01

Warsaw School of Economics

Data Mining

using decison tree and random forest in R

Prasenjeet Rathore - 110092

→

02

Mushroom Dataset

This data set includes descriptions of samples corresponding to 23 species of gilled mushrooms

Each species is identified as definitely edible or definitely poisonous.
(1 response variable chosen = "class")

There are in total 22 attributes of a mushroom. (22 explanatory variables)

Source: https://archive.ics.uci.edu/ml/datasets/mushroom

03

Objectives

01. **To Explore Data**

aim of this step is to investigate the characteristics of the dataset in this step

02. **Apply Machine Learning models**

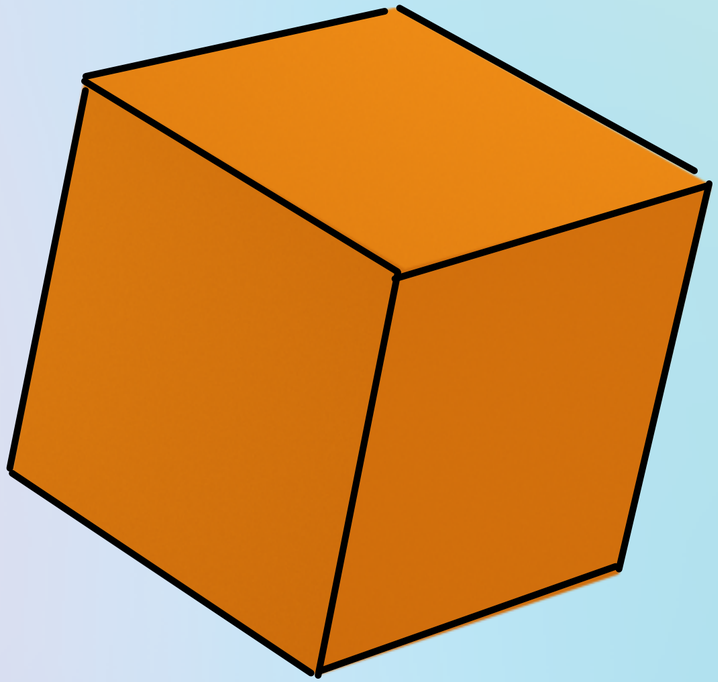
to discover patterns in the data and predict whether a mushroom is edible or poisonous

03. **Compare Models**

Compare the accuracy of the ML models used in determining the whether a mushroom is edible or poisonous

04

Packages Used



01. **randomForest**

Breiman and Cutler's Random Forests for Classification and Regression

02. **rpart**

It builds classification or regression models of a very general structure using a two stage procedure; the resulting models can be represented as binary trees.

