**INT247**

NAME: Mahadasu Gowtham Kishore

Reg. No.: 11801832

Roll no.: B44

Mushroom Classification (Poisonous)

**INTRODUCTION:**

This Project is on classification of poisonous mushrooms to find the accuracy and trained using SVM and Logistic Regression Model.

**Dataset description:**

This dataset contains 8124 entries corresponding to 23 species of gilled mushrooms from North America. Each species is identified as definitely edible (e), definitely poisonous (p), or of unknown edibility and not recommended (also p). Each entry has 22 features related to the physical characteristics of the mushroom.

**Modules, Features and Analysis:**

### Converting categorical data to numerical

Most Machine Learning algorithms require numerical features. However, our dataset is composed of categorical features.

#### **Label Encoding**

A typical approach is to perform Label Encoding. This is nothing more than just assigning a number to each category, that is:

(cat\_a, cat\_b, cat\_c, etc.) → (0, 1, 2, etc.)

This technique works:

When the features are binary (only have 2 unique values)

When the features are ordinal categorical (that is, when the categories can be ranked). A good example would be a feature called t-shirt size with 3 unique values small, medium or large, which have an intrinsic order.

**However**, in our case, only some of our features have 2 unique values (most of them have more), and none of them are ordinal categorical (in fact they they are nominal categorical, which means they have no intrinsic order).

#### **One Hot Encoding**

For the remaining features, a technique called One Hot Encoding is used.

Essentially, this consists on creating a new binary feature representing each category. For instance, from the feature cap surface, which has 4 unique values (f, g, y and s), creating 4 binary features (cap\_surface\_f, cap\_surface\_g, cap\_surface\_y and cap\_surface\_s) indicating whether the category they represent was indeed that one or not.

### Separating labels from features

X will now contain our features, and y our labels (0 for edible and 1 for poisonous/unknown)

### Standardising features

It is generally considered a good practice to standardise the features (converting them to have zero-mean and unit variance). Most of the times, the difference will be small.

### Creating training and sets sets

The data is separated into a training set (70%) and a test set (30%). This is a very standard approach in Machine Learning.

The stratify option ensures that the ratio of edible to poisonous mushrooms in the dataset remains the same in both training and test sets. The random\_state parameter is simply a seed for the algorithm to use.

**Logistic Regression**

A logistic regression simply predicts the probability of an instance (row) belonging to the default class, which can then be snapped into a 0 or 1 classification.

Logistic Regression is used again using cross-validation, to ensure that we are not overfitting the data.

**Support Vector Machine**

**SVM** is a supervised **machine learning** algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems.

**This dataset ensures:**

All mushrooms with an almond or anise odor are edible

All mushrooms with a creosote, fishy, musty, pungent or spicy odor are poisonous (or unknown edibility)

Most mushrooms with no odor are edible.

Finally,

Accuracy of Logistic Regression classifier on the test set: 1.00

Accuracy of Logistic Regression classifier using 10-fold cross-validation: 0.9997655334114889

