

Prepared by: Abu Naweem Khan & Sayed Muhammad Saifuddin

Decision Tree Construction via ID3

Objective:

The main goal is to implement ID3 algorithm for generating Decision Tree classifier. So, we have to write a program that generates a Decision Tree via the ID3 algorithm.

Background:

In this project, we were given a Tutorial describing the operation of the ID3 algorithm and five sample datasets. Among those datasets, we have used four of them and we have chosen the datafile format as below:

- Header: metadata describing the dataset
attributeName: attributeValues # list of attributes with their possible values
targetName: targetValues # values the target can take
- Data: Training set
attributeValues, targetValue # one example/line

The sample datasets that are involved to this classification project are:

- **Treatment** (HW #7): predict how a patient will respond to a new treatment
- **CAD**: predict if a patient is likely to incur coronary artery disease
- **Fishing**: is it a good day to fish?
- **Contact-lens**: suggest the best (out of three) contact-lens option

We have implemented our program as “non-recursive” first and then extended it to “recursive” for further investigation. We have written re-usable functions to generalize our program that it can work with either dataset. Furthermore, we have used **pyvis.network** package for the visualization. To validate the correctness of our program, we have tested it against the `DecisionTreeClassifier()` found in **scikit.learn**. For non-recursive part, the program was executed on the **Treatment** dataset. Then we have executed our recursive program to the **Fishing**, **CAD**, and **Contact-lens** datasets for further investigation. In addition, we have added “Classification mode” to our program for the unseen input instance and used our decision tree to output a prediction/classification. And finally, we were able to see the growing tree by using our program.

Methodology:

We have implemented the basic ID3 algorithm by using the below metrics for creating a decision tree classifier -

- A measure of purity:

$$\text{Entropy}(S) \equiv -\sum_{i=1}^k p_i \log_2 p_i$$

where S is the collection of examples, k is the number of categories, and p_i is the ratio of the cardinality of category i to the cardinality of S , as in $p_i = N_i / N$

Prepared by: Abu Naweem Khan & Sayed Muhammad Saifuddin

- The formula for Information Gain:

$$\text{Gain}(S, a) = \text{Entropy}(S) - \sum_{v=\text{values}(a)} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

where $\text{values}(a)$ is the set of all possible values for attribute a , and S_v is the subset of set S for which attribute a has value v .

The ID3 algorithm is –

ID3 (S)

```

if all examples in  $S$  are of the same class
    return a leaf with that class label
else if there are no more attributes to test
    return a leaf with the majority class label
else
    choose the attribute  $a$  that maximizes the Information Gain of  $S$ 
    let attribute  $a$  be the decision for the current node
    add a branch from the current node for each possible value  $v$  of attribute  $a$ 
    for each branch
        “sort” examples down the branches based on their value  $v$  of attribute  $a$ 
        recursively call ID3( $S_v$ ) on the set of examples now at each branch

```

```

def __tree_at(self, node: dict, attrs: list):
    """
    Uses Depth-first-search approach to build the tree
    recursively using ID3 algorithm.
    :param node: the current node to split
    :return:
    """
    if node['S']['entropy'] == 0 or len(attrs) == 0:
        self.__make_leaf(node)
        return

    self.__split(node, attrs)

    # TODO: remove split attr from list and recursive call
    new_attrs = attrs.copy()
    new_attrs.remove(node['split']['attr'])

    if node['branches'] is not None:
        for key in node['branches']:
            self.__tree_at(node['branches'][key], new_attrs)

```

Prepared by: Abu Naweem Khan & Sayed Muhammad Saifuddin

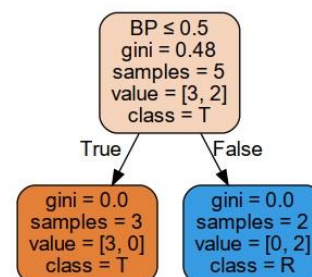
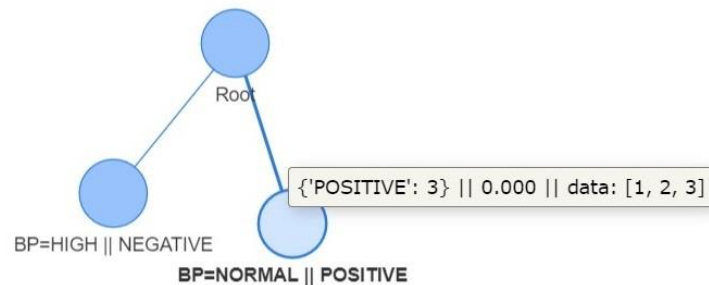
Visualization: [Display the Output]

