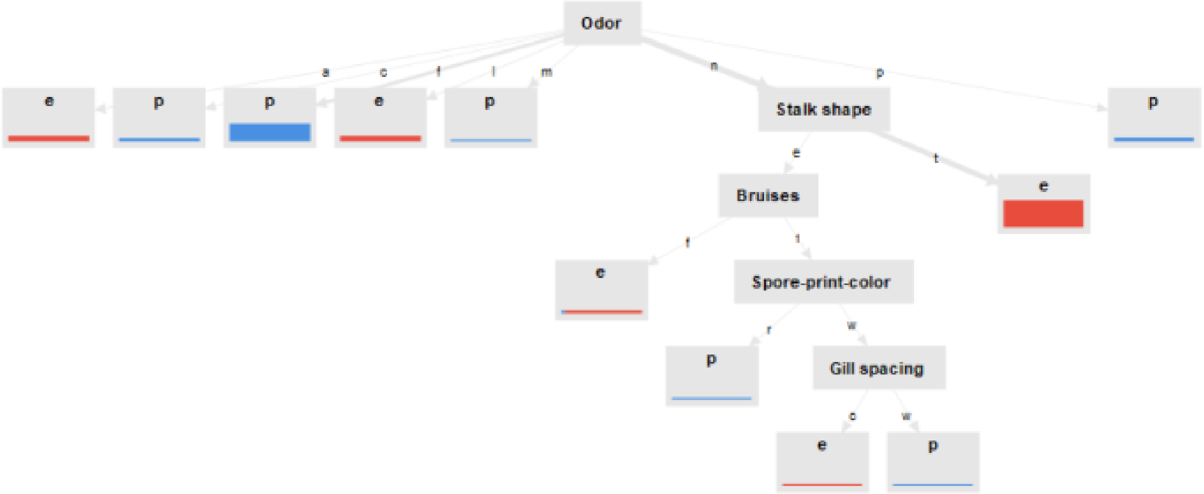
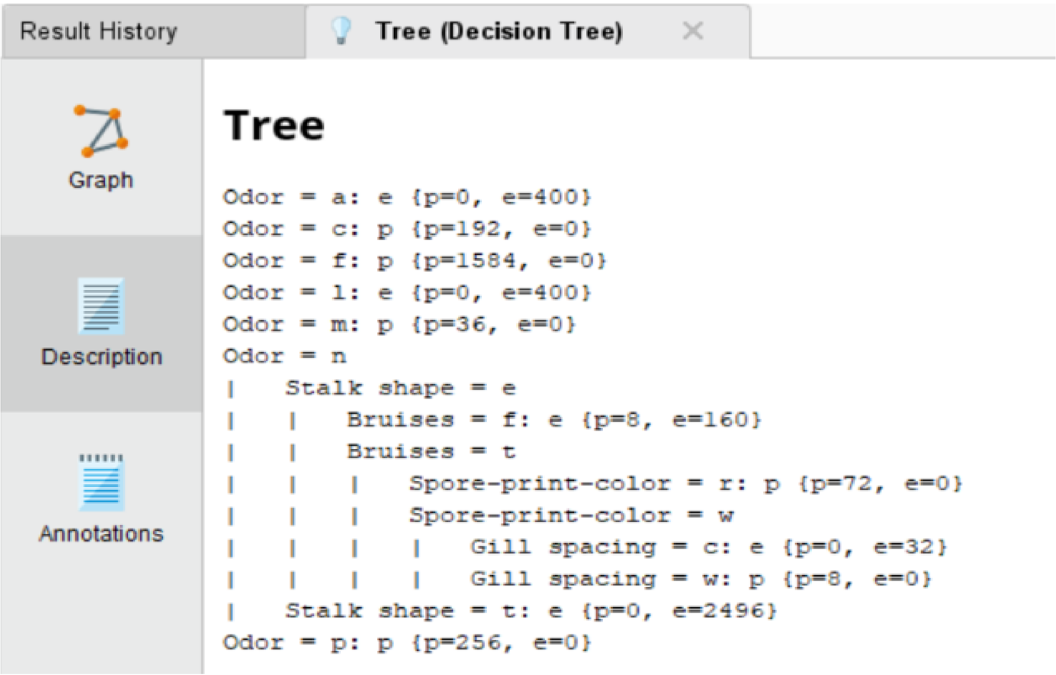