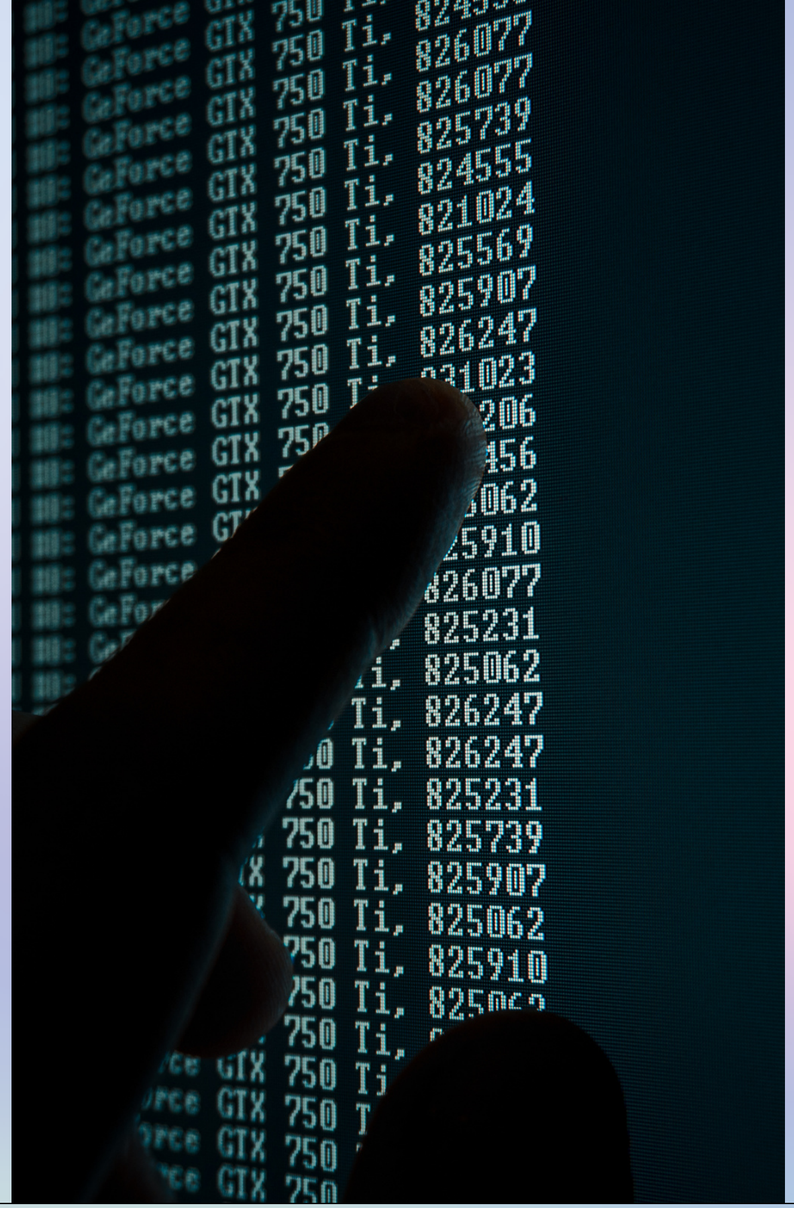
03. **tidyverse**

The tidyverse is an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures.

Data Cleaning

- All the variables in the dataset have class as "chr" that is character class and in order to analyze the dataset it must be converted into factor

- loading dplyr package which is a part of tidyverse package pipe function "%>% and "as.factor()" was used to change class of variables from "chr" to "factor"



Data Insight

- Number of observations in the dataset : 8124
- Balance of response variable "class"

