Machine Learning Project

Eder Tarifa Fernandez

1 Introduction

This project will involve building several supervised learning models using different algorithms and selecting the one that yields the best results, in addition to analyzing all the models we have created. For this, we will use a dataset to train and test our models.

2 Problem Description

The selected dataset includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms from the Agaricus and Lepiota families. Each species is classified as either definitely edible, definitely poisonous, or of unknown edibility and not recommended. The latter class has been combined with the poisonous one. Details of how the dataset is structured and the changes we have made are explained in the file "supervised_learning_notebook.ypynb." This information is crucial as it helps us understand our data. Additionally, an encoder has been applied to the data to allow us to build our models. Instances containing missing values have also been removed.

3 Metodology

Firstly, we created a model using K-NN, for which we applied cross-validation to search for the best value of k between one and seventy. We obtained the best result using a wide range of k values, as shown in Figure 1. Additionally, we used the Euclidean distance to measure the distance between the different instances and applied weights to these distances, increasing the value for neighbors that are closer.

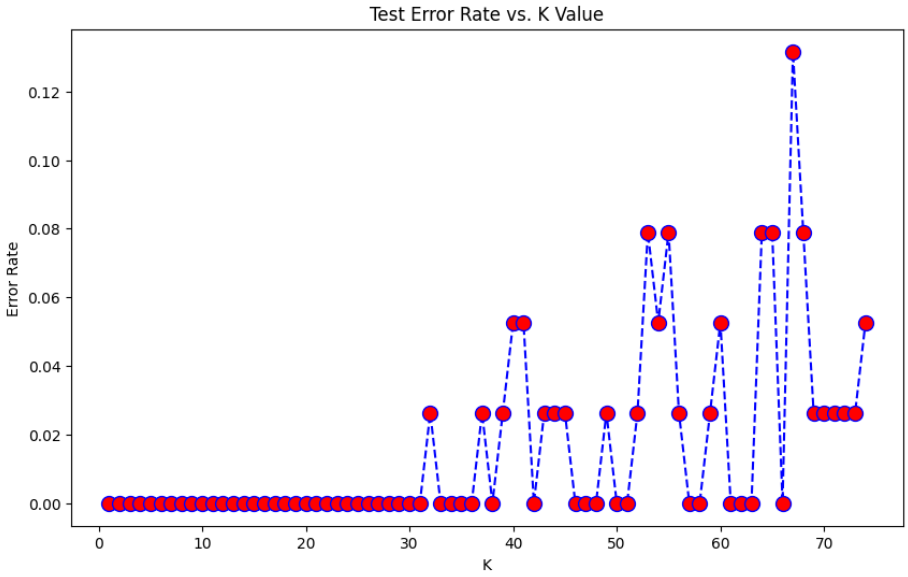

Figure 1: Error rate vs k for K-NN

Next, we built a logistic regression model using the following possible parameters: 0.001, 0.01, 0.1, 1, 10, and 100. In the results, we can see that the parameter with which we were able to build the best model was 100. Therefore, we know that our model has little regularization, meaning it allows the model coefficients to be more free to fit the training data without significant penalty for model complexity.

The next model we built was the SVM. Similar to the previous one, we selected the parameters 0.001, 0.01, 0.1, 1, 10, and 100 for both C and γ. Using cross-validation to obtain the optimal parameters, we achieved our best model with the value of 1 for both parameters. With C=1, we know that the model accepts a certain amount of error in the training set with the aim of obtaining a smoother decision function. With $\gamma = 1$, we know that the hyperplane will fit more strictly to the nearby points.

Finally, we created a model using decision trees. For this, we again performed cross-validation to obtain the best parameters, considering the following tree characteristics that produced the best results:

- Minimum samples per leaf: 5, 10
- Minimum samples per split: 5, 10 and 20
- Maximum depth: 10, 100
- Criterion: entropy and Gini

- ccp α: 0

This allowed us to achieve our best model, which divided the data based on the following Figure 2:

