**Problem Statement:**

Draw decision tree using ID3 algorithm on golf playing dataset.

**Dataset Description:**

The dataset you provided appears to be related to predicting whether or not people will play golf based on different weather conditions.

Columns:

Outlook: Describes the weather outlook.

Categories: Sunny, Overcast, Rain.

Temp. (Temperature): Describes the temperature during the day.

Categories: Hot, Mild, Cool.

Humidity: Describes the humidity level.

Categories: High, Normal.

Wind: Describes the wind speed.

Categories: Weak, Strong.

Decision: The decision of whether or not to play golf.

Categories: Yes, No.

The dataset contains 14 records, each representing one instance of weather conditions and the corresponding decision regarding playing golf. The columns are a mix of categorical features (Outlook, Temp., Humidity, Wind) and a binary target variable (Decision).

**Procedure:**

Step 1: Begin with all the data points in the root of the tree.

Step 2: Calculate entropy for the dataset.

Step 3: For each attribute, calculate the information gain.

Step 4: Select the attribute with the highest information gain to split the dataset.

Step 5: Split the dataset based on the chosen attribute’s values.

Step 6: Repeat the process recursively for each subset.

Step 7: Stop when one of the stopping conditions is met (pure subsets or no attributes left to split).

**Source code:**

import numpy as np

import pandas as pd

import math

from collections import Counter

import graphviz

# Load dataset

df\_data = pd.read\_csv('GolfPlay.csv')

# Handle missing values by filling with the mode of each column

df\_data.fillna(df\_data.mode().iloc[0], inplace=True)

# Remove duplicate rows

df\_data.drop\_duplicates(inplace=True)

# Convert categorical data into numerical format

df = df\_data.apply(lambda x: x.astype('category').cat.codes if x.dtype == 'object' else x)

# Entropy Calculation

def entropy(y):

counter = Counter(y)

total = len(y)

return -sum((count/total) \* math.log2(count/total) for count in counter.values())

# Information Gain Calculation

def information\_gain(df, feature, target):

total\_entropy = entropy(df[target])

values = df[feature].unique()

weighted\_entropy = sum((len(df[df[feature] == v]) / len(df)) \* entropy(df[df[feature] == v][target]) for v in values)

return total\_entropy - weighted\_entropy

# ID3 Algorithm - Recursively Build the Decision Tree

def id3(df, features, target, tree=None):

if len(df[target].unique()) == 1:

return df[target].iloc[0]

if len(features) == 0:

return df[target].mode()[0]

best\_feature = max(features, key=lambda f: information\_gain(df, f, target))

tree = {best\_feature: {}}

for value in df[best\_feature].unique():

subset = df[df[best\_feature] == value]

tree[best\_feature][value] = id3(subset, [f for f in features if f != best\_feature], target)

return tree

# Build the decision tree

features = df.columns[:-1]

target = 'Decision'

decision\_tree = id3(df, features, target)

# Print Decision Tree

import pprint

pprint.pprint(decision\_tree)

# Function to visualize the tree

def visualize\_tree(tree, parent=None, graph=None):

if graph is None:

graph = graphviz.Digraph(format="png")

if isinstance(tree, dict):

for node, sub\_tree in tree.items():

if parent is None:

graph.node(node, label=node, shape="diamond")

else:

graph.node(node, label=node, shape="diamond")

graph.edge(parent, node)

for value, branch in sub\_tree.items():

child\_name = f"{node}\_{value}"

graph.node(child\_name, label=str(value))

graph.edge(node, child\_name)

visualize\_tree(branch, parent=child\_name, graph=graph)

else:

graph.node(str(tree), label=str(tree), shape="box")

graph.edge(parent, str(tree))

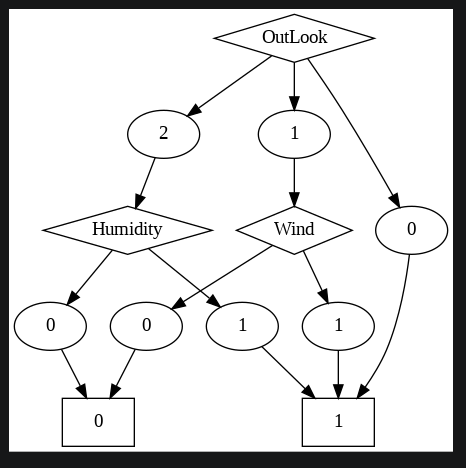
return graph

# Draw the decision tree

graph = visualize\_tree(decision\_tree)

graph.render("decision\_tree", view=True)

**Output:**



**Discussion:**

ID3 is a simple yet powerful decision tree algorithm that can be an excellent choice for classification tasks where interpretability is important. While it has some limitations, such as a tendency to overfit and favor features with more categories.