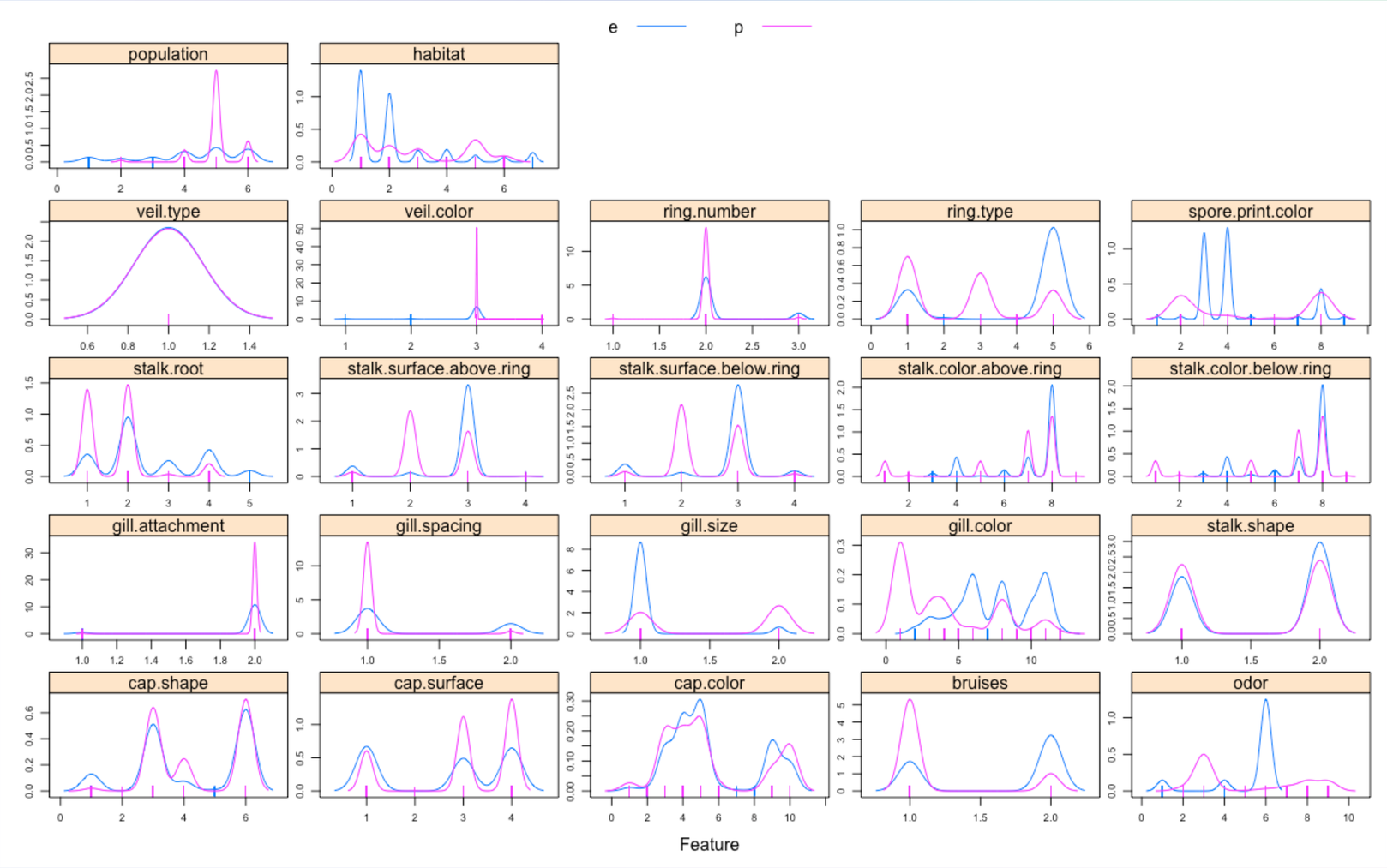
e	p
4208	3916

e = edible, p = poisonous

the response variable is fairly balanced so no adjusting is needed in order to handle bias

- There are no missing values in the dataset

07

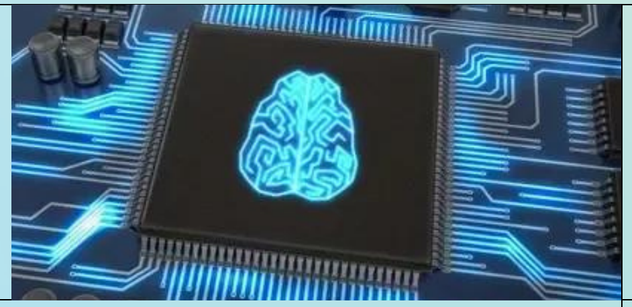


featurePlot was used to analyze the density of values for "**class**" response variable with values "**e**" & "**p**" in relation to all explanatory variables. There were no major distinct values found although for **habitat**, **ring.type**, **spore.print.color**, **population** there was decent separation and for **viel.type** and **stalk.shape** both had very similar values.

Sampling and Splitting Data

randomizaiton

using r's random number generator function "set.seed()" random numbers are generated and using function "sample()" the random number generated are used to randomise data



Training Data

70% of the randomised data is split and allocated to the training data which will be used to train machine learning models