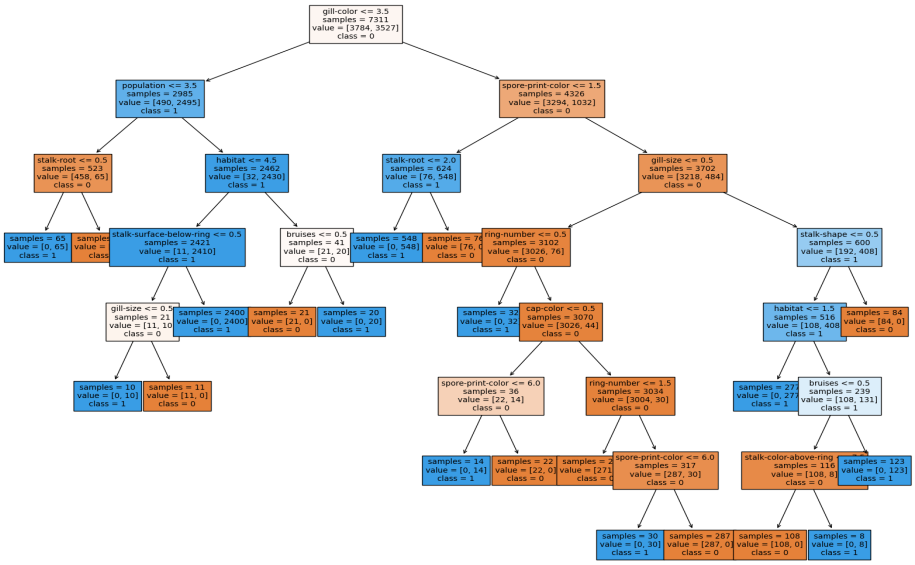

Figure 2: Visual result of decision tree

The primary distinction made by the model was based on the gill color of the mushrooms, dividing the data between those with a value above 3.5, that is, mushrooms with gill colors of gray, purple, buff, and green on one side, and all other colors on the other.

On one branch of the tree, a quick decision was made for mushrooms with these four colors, including those with abundant appearance and bulbous stem shape, among other features. The bulbous factor was particularly significant, as mushrooms with this trait were classified as poisonous. However, the most decisive factor on this main branch was the "stem-surface-below-ring" feature. More than 2,000 poisonous instances were identified when this surface was scaly, silky, or smooth. Additional factors such as gill size and bruises were also considered, though they were less influential.

On the other branch of the tree, the most important differentiator was spore print color. Among mushrooms with black, chocolate, or buff spore prints, those with a "club" or "bulbous" stem-root were identified as poisonous, while those with other stem-root types were considered edible. For mushrooms with other spore print colors and a broad gill size, the presence of at least one ring tended

to indicate edibility (depending on the spore print color), whereas those with no ring were categorized as poisonous. On the other hand, mushrooms with a narrow gill size presented numerous possible cases, with a total of 600 samples in our model. A quick observation, without needing to account for multiple variables (though it could be clarified visually), is that mushrooms with an "adjusted" stem shape are considered edible.

4 Results

Below, we present the results obtained from the different models we created:

Modelo	**Accuracy**	**Precision**	**Recall**	**F1**
K-NN	1	1	1	1
Regresión logística	0.959	0.967	0.947	0.957
SVM	1	1	1	1
Árboles de decisión	1	1	1	1

Table 1: Results obtained from different models

5 Discussion

We observe that the models created with K-NN, SVM, and decision trees achieved perfect results in the tests we conducted using 20% of the data. On the other hand, the logistic regression model showed the lowest performance. While these results could be considered good in another context, it is easy to rule out logistic regression as the best model given that we have three other models with superior performance.

Furthermore, we have a situation where three models obtained perfect scores in terms of accuracy, precision, recall, and F1. All three models performed excellently, demonstrating flawless predictions on the data. However, the model from which we could extract the most information—and the one that is most interpretable—is the decision tree model, as it provides insight into how decisions are made based on the different features of the mushrooms.

This information is extremely valuable because, for instance, if we needed to present the model, we could cross-check its decision-making process with a group of mushroom experts. This would allow us to demonstrate that our model follows clear criteria for determining whether a mushroom is poisonous or edible.

6 Conclusion

In conclusion, we chose the decision tree model. Although we have four highly accurate models for making predictions, this model provides interpretability, which can be essential in certain cases to understand how the model arrived at a particular conclusion.

7 References

Database: http://archive.ics.uci.edu/dataset/73/mushroom