Treatment Dataset:

```

2 Enter dataset name: treatment.data
3 *****
4 Custom implementation of Decision Tree.
5 *****
6
7 Step 1: Pre-processing and Initialization.
8 <class 'pandas.core.frame.DataFrame'>
9 RangeIndex: 5 entries, 0 to 4
10 Data columns (total 4 columns):
11 # Column Non-Null Count Dtype
12 ---
13 0 PULSE 5 non-null object
14 1 BP 5 non-null object
15 2 AGE 5 non-null object
16 3 TREAT 5 non-null object
17 dtypes: object(4)
18 memory usage: 288.0+ bytes
19 None
20
21 Step 2: Training...Done
22
23 Step 2: Plotting Tree...Done
24
25 Step 3: Prediction.
26 PULSE: normal
27 BP: normal
28 AGE: <25
29 Root -> BP=NORMAL || POSITIVE
30 Press "q" to quit, any key to continue:
31 PULSE: normal
32 BP: Low
33 AGE: <25
34 Root -> Not predictable: {'NEGATIVE': 'Prob: 0.400', 'POSITIVE': 'Prob: 0.600'}
35 Press "q" to quit, any key to continue: q
36 *****
37 sklearn implementation of Decision Tree.
38 *****
39
40 Step 1: Pre-processing and Initialization.
41 <class 'pandas.core.frame.DataFrame'>
42 RangeIndex: 5 entries, 0 to 4
43 Data columns (total 4 columns):
44 # Column Non-Null Count Dtype
45 ---
46 0 PULSE 5 non-null object
47 1 BP 5 non-null object
48 2 AGE 5 non-null object
49 3 TREAT 5 non-null object
50 dtypes: object(4)
51 memory usage: 288.0+ bytes
52 None
53
54 PULSE BP AGE ... BP_encoded AGE_encoded TREAT_encoded
55 0 NORMAL NORMAL <25 ... 0 0 0
56 1 NORMAL NORMAL 25-40 ... 0 1 0
57 2 RAPID NORMAL >40 ... 0 2 0
58 3 NORMAL HIGH >40 ... 1 2 1
59 4 RAPID HIGH >40 ... 1 2 1
60 [5 rows x 8 columns]
61
62 Step 2: Training...Done
63
64 Step 2: Plotting Tree...Done
65
66 Step 3: Prediction.
67 PULSE: Normal
68 BP: Normal
69 AGE: <25
70 Prediction: POSITIVE
71 Press "q" to quit, any key to continue: q
72
73 Process finished with exit code 0
74

```



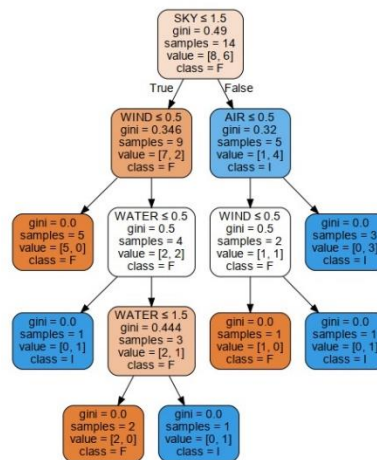
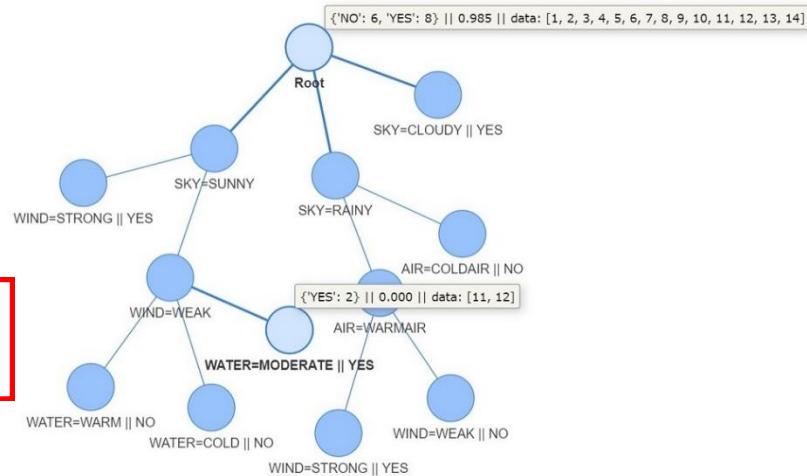
Prepared by: Abu Naweem Khan & Sayed Muhammad Saifuddin

