DS5097

HW6 – Data Preprocessing and Intro to ML

Introduction

Python's SK-Learn Library for Machine Learning does not work well with data that is not numerical. Your job is to convert the Mushroom data set from a 'char' dataset into an integer dataset using Label Encoding.

Label Encoding is a technique that is used to convert categorical columns into numerical ones so that they can be fitted by machine learning models which only take numerical data. It assigns a unique integer to each category in the data.

Once the data is converted, you then need to run three different ML algorithms; SVM, GNB, and Decision Trees, on the dataset to try to predict if a mushroom is poisonous or edible.

Step 1 (20 points) Understanding the data

The first part of any data science project is to understand what the data that you are working with is. Read through the documentation to understand the type of data that you are working with.

Answer the following questions (**5 points each**):

1) What are the columns representing?

2) Which column is your "target" Column. The column that you are going to try to predict.

3) List out what each unique 'char' represents.

4) Create a mapping of 'chars' to their numerical representation.

Step 2 (20 points) Pre-Processing

Convert the mushroom dataset into a numerical dataset using label encoding and save it as a separate dataset.

Step 3 (40 points) Machine Learning

Run three different ML algorithms (SVM, GNB, and Decision Trees) on the dataset and compare the results of the three algorithms. Report on the overall accuracy using a Confusion Matrix for each algorithm.

Step 4 (20 points) Report

Write a report as either a .pdf or markdown file outlining the results of your algorithms. Explain your results as if you are talking to a non-technical person. How do you interpret the results? Which algorithm did better? Why do you think so? Include any graphs or screenshots in your report. Your report should also include your answers from part 1.

What to turn in

Push your dataset and report to GitHub and submit a link to your GitHub submission in canvas.