MUSHROOM CLASSIFICATION

Objective

- The aim is to develop a application by using machine learning algorithm that will determine if a certain mushroom is edible or poison by its specifications like cap shape, cap color, gill shape, gill size etc.

Data Sharing Agreement

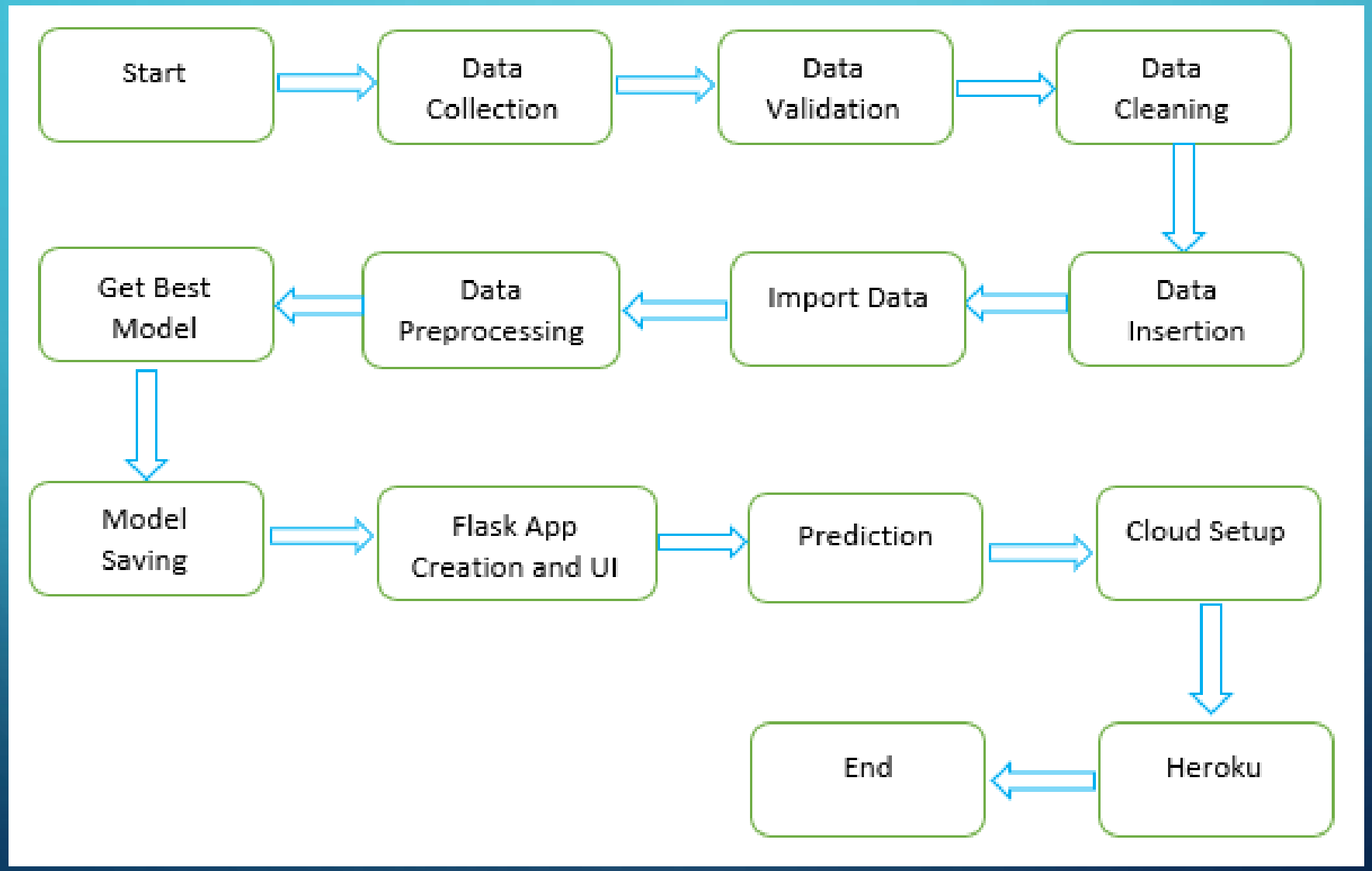
- File name(mushrooms.csv)
- Number of columns
- Column names
- Column details
- Column data types

Data Description:

- 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

Architecture



Data Validation

- File name validation: File name validation as per the Data Sharing Agreement.
- Name and number of columns: It will check for name and number of columns.
- Data types of columns: The data type of columns are categorical.
- Meaningless observation is converted into meaningful.

