MUSHROOM CLASSIFICATION

Detailed Project Report

Oviyashri B
Data Science Intern at Ineuron.ai

INTRODUCTION

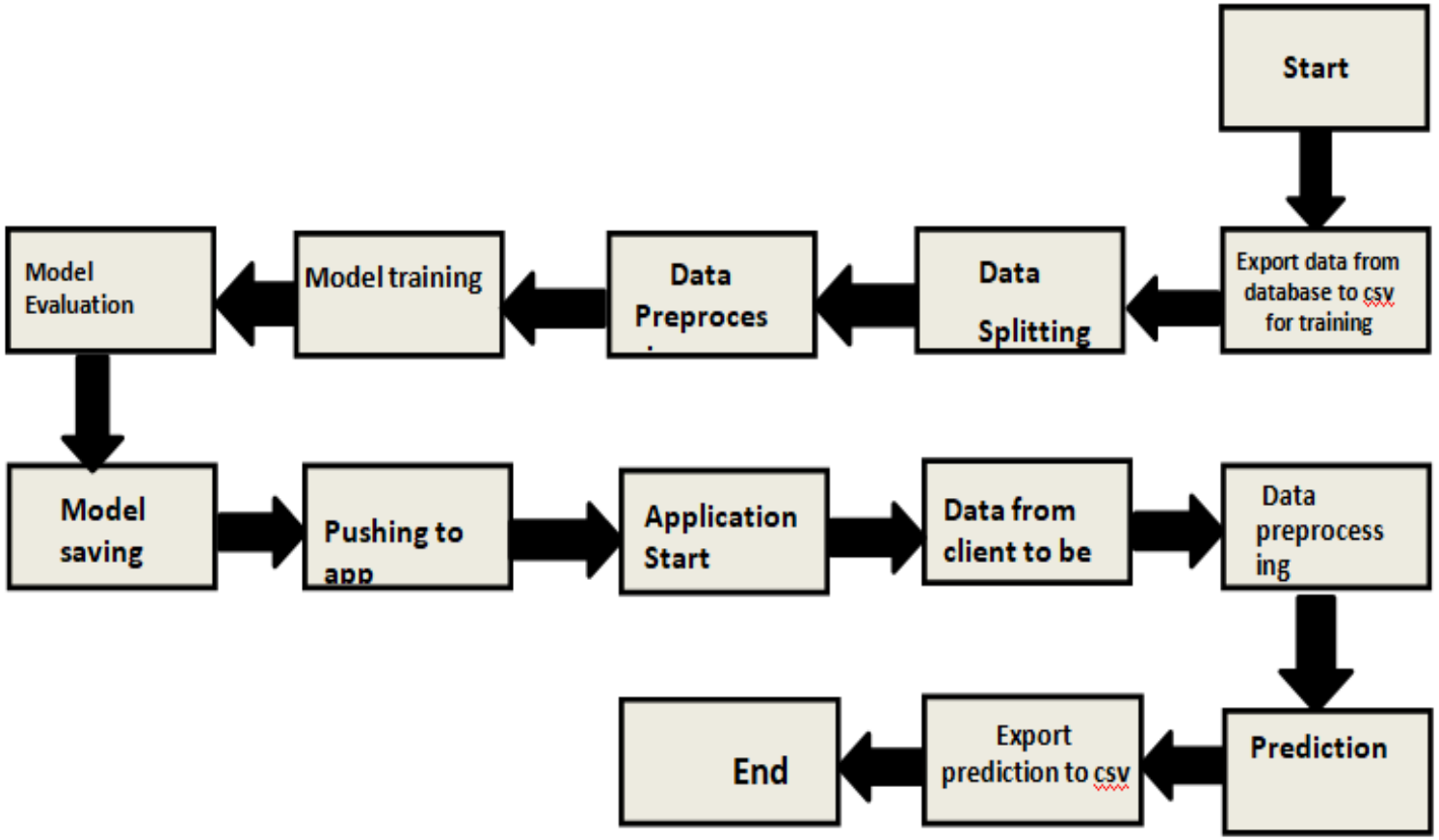
Mushrooms can be found extensively in a variety of natural environments and visual identification of mushroom species is well established. Some mushrooms are known because of their nutritional and therapeutic properties. Some species are known all over the world because of their toxicity that causes fatal accidents every year mainly due to misidentification. Some of the edible mushrooms are *Ganodeíma spp, Canthaíellus spp, Agaricusspp, Pleuíotus spp, Russula spp, Auricularia spp and ľermitomyces spp;* but the ornamentals are the beautifullyringed*Micropoíous spp. Amanita spp, Lepiota cristata, Lepiota brunneoincarnata and Inocybe asterospoía, Coprinusspp* are among the most important species responsible for mushroom poisoning. Morphological and chemical analyses for mushrooms are occasionally required in forensic science practice. In this work, the characteristics of the representative toxic mushrooms and some chemical methods for their toxins are presented. Mushrooms are identified traditionally by their appearance, taste, colour, odour, presence of scales etc.