Figure 2: Classification Diagram



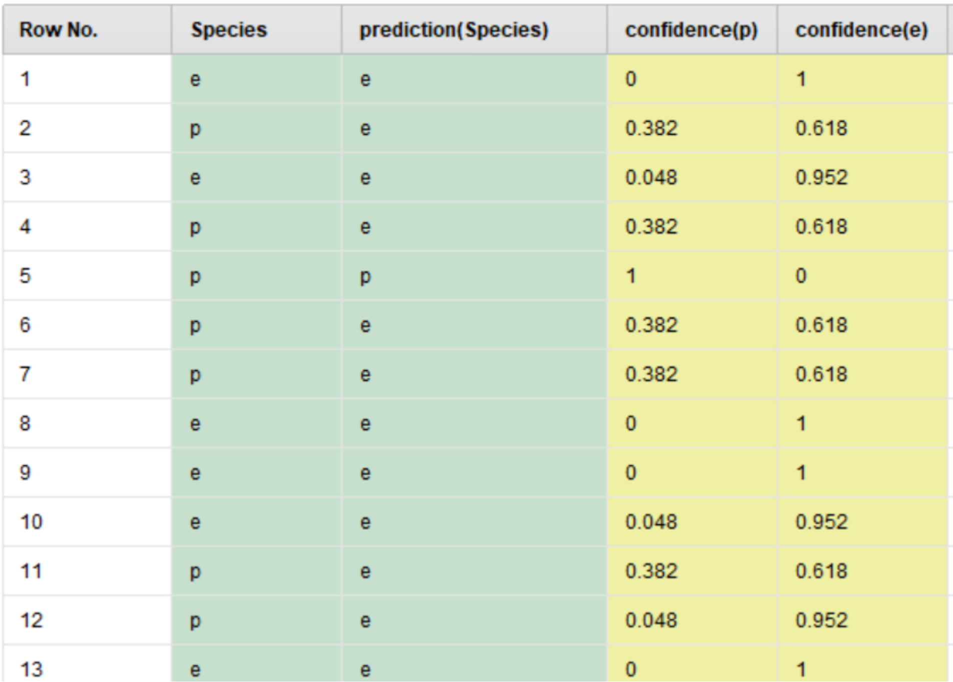
Tree

```
Odor = a: e {p=0, e=400}
Odor = c: p {p=192, e=0}
Odor = f: p {p=1584, e=0}
Odor = l: e {p=0, e=400}
Odor = m: p {p=36, e=0}
Odor = n
|   Stalk shape = e
|   |   Bruises = f: e {p=8, e=160}
|   |   Bruises = t
|   |   |   Spore-print-color = r: p {p=72, e=0}
|   |   |   Spore-print-color = w
|   |   |   |   Gill spacing = c: e {p=0, e=32}
|   |   |   |   Gill spacing = w: p {p=8, e=0}
|   Stalk shape = t: e {p=0, e=2496}
Odor = p: p {p=256, e=0}
```

Figure 3: Tree Structure

ii. Applying the Model

After obtaining the decision tree, we need to predict the accuracy of the dataset. In this process the process operators such as Multiply, Filter Examples, Decision Tree and Apply Model were used. Multiply operators were used in this process because in RapidMiner, one output-port can be connected to only one other port. So, multiply operators help to create a copy of the data to use for the model and one more to the filter examples from the decision tree. Filter examples that did not contain '?' was connected to decision tree while filter examples that contain '?' was connected to apply model operators. Then, after we ran the process, we got the prediction of mushroom species and the precision value. Based on Figure 4, species column was from our original data while the prediction (species) column was predicted by the system using RapidMiner. The confidence (p, e) column tells us about the precision value for edible and poisonous species of mushrooms.



Row No.	Species	prediction(Species)	confidence(p)	confidence(e)
1	e	e	0	1
2	p	e	0.382	0.618
3	e	e	0.048	0.952
4	p	e	0.382	0.618
5	p	p	1	0
6	p	e	0.382	0.618
7	p	e	0.382	0.618
8	e	e	0	1
9	e	e	0	1
10	e	e	0.048	0.952
11	p	e	0.382	0.618
12	p	e	0.048	0.952
13	e	e	0	1

Figure 4: Prediction Analysis

iii. Training and Validation

In this section, we used the filter examples and cross validation operators. When we double clicked on cross validation operators, it split into 2 processes which were training and testing as shown in Figure 6. For training, the decision tree operators were used while for testing, we used the apply model and performance operators. The number of folds used in this process was 10 which means that the datasets were divided into 10 parts. Only one was used in the testing process and the rest of it were for training.

