Mushroom Classification System

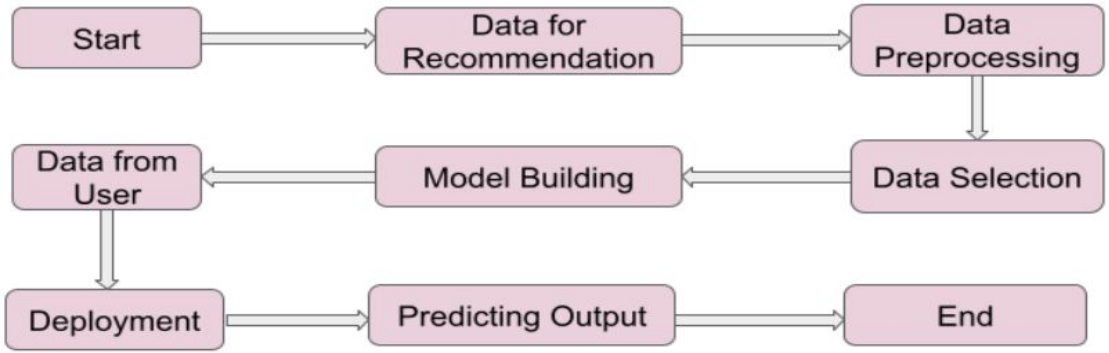
Objective:

The main objective of the project is to help people who are not familiar with the mushroom types by classifying the mushroom is edible or poisonous

Benefits:

- Identify mushroom edibility
- Preventing health issues
- Enriching nutrition intake
- Promoting good mushrooms

Architecture Diagram



Data Validation and Data Transformation :

- Data type of columns - The data type of columns is in the form of String .When user select a category of each column , it take as a String
- Null values in columns – If user forget to give value of any column , then default value will consider as a selected value
- Name of columns - The name of the columns are as per the columns name of the dataset

Model Training:

- Data from Kaggle:

  Data got from Kaggle as csv file used to train the model

- Data Preprocessing
  - Performing EDA to get insight of data like identifying imbalance data and distribution
  - Check for null values in the columns. If present impute the null values.
  - Encode the categorical values with numeric values.
  - Performing statistical hypothesis test to analyze correlation between categorical variable
- Model Selection

  After the data is encoded ,By using 4 algorithms 'SVM','RandomForestClassifier','XGBClassifier' and 'LogisticRegression'. We done GridSearch through all models to calculate the AUC scores for all models and select the model with the best score.

Prediction:

- Select the Test data
- We performed data pre-processing techniques on test data.
- Model is used to predict the Test data .
- Once the prediction is done . We measure the accuracy of the predictions using confusion matrix and accuracy matrix . Confusion matrix saved as a image file

Q & A:

Q1) What's the source of data?

We got data from Kaggle platform

Q 2) What was the type of data?

The data was Categorical values.

Q 3) What's the complete flow you followed in this Project?

Refer slide 4th for better Understanding

Q 4) What database used ?

We are not used any database, as we store the model as a binary file in the localstorage

Q 5) Where the application is hosted ?

The application source code is hosted in GitHub, Live application is running in Heroku

Q 6) Does user needs to register to access this application ?

No, it is open to all to access free of cost and free of registration

Q7) What techniques were you using for data pre-processing ?

- Analyze data for imbalance
- Perform statistical test to identify the relationship between target variable and features
- Check for null values . Impute if any
- Converting categorical data into numeric values(Encoding)

Q 8) How training was done or what models were used ?

- The categorical columns were encoded
- After training data is ready , Algorithms like SVM , XGBoost, Random Forest, Logistic Regression were used through GridSearch and selected best model base on score

Q 9) How Prediction was done ?

The data collected from user .We Perform the same life cycle till the data is Encoded . Then load the model and do prediction.

Q 10) What are the different stages of deployment?

- When the model is ready we deploy it in Fire environment .Where SIT and UAT is performed over it.
- Once We get Sign off from Fire we deploy in Earth and UAT is performed over it.
- After getting the sign off from Earth we deploy in production