The Audubon Society Field Guide to North American Mushrooms contains descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family Mushroom (1981). Each species is labelled as either definitely edible, definitely poisonous, or may be edible but not recommended. This last category was merged with the toxic category. The Guide asserts unequivocally that there is no simple rule for judging a mushroom's edibility, such as "leaflets three, leave it be" for Poisonous Oak and Ivy.

OBJECTIVE

The main goal is to predict which mushroom is poisonous & which is edible. The classical machine learning tasks like Data Exploration, Data Cleaning, Feature Engineering, Model Building and Model Testing.

ARCHITECTURE

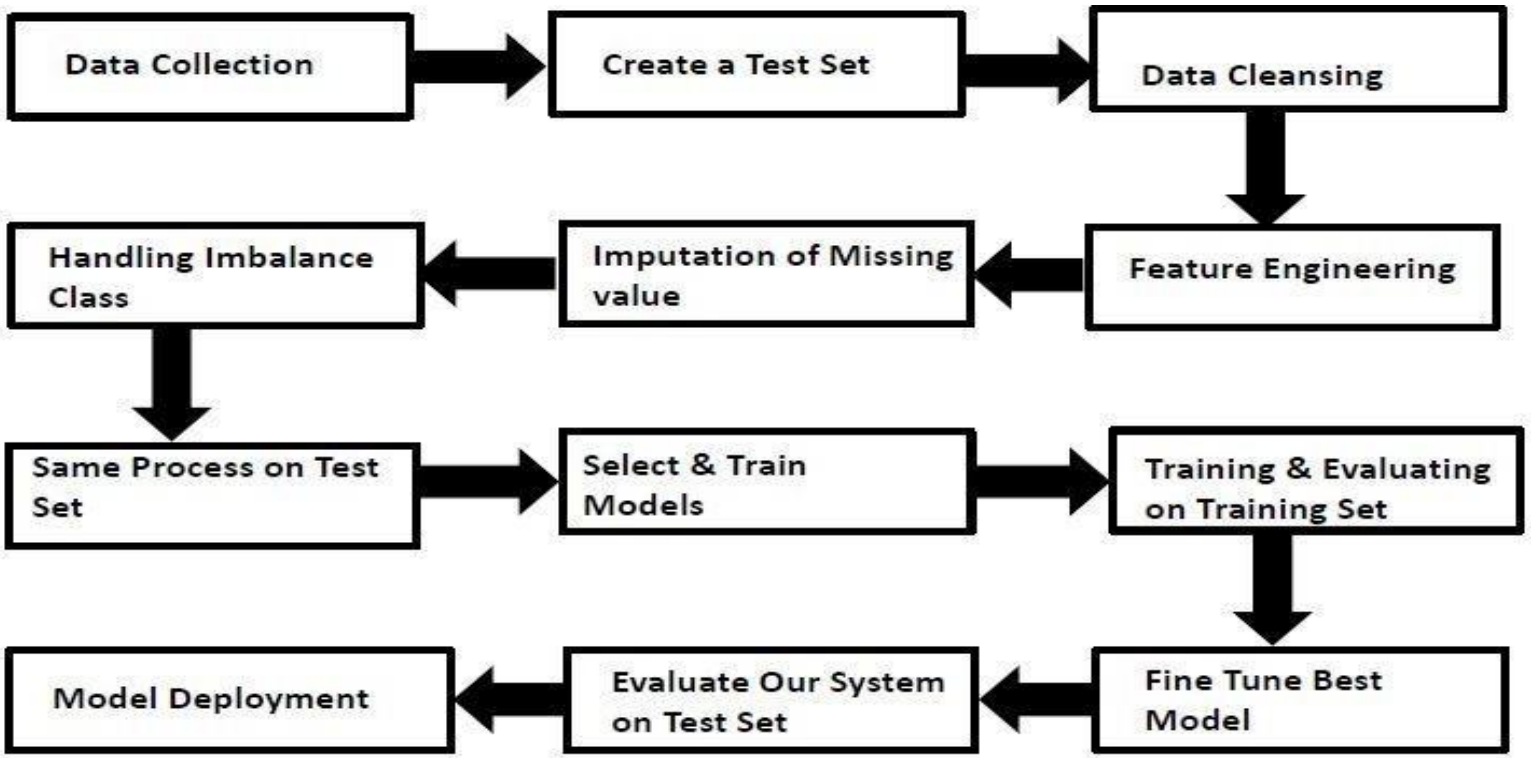


DATASET

Attribute Information: (classes: edible=e, poisonous=p)