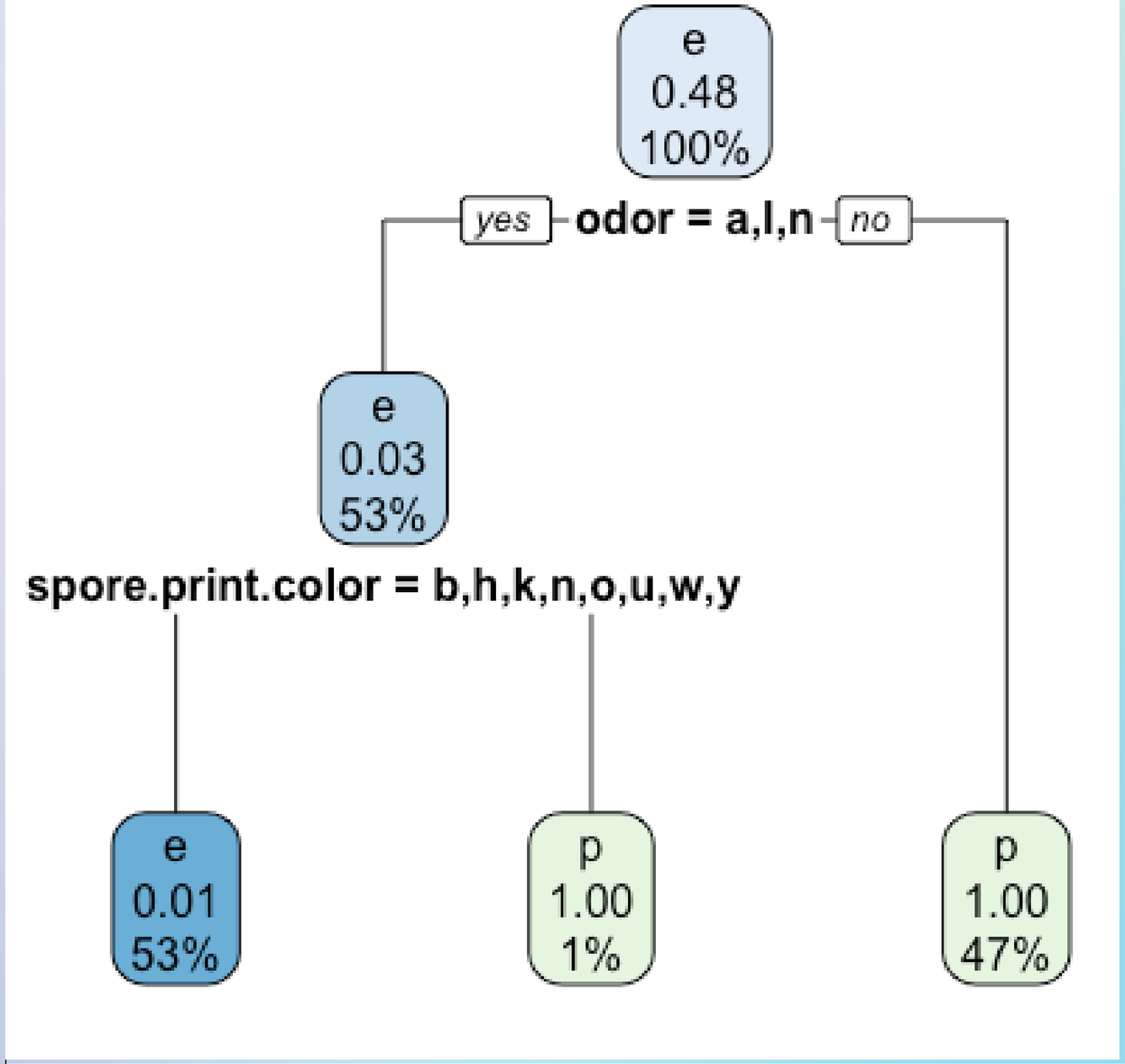
Test Data

Remaining 30% of the data will be allocated to the test data which will be used to measure the accuracy of the ML models trained using the training data.

Model I
Decision Tree

rpart function was used to create a model from the dataset using decision tree

n = 5686
*n = number of observations used in the model



The decision tree generated shown in the figure used odor as the most important predictor to predict and classify mushroom if they were edible "e" or poisonous "p".