Accuracy of Support Vector Machine: 0.9995898277276456

## **Conclusion**

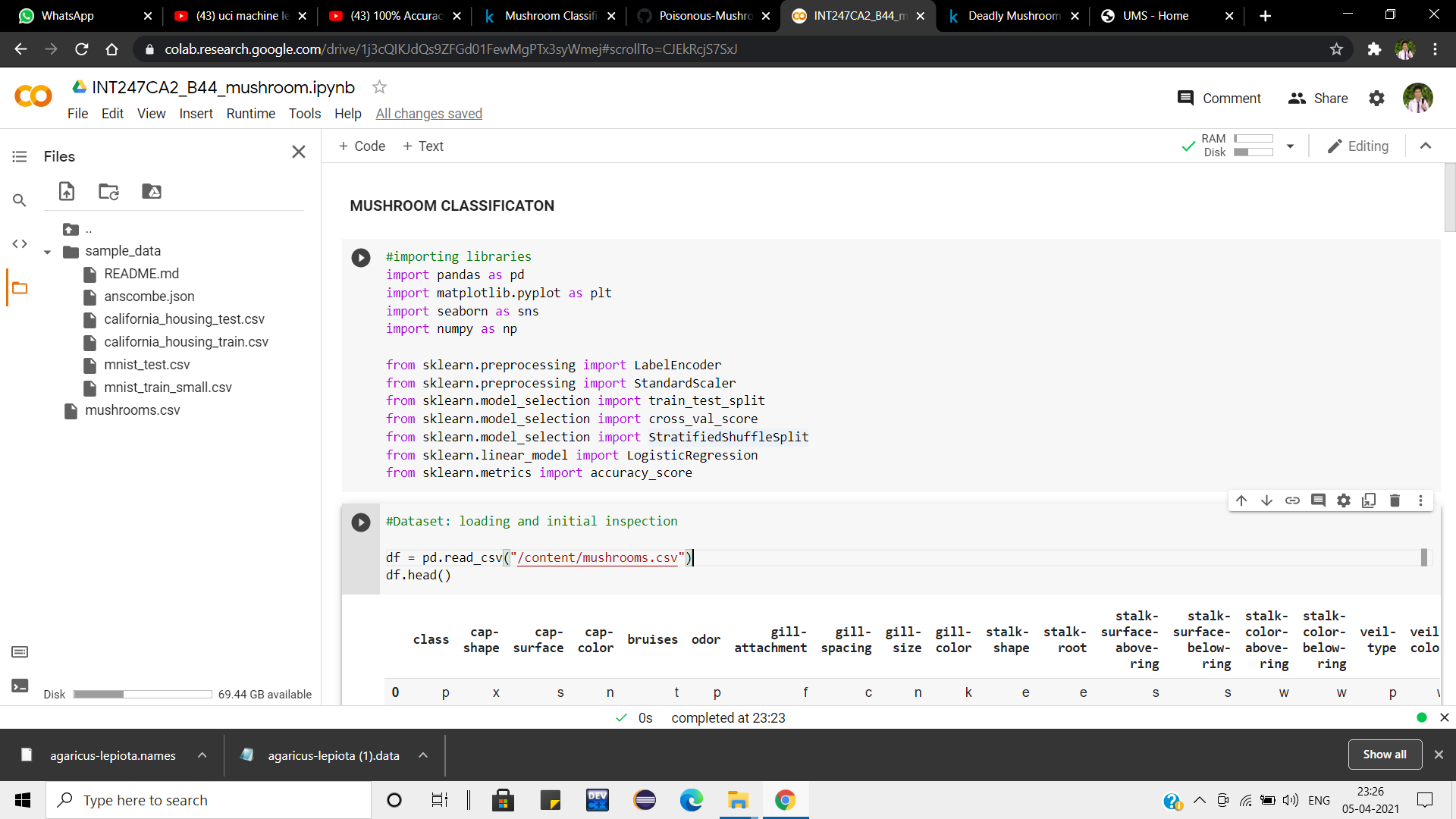
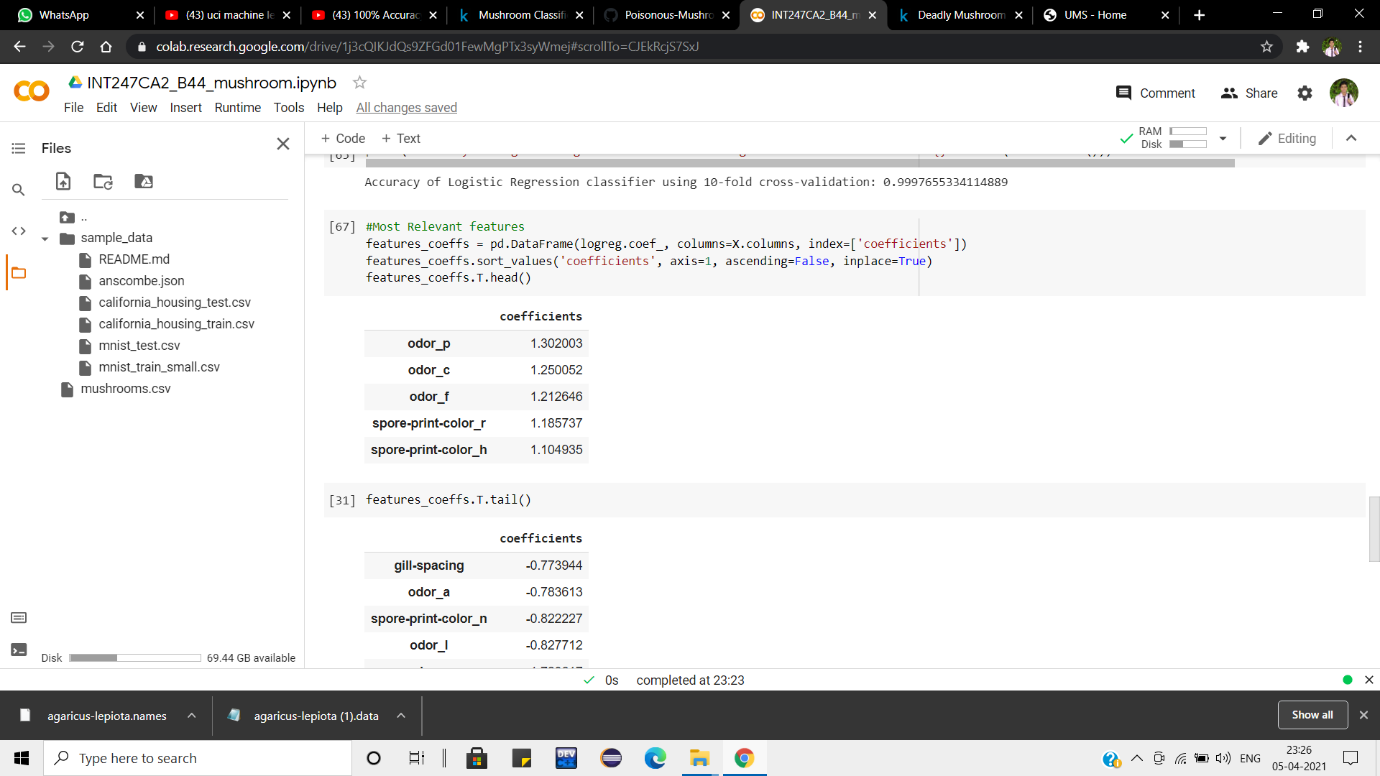
Fitted a logistic regression model and achieved near perfect accuracy, so there was no need to try with more complex models.

This algorithm identified specific traits (particularly regarding odor) that seem to heavily influence the chance that a mushroom is edible or not.

Even though experts have determined that is that there is no simple set of rules to determine whether a mushroom is edible or not, it seems like with this algorithm we can get pretty close.

**Google Colab link:** <https://colab.research.google.com/drive/1j3cQIKJdQs9ZFGd01FewMgPTx3syWmej?usp=sharing>

**Screenshots:**

****