Fishing Dataset:

```

2 Enter dataset name: fishing.data
3 *****
4 Custom implementation of Decision Tree.
5 *****
6
7 Step 1: Pre-processing and Initialization.
8 <class 'pandas.core.frame.DataFrame'>
9 RangeIndex: 14 entries, 0 to 13
10 Data columns (total 5 columns):
11 # Column Non-Null Count Dtype
12 ---
13 0 WIND 14 non-null object
14 1 AIR 14 non-null object
15 2 WATER 14 non-null object
16 3 SKY 14 non-null object
17 4 FISH 14 non-null object
18 dtypes: object(5)
19 memory usage: 688.0+ bytes
20 None
21
22 Step 2: Training...Done
23
24 Step 2: Plotting Tree...Done
25
26 Step 3: Prediction.
27 WIND: Strong
28 AIR: WarmAir
29 WATER: Warm
30 SKY: Sunny
31 Root -> SKY=SUNNY -> WIND=STRONG || YES
32 Press "q" to quit, any key to continue:
33 WIND: Moderate
34 AIR: WarmAir
35 WATER: Warm
36 SKY: Sunny
37 Root -> SKY=SUNNY -> Not predictable: {'NO': 'Prob: 0.250', 'YES': 'Prob: 0.750'}
38 Press "q" to quit, any key to continue: q
39 *****
40 sklearn implementation of Decision Tree.
41 *****
42
43 Step 1: Pre-processing and Initialization.
44 <class 'pandas.core.frame.DataFrame'>
45 RangeIndex: 14 entries, 0 to 13
46 Data columns (total 5 columns):
47 # Column Non-Null Count Dtype
48 ---
49 0 WIND 14 non-null object
50 1 AIR 14 non-null object
51 2 WATER 14 non-null object
52 3 SKY 14 non-null object
53 4 FISH 14 non-null object
54 dtypes: object(5)
55 memory usage: 688.0+ bytes
56 None
57 WIND AIR WATER ... WATER_encoded SKY_encoded FISH_encoded
58 0 STRONG WARMAIR WARM ... 0 0 0
59 1 WEAK WARMAIR WARM ... 0 0 1
60 2 STRONG WARMAIR WARM ... 0 1 0
61 3 STRONG WARMAIR MODERATE ... 1 2 0
62 4 STRONG COLDAIR COLD ... 2 2 1
63
64 [5 rows x 10 columns]
65
66 Step 2: Training...Done
67
68 Step 2: Plotting Tree...Done
69
70 Step 3: Prediction.
71 WIND: Strong
72 AIR: WarmAir
73 WATER: Warm
74 SKY: Sunny
75 Prediction: YES
76 Press "q" to quit, any key to continue: q
77
78 Process finished with exit code 0
79

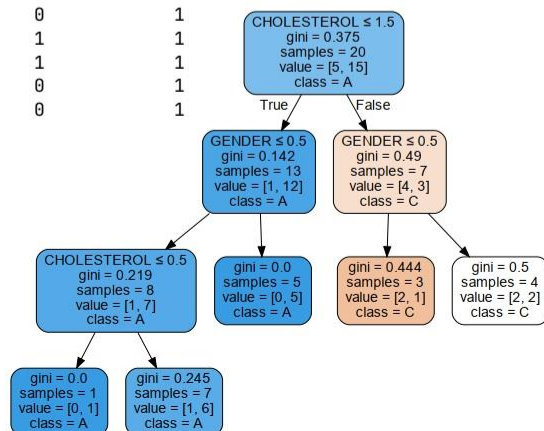
```



