

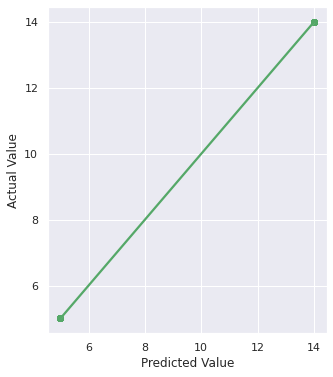
|  |
| --- |
| PROJECT DESCRIPTION |
|  |
| January 4  Presented By  Gabriel Sundalkar |

# Let us understand the objective and approach of our project

|  |
| --- |
| There are three parts of our project:  1. Exploratory data analysis (EDA)  * Reading the data, working on datatypes, null values * Visualization: Count plots, heatmap etc.  1. Machine learning models  * Logistic Regression * KNN * Decision Tree Classifier * Random Forest Classifier  1. Recommendation Model Using PySpark |
| *We have considered mushroom dataset for EDA, machine learning techniques and PySpark recommendation model.*  *Source of Dataset:* [*https://www.kaggle.com/*](https://www.kaggle.com/)  *Our dataset has 23 columns and 8124 rows.* |
| Below is the data: Data Storage Platform: We have considered mushroom dataset from Kaggle website.  Our dataset is structured and we can store it in any relational database such as SQL.  In the use case that we have considered, we have used python libraries such as PySpark, Sklearn, Seaborn etc. Data Transformation: Since, ours is the structured data that is capable of storing in any relational database(SQL), we have used dataframe to present and manipulate the data according to the use case.  We have calculated the total number of null values in each column. Our dataset does not have any null value in it. We have used multiple functions to print all basic parameters of each column of dataset.  **Visualization and data analytics:** We have used seaborn library to present various plots and graphs. We have trained various ML models along with a recommendation system. Let us have an overview PART 1: EXPLORATORY DATA ANALYSIS Below are the screenshots of countplots:    We have presented the only four screenshot of countplot above. We have created the count plot for each column this way.  Below is the heatmap:      **Below are the distplots:**  We have used distplot( ) function from seaborn library.      We have presented above the screenshots of 4 distplots only. We have created the dist plot for each column this way. PART 2: MACHINE LEARNING MODELS We have implemented 4 machine learning model in our project. Below are the screenshots of regression plot of each of the machine learning model:   1. Logistic Regression:   Accuracy of the model: 94.26%. We have also presented the confusion matrix. |

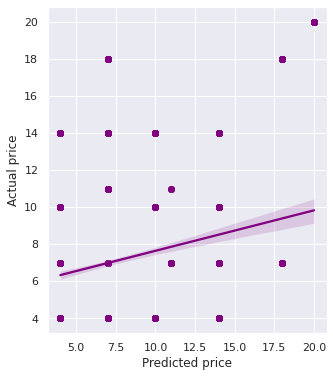
1. KNN:

Accuracy of the model is 100%



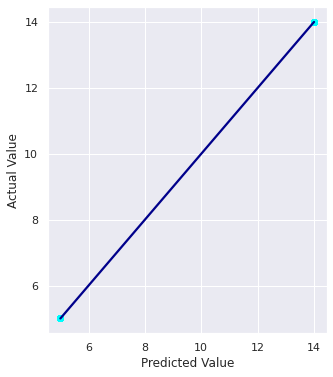
1. Decision Tree Classifier:

Accuracy of the model is 73.1%



1. Random Forest Classifier

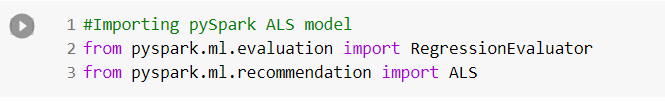
Accuracy of the model: 100%

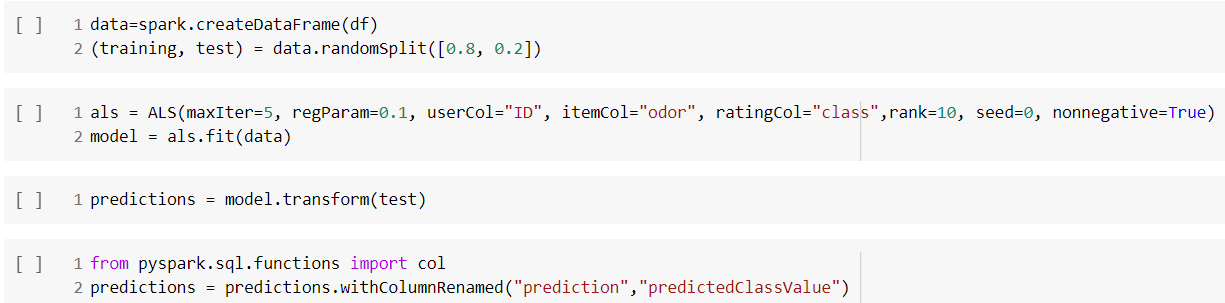


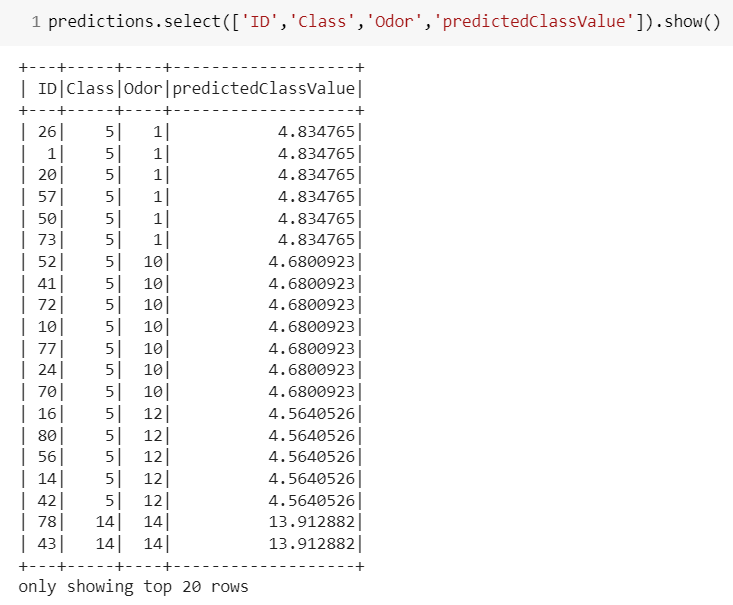
We have trained models that we can use further for prediction.

# PART 3: RECOMENDATION MODEL (USING PYSPARK ALS)

Apart from machine learning models, we have also generated a recommendation system that is capable of predicting the class of mushroom based on odor and other parameters.







We can further use this model to predict the class of mushroom.

*For each step by step execution, please refer the attached the code file(.ipynb).*

*Requirements: Jupyter notebook (alternatively, Google Collaboratory)*