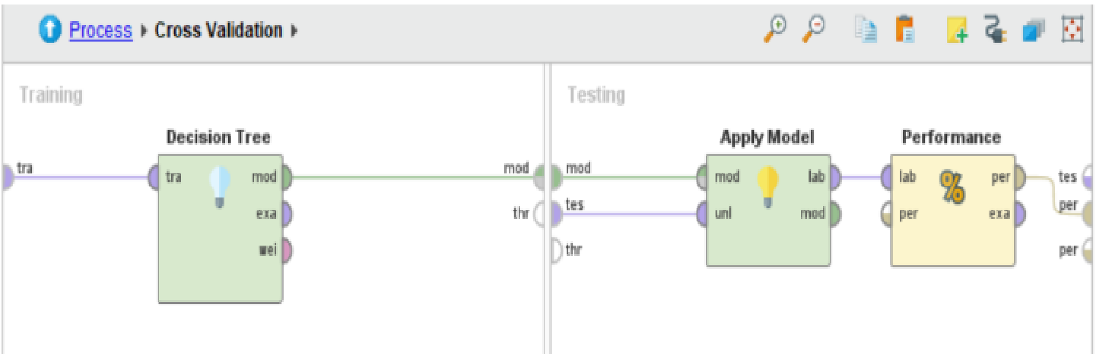
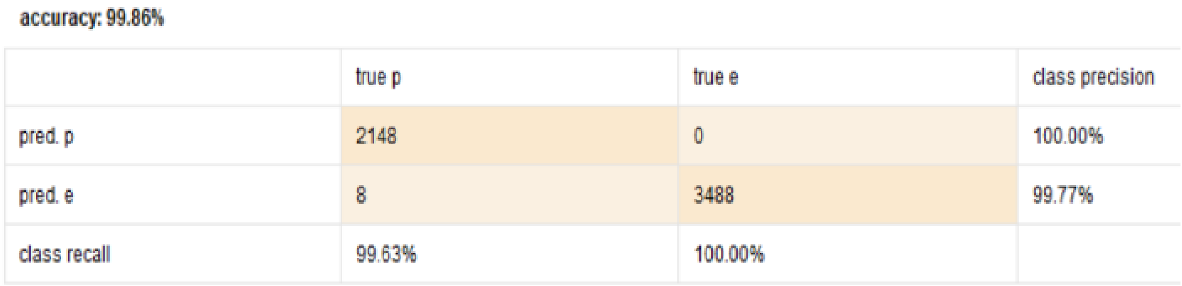


Figure 6: Cross Validation

iv. Evaluating the Model

In this section, we want to know about the accuracy of the model. Using the same operators from the previous step and adding the performance operator, we got the result as depicted in Figure 5. However, the result obtained is not the final result as it only gave the output of the given dataset that we want to test. So, the validation process was needed for unknown dataset and to know the real accuracy.



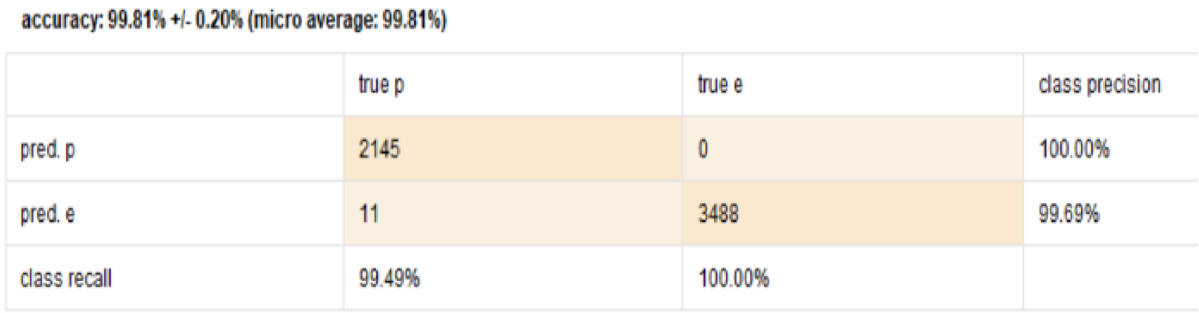
accuracy: 99.86%

	true p	true e	class precision
pred. p	2148	0	100.00%
pred. e	8	3488	99.77%
class recall	99.63%	100.00%	

Figure 5: Accuracy of testing model

4 Results and Discussion

The result of the accuracy that were obtained from cross validation that involved training and testing is shown in Figure 7. This is the accuracy for the species of mushrooms which are edible or poisonous. The total for poisonous mushrooms was 2156 (2145+11) and for edible mushrooms it was 3488. We can observe that 2145 predicted poisonous mushrooms were actually poisonous and 3488 predicted edible mushrooms were actually edible. The class precision column tells us about the percentage of predictions that were correct. From the class recall row, the model was able to predict 99.49 out of 100 for poisonous mushrooms and predicted 100 out of 100 for edible mushrooms.



accuracy: 99.81% +/- 0.20% (micro average: 99.81%)

	true p	true e	class precision
pred. p	2145	0	100.00%
pred. e	11	3488	99.69%
class recall	99.49%	100.00%	

Figure 7: Accuracy using Cross Validation

It can be analyzed from the output that the proposed model has an accuracy of 99.81% for predicting the species of the mushrooms (edible or poisonous). In the case of predicting edible mushrooms, the accuracy was 100% and for poisonous mushrooms it was 99.49%. In comparison, Alkronz et al. achieved a lower accuracy result of 99.25% using the same dataset [6].

