Lab - Decision Trees

Q1. GINI Calculations

- Calculate the GINI value for a node with a 9/7 split (9 instances of class1, 7 instances of class2)
- Calculate the GINI values of the nodes with a 1 / 4 split, 2/1 split and 1/1 split
- Calculate the overall GINI value of a split into the above three nodes (weighted sum of the individual GINI values)

Q2a. Decision Tree for Iris data.

- Load iris data from CSV file
- Explore the data
  - df.head()
  - df.shape
  - df.species.value_counts()
  - df.describe()
  - scatter_matrix()
- Draw a scatter plot for sepal_length and sepal-width with different species identified. Use the code below.
- Do the same for petal_length and petal_width

```
# sepal_length and sepal_width for each species?
d = np.array(df)
plt.scatter(
   d[d[:,4] == "setosa", 0], d[d[:,4] == "setosa", 1],
   c='lightgreen', marker='s', edgecolor='black',
   label='setosa'
)
plt.scatter(
   d[d[:,4] == "versicolor", 0], d[d[:,4] == "versicolor", 1],
   c='orange', marker='o', edgecolor='black',
   label='versicolor'
)
plt.scatter(
   d[d[:,4] == "virginica", 0], d[d[:,4] == "virginica", 1],
   c='lightblue', marker='v', edgecolor='black',
   label='virginica'
)
plt.xlabel('sepal_length')
plt.ylabel('sepal_width')
plt.legend(loc="upper left")
plt.show()
```

Q2b.

Continue the iris example

- Split into training and test data
- Fit to a DecisionTreeClassifier
- Find the accuracy and confusion matrix.
- Plot the decision tree.
- Experiment with different values of max_depth.

model = DecisionTreeClassifier(max_depth=10)

Q3. Decision Tree – Diabetes Data

Create a decision tree model for the diabetes data set. The diabetes data set has no column headers.

```
# How to read a csv file with no column names and set the names of the Dtaframe
col_names = ['pregnant', 'glucose', 'bp', 'skin', 'insulin', 'bmi', 'pedigree', 'age', 'label']
df = pd.read_csv("data/diabetes.csv", header=None, names=col_names)
```

label is the target variable. 0 is a negative, 1 a positive.

- Load diabetes data from CSV file
- Explore the data
- Draw a scatter plot for glucose and insulin with different labels identified.
- Split into training and test data
- Fit to a DecisionTreeClassifier.
- Find the accuracy and confusion matrix.
- Plot the decision tree.
- Experiment with different values of max_depth.

Q4. Decision Tree – Iris Cross Validation

- Build a DecisionTreeClassifier for the iris data set and use cross validation to pick the optimal value of max_depth

Q5. Decision Tree – Diabetes Cross Validation

- Build a DecisionTreeClassifier for the diabetes data set and use cross validation to pick the optimal value of max_depth.

Q6. Decision Tree -Mushroom Data

- Load mushroom data from CSV file
- One hot ecode the features (X = get_dummies(X))
- Split into training and test data
- Fit to a DecisionTreeClassifier.
- Use cross validation to determine the optimal depth of the decision tree.
- For that depth find the accuracy and confusion matrix and plot the decision tree.