Exploratory Data Analysis

- Number of rows and columns.
- Information about the data.
- Describing the data.
- Graphical representation and interpretation.

Data Preprocessing and Model Training

- Removing unwanted attributes.
- If the null values in the dataset are small then values are removed or replaced.
- Visualizing relation of independent variables with each other and output variables.
- Feature scaling is done.
- Meaningless observations are converted into meaningful observation.
- Using label encoding method for converting categorical data into numeric values.
- Splitting the data into train and test.
- Applying the data to different classification machine learning models and hyper parameter tuning is done.
- Comparing the accuracy of different machine learning models.

Model Selection

- Compute confusion metrics for model evaluation.
- Compute AUC value for each model.
- Hyper parameter tuning has been done for every model.
- After testing several classification algorithms and comparing there performance, Random Forest is selected for model building with 100% accuracy.

Prediction

- The testing files are shared and perform the same validation operations, data transformation and data insertion on them.
- The accumulated data from database is exported in csv format for prediction.
- We perform data pre-processing techniques in it.

Question and Answers

Q1) Explain about project

Ans: This project will help the users get to know which type of mushroom is good for health and which is not without having deep knowledge about it. As a data scientist I am involved in every phase of the project. My responsibility is to collect the data, importing the data as csv file, Exploratory Data Analysis, data preprocessing, model training, prediction and model deployment in the cloud.

Q2) What is source and size of the data?

Ans: The data is taken from the Kaggle.com and the size is 374KB.

Q3) What was the type of the data and what is the output?

Ans: The data is categorical. Output column consists two categories edible and poisonous.

Q4) How logs are managed?

Ans: We are using different logs as per the steps that we follow in validation and modeling like File validation log , Data Insertion ,Model Training log , prediction log etc.

Q5) What techniques were you using for data pre-processing?

Ans: Following are the data pre-processing techniques used for the project.

- Removing unwanted attributes.
- Visualizing relation of independent variables with each other and output variables.
- Meaningless observations are converted into meaningful observation.
- Converting categorical data into numeric values.

Q6) What's the complete flow you followed in this Project?

Ans: Refer slide 6th for better Understanding

Q7) What are models were used for this project ,which model performs better and why?

Ans: For this project Decision Tree, Random Forest, Adaptive boost, Gradient boost and Extreme gradient boosting techniques are used. Random forest model is considered as best model.

Random forest model is used for the deployment because:

- It is not overfitting.
- It uses row wise and columns wise sampling therefore it is robust to both outliers and missing values.
- It uses Decision tree as base model.

Q8) What is confusion metrics?

Ans: A confusion metrics is the table which is used to measure the performance of the classification algorithm for supervised learning. Actual value and the predicted value are the two parameters used in confusion metrics.

Q9) Briefly explain about under fitting and overfitting.

Ans: If the machine learning model does not performs well on both training set and test set then it is under fitting problem. To overcome from this problem we use regularization techniques.

If the machine learning model performs well on training set and does not perform well on the test set then it is called as overfitting.

Reasons for overfitting

- Small dataset with more number of parameters.
- Model is complex.
- Variance is high and bias is low.

By using cross-validation we can avoid overfitting.

Q10) What are the different steps of deployment process?

- Create the pickle file of the model.
- Create index.html, app.py, procfile and requirement.txt files.
- Upload the files on GitHub with new repository.
- Import the files on Heroku and create new application.
- After the deployment of the branch, new web application page has been created.
- Fill the input options and click on predict button then result will be displayed.


Thank You