The model predicted that almost 47% of all mushrooms are poisonous if they don't have the following odor "a","l", "n" (which are almond, anise and none respectively)

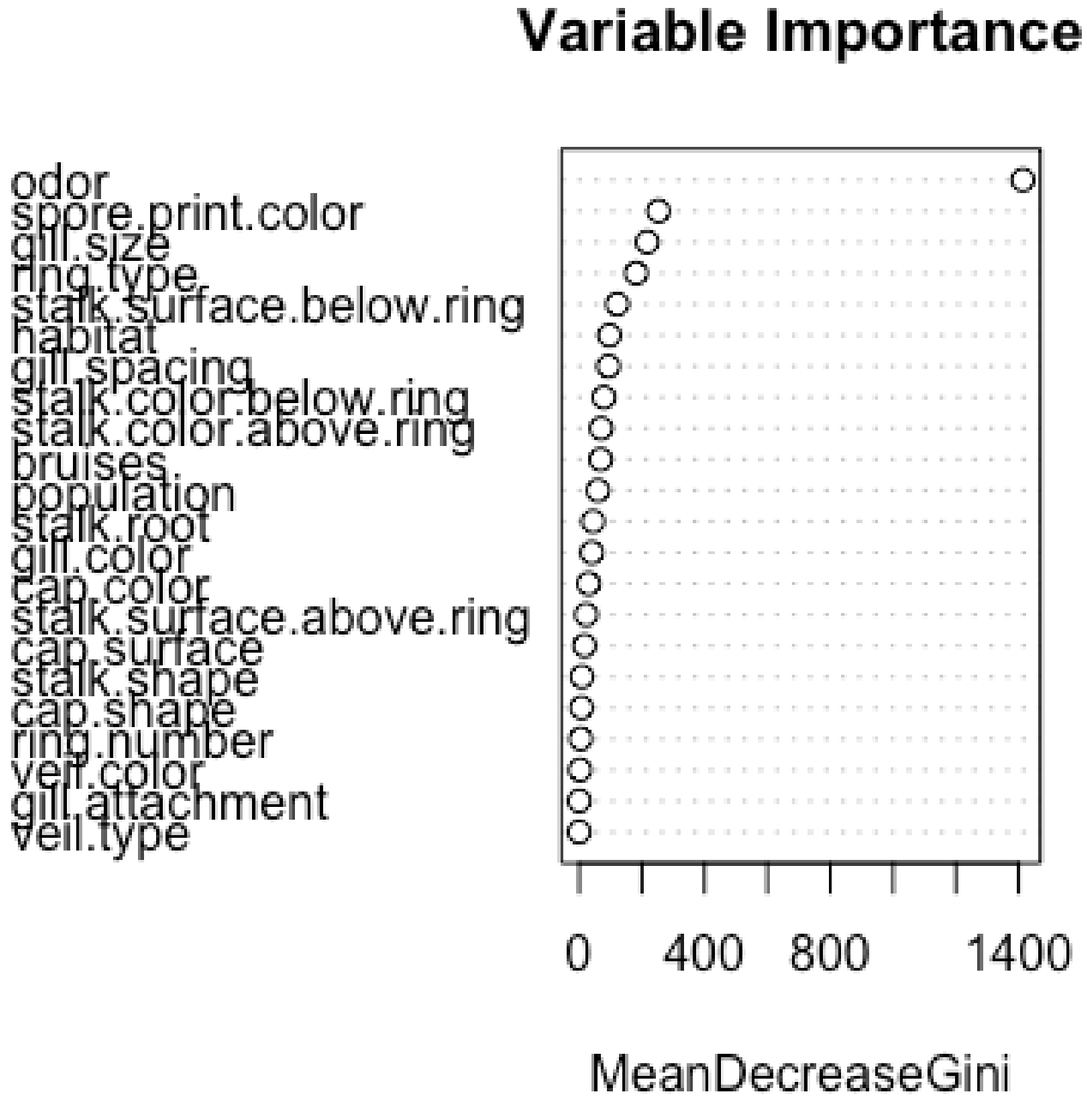
Remaining 53% of the mushrooms were further split on the basis of spore.print.color.