5 Conclusion

In conclusion, this study presented descriptive analytics and predictive analytics by using RapidMiner. These analytics were used to summarize the data, to predict the accuracy of the data and to visualize the data. The summarization of data was built using Decision Tree while predicting accuracy was done using Apply Model for known dataset with an achieved accuracy of 99.86% and Cross Validation for unknown dataset with an accuracy of 99.81%.

Acknowledgements

This work was supported by the Fundamental Research Grant Scheme (FRGS) under the Ministry of Education (MOE) with grant number 600-IRMI/FRGS 5/3 (234/2019).

References

[1] M. Soković, J. Glamočlija, A. Ćirić, J. Petrović, and D. Stojković, “Chapter 5 - Mushrooms as Sources of Therapeutic Foods,” in *Therapeutic Foods*, A. M. Holban and A. M. Grumezescu, Eds. Academic Press, 2018, pp. 141–178.

[2] M. Kuo, “Lepiotoid Mushrooms,” *MushroomExpert.Com*. https://www.mushroomexpert.com/lepiotoid.html.

[3] P. Maurya and N. P. Singh, “Mushroom Classification Using Feature-Based Machine Learning Approach,” in *Proceedings of 3rd International Conference on Computer Vision and Image Processing*, 2020, pp. 197–206.

[4] M. E. Valverde, T. Hernández-Pérez, and O. Paredes-López, “Edible mushrooms: improving human health and promoting quality life.,” *Int. J. Microbiol.*, vol. 2015, p. 376387, 2015, doi: 10.1155/2015/376387.

[5] S. Ismail, A. R. Zainal, and A. Mustapha, “Behavioural features for mushroom classification,” in *2018 IEEE Symposium on Computer Applications Industrial Electronics (ISCAIE)*, 2018, pp. 412–415, doi: 10.1109/ISCAIE.2018.8405508.

[6] M. G. Eyad Sameh Alkronz, Khaled A. Moghayer, Mohamad Meimeh, “Classification of Mushroom Using Artificial Neural Network .,” *Int. J. Acad. Appl. Res. (IJAAR).*, vol. 3, no. 2, pp. 1–5, 2019, [Online]. Available: http://www.ijeais.org/ijaar.

[7] C. Dua, Dheeru and Graff, “{UCI} Machine Learning Repository,” *University of California, Irvine, School of Information and Computer Sciences*, 2017. http://archive.ics.uci.edu/m.

[8] “Classification,” in *Data Mining Concepts*, Oracle, 2021, p. 14.

[9] L. Smolskaitė, P. R. Venskutonis, and T. Talou, “Comprehensive evaluation of antioxidant and antimicrobial properties of different mushroom species,” *LWT - Food Sci. Technol.*, vol. 60, no. 1, pp. 462–471, 2015, doi: https://doi.org/10.1016/j.lwt.2014.08.007.

[10] K. M. Schenk-Jaeger, C. Rauber-Lüthy, M. Bodmer, H. Kupferschmidt, G. A. Kullak-Ublick, and A. Ceschi, “Mushroom poisoning: a study on circumstances of exposure and patterns of toxicity.,” *Eur. J. Intern. Med.*, vol. 23, no. 4, pp. e85-91, Jun. 2012, doi: 10.1016/j.ejim.2012.03.014.

[11] J. N. Reed, S. J. Miles, J. Butler, M. Baldwin, and R. Noble, “AE—Automation and Emerging Technologies: Automatic Mushroom Harvester Development,” *J. Agric. Eng. Res.*, vol. 78, no. 1, pp. 15–23, 2001, doi: https://doi.org/10.1006/jaer.2000.0629.

[12] A. Gürgen, S. Z1le1z, and Ü. C. Y1ld1z, “Determination of mushroom consumption preferences by using fuzzy analytic hierarchy process,” *J. For. Sci.*, vol. 6, pp. 25–34, 2018.

[13] A. Al-Momany and G. Saleh, “A comprehensive study on Agaricus species of North Cyprus.,” *World J. Agric. Sci.*, vol. 5, pp. 195–200, 2009.

[14] C. Eusebi, C. Gliga, D. John, and A. Maisonave, “Data Mining on a Mushroom Database,” 2008.

[15] “RapidMiner Studio,” 2021. https://rapidminer.com/products/studio/feature-list/.

[16] P. Vadapalli, “Decision Trees in Machine Learning: Functions, Classification, Pros & Cons,” 2021. https://www.upgrad.com/blog/decision-trees-in-machine-learning/.