- cap-shape: bell=b,conical=c,convex=x,flat=f, knobbed=k,sunken=s
- cap-surface: fibrous=f,grooves=g,scaly=y,smooth=s
- cap-color: brown=n,buff=b,cinnamon=c,gray=g,green=r,pink=p,purple=u,red=e,white=w,yellow=y
- bruises: bruises=t,no=f
- odor: almond=a,anise=l,creosote=c,fishy=y,foul=f,musty=m,none=n,pungent=p,spicy=s
- gill-attachment: attached=a,descending=d,free=f,notched=n
- gill-spacing: close=c,crowded=w,distant=d
- gill-size: broad=b,narrow=n
- gill-color: black=k,brown=n,buff=b,chocolate=h,gray=g, green=r,orange=o,pink=p,purple=u,red=e,white=w,yellow=y
- stalk-shape: enlarging=e,tapering=t
- stalk-root: bulbous=b,club=c,cup=u,equal=e,rhizomorphs=z,rooted=r,missing=?
- stalk-surface-above-ring: fibrous=f,scaly=y,silky=k,smooth=s
- stalk-surface-below-ring: fibrous=f,scaly=y,silky=k,smooth=s
- stalk-color-above-ring: brown=n,buff=b,cinnamon=c,gray=g,orange=o,pink=p,red=e,white=w,yellow=y
- stalk-color-below-ring: brown=n,buff=b,cinnamon=c,gray=g,orange=o,pink=p,red=e,white=w,yellow=y
- veil-type: partial=p,universal=u
- veil-color: brown=n,orange=o,white=w,yellow=y
- ring-number: none=n,one=o,two=t
- ring-type: cobwebby=c,evanescent=e,flaring=f,large=l,none=n,pendant=p,sheathing=s,zone=z
- spore-print-color: black=k,brown=n,buff=b,chocolate=h,green=r,orange=o,purple=u,white=w,yellow=y
- population: abundant=a,clustered=c,numerous=n,scattered=s,several=v,solitary=y
- habitat: grasses=g,leaves=l,meadows=m,paths=p,urban=u,waste= w,woods=d

MODEL TRAINING AND EVALUATION WORKFLOW



MODEL TRAINING AND EVALUATION

Data Collection

- Mushroom Classification Data Set from Kaggle Repository
- For Data Set: https://www.kaggle.com/uciml/mushroom-classification

Data Pre-Processing

- Categorical features handling

MODEL TRAINING AND EVALUATION

Model Training and Evaluation

- Classification algorithm - SVC is tested since it given better result and was chosen for model training and testing.
- Model performance evaluated based on accuracy, classification report.

WORKFLOW

Data Description

We will be using Mushroom Prediction Data Set present in Kaggle Repository. This Data set is satisfying our data requirement. Total 8120 instances present in different batches of data.

Export Data from database to CSV for Training

Here we will be exporting all batches of data from database into one csv filefor training.

Data Splitting

We split the data here for our train and test data for further uses.

Data Preprocessing

We will be exploring our data set here and perform data preprocessing depending on the data set. We first explore our data set in Jupyter Notebook and decide what pre-processing and validation we have to convert all those to numerical values by label encoding and then we have to write separate modules according to our analysis, so that we can implement that for training as well as prediction data.

Model Training

We trained various model in our notebook and SVC was good on it. We trained with our processed data.

Model Saving

We will save our models so that we can use them for prediction purpose.

Push to app

Here we will do cloud setup for model deployment. We also create our streamlit app and user interface and integrate our model with streamlit app and UI.

Data from client side for prediction purpose

Now our application on cloud is ready for doing prediction. The prediction data which we receive from client side.