II

Model II
Random Forest

" Random Forests grows many classification trees. To classify a new object from an input vector, put the input vector down each of the trees in the forest. Each tree gives a classification, and we say the tree "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest)."

source: https://www.stat.berkeley.edu/~breiman/RandomForests



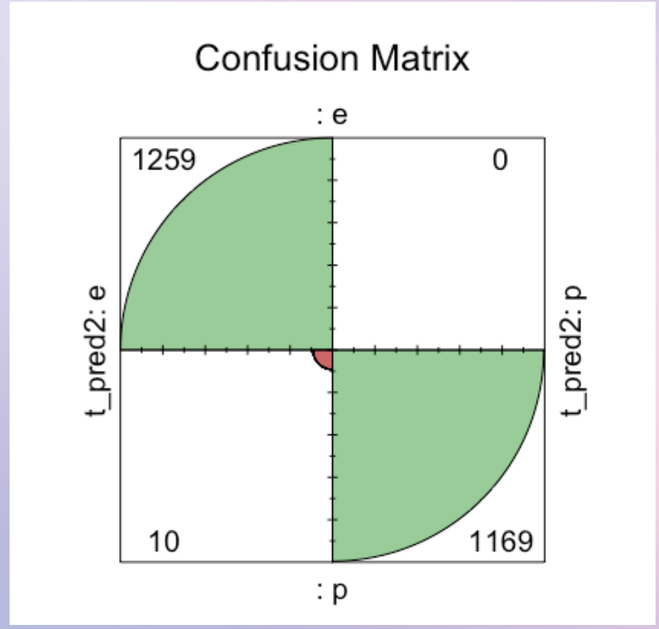
Random forest model predicted that the odor is the most importance variable in detecting whether a mushroom is edible "e" or poisonous "p". Very similar conclusion to the decision tree model.

spore.print.color , gill.size and ring.type of a mushroom were the next most important predictors after odor according to the model.

There were 20 trees generated during training in order to perform the classification.

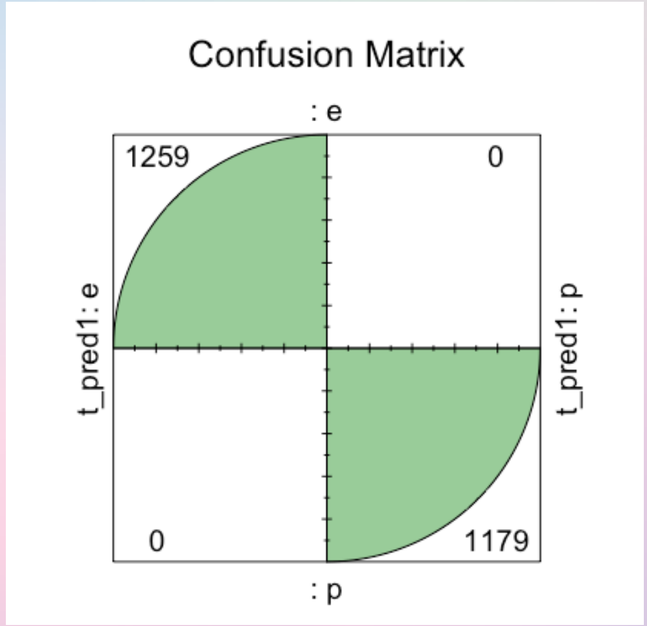
Comparing both models

Decision Tree



Decision tree correctly predicted 1259 mushrooms as edible and 1169 as poisonous. It had an accuracy of 99.58%. Although it predicted 10 mushrooms as edible which were actually poisonous.

Random Forest



Random forest correctly predicted 1259 mushrooms as edible and 1179 as poisonous. It had an accuracy of 100%.

⟶

Conclusion

Both the models did a very good job at predicting whether a mushrooms was edible or poisonous , however decision tree had a type II error which is fairly dangerous in comparison to the random forest model. For example in population of 30 million if decision tree was used to select mushrooms that would poison about 123051 people.

Model preffered

Random forest due to better accuracy and no type II error.

Suggestion

Decision Tree

Type II error can be reduced in the model if a larger dataset is provided to train the model.

Thank You