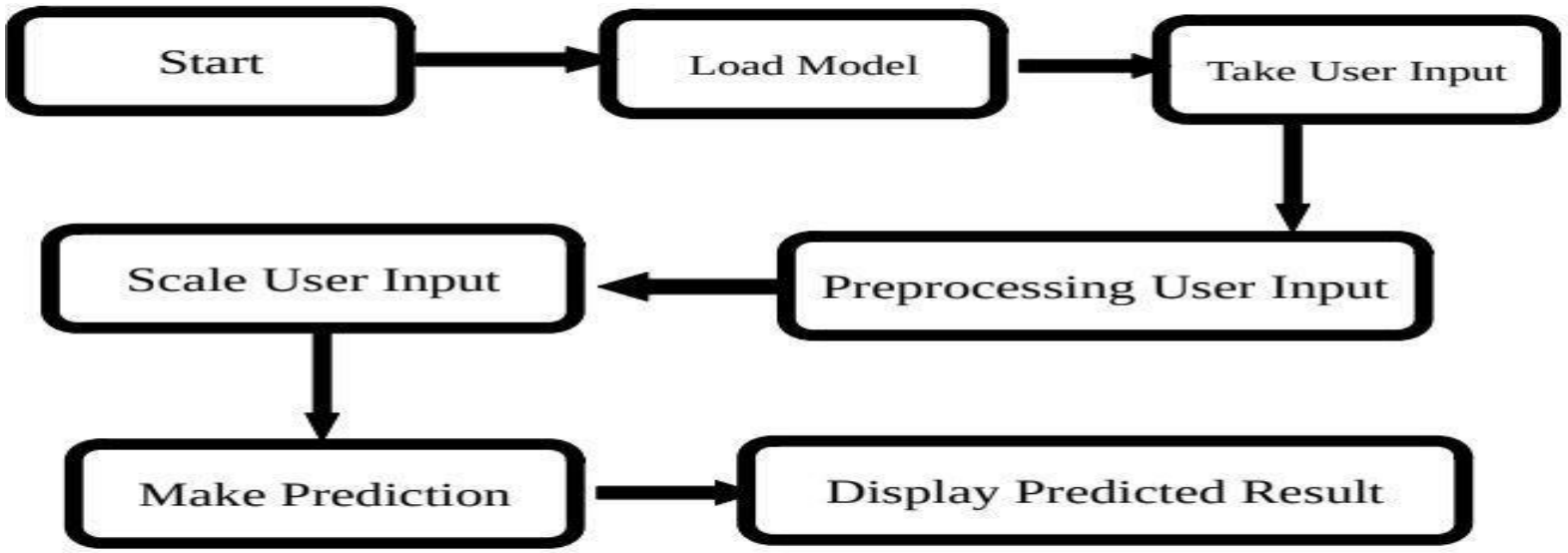
Data processing

Client data will also go along the same process Data pre-processing and according to that we will predict those data.

Export Prediction to CSV

Finally when we get all the prediction for client data, then our final task is toexport prediction to csv file and hand over it to client.

Model Deployment



Model Deployment

- The final model is deployed on Heroku using Streamlit framework.

FREQUENTLY ASKED QUESTIONS

Q1) What is the source of data?

Kaggle for Dataset

URL : https://www.kaggle.com/uciml/mushroom-classification

Q2) What was the type of data?

The data was the combination of numerical and Categorical values.

Q3) What's the complete flow you followed in this Project?

Refer slide 7th, 8th and 9th for better understanding.

Q4) After the File validation what you do with incompatible file or files which didn't pass the validation?

Files like these are moved to the Achieve Folder and a list of these files has been shared with the client and we removed the bad data folder.

Q5) How logs are managed?

We are using different logs as per the steps that we follow in training and prediction like model training log and prediction log etc. And then sub log are inside those folder.

Q 6) What techniques were you using for data pre-processing?

- Removing unwanted attributes
- Visualizing relation of independent variables with each other and output variables
- Checking and changing Distribution of continuous values
- Cleaning data and imputing if null values are present.
- Converting categorical data into numeric values.

Q 7) How training was done or what models were used?

- First Data validation done on raw data and then good data insertion happen in DB.
- Then Data preprocessing done on final CSV file received from DB.
- Various model such as Logistic, SVM models are trained on all andbased on performance, model is saved.

Q 8) How Prediction was done?

- The testing files are shared by the client .We Perform the same life cycle till the data is clustered .
- Then on the basis of cluster number model is loaded and perform prediction. In the end we get the accumulated data of predictions.

Q 9) What are the different stages of deployment?

- After model training and finalizing all models. We created required files for deployment.
- Finally deployed our model cloud platform Heroku.

Q 10) How is the User Interface present for this project?

- For this project I have made for one user input prediction.
- UI is very user friendly and easy to use.

THANK YOU