CAD Dataset:

Decision tree structure:

- Root
 - CHOLESTEROL=HIGH
 - GENDER=M || YES
 - GENDER=F || NO (highlighted in red box)
 - GENDER=M || NO
 - CHOLESTEROL=HIGH
 - GENDER=F || NO
 - GENDER=M || NO
 - CHOLESTEROL=NORMAL || NO
 - GENDER=F || NO
 - GENDER=M || NO



Contact-lens Dataset:

[illegible]

```

graph TD
    Root["TEAR-RATE ≤ 0.5  
gini = 0.538  
samples = 24  
value = [5, 4, 15]  
class = N"]
    Root -- True --> L1["gini = 0.0  
samples = 12  
value = [0, 0, 12]  
class = N"]
    Root -- False --> L2["ASTIGMATISM ≤ 0.5  
gini = 0.653  
samples = 12  
value = [5, 4, 3]  
class = C"]
    L2 -- True --> L3["AGE ≤ 1.5  
gini = 0.278  
samples = 6  
value = [5, 0, 1]  
class = C"]
    L2 -- False --> L4["PRESCRIPTION ≤ 0.5  
gini = 0.444  
samples = 6  
value = [0, 4, 2]  
class = O"]
    L3 -- True --> L5["gini = 0.0  
samples = 4  
value = [4, 0, 0]  
class = C"]
    L3 -- False --> L6["PRESCRIPTION ≤ 0.5  
gini = 0.5  
samples = 2  
value = [1, 0, 1]  
class = C"]
    L4 -- True --> L7["gini = 0.0  
samples = 3  
value = [0, 3, 0]  
class = O"]
    L4 -- False --> L8["AGE ≤ 0.5  
gini = 0.444  
samples = 3  
value = [0, 1, 2]  
class = N"]
    L6 -- True --> L9["gini = 0.0  
samples = 1  
value = [0, 0, 1]  
class = N"]
    L6 -- False --> L10["gini = 0.0  
samples = 1  
value = [1, 0, 0]  
class = C"]
    L8 -- True --> L11["gini = 0.0  
samples = 1  
value = [0, 1, 0]  
class = O"]
    L8 -- False --> L12["gini = 0.0  
samples = 2  
value = [0, 0, 2]  
class = N"]
  
```

Prepared by: Abu Naweem Khan & Sayed Muhammad Saifuddin

Analysis and Discussion:

We can see from the above visualization part that the output generated using the implemented algorithm and the output generated using the DecisionTreeClassifier() found in **scikit.learn** are same. So, we can say that our implemented algorithm is validated by a library package.

Future Work:

In future, we can extend our program by implementing the capability to process numeric (continuous-valued) attributes and extract the rule-base (IF-THEN) from our decision tree.

[Note: Please turn next page for the Source Code]

Prepared by: Abu Naweem Khan & Sayed Muhammad Saifuddin

Source Code:

3/15/22, 1:16 PM

decision_tree.py

```
1 import math
2 from collections import defaultdict
3 from queue import Queue
4
5 import pandas as pd
6 from pyvis.network import Network
7
8
9 class DecisionTree:
10     __data = None
11     __attrs = None
12     __cls = None
13     __tree = defaultdict(dict)
14
15     def __init__(self, df: pd.DataFrame, attrs=None):
16         self.__data = df
17         self.__gen_attr_set(attrs)
18
19     def __gen_attr_set(self, attrs=None):
20         """
21         If attrs is None, generates the unique values for each attribute
22         from the dataset. If any value of an attribute is not present in
23         the dataset, this will fail to discover all possible values of an
24         attribute.
25         In that case attrs must not be None.
26
27         Considers class to be the last column.
28         :return:
29         """
30         self.__attrs = dict()
31         self.__cls = dict()
32
33         if attrs is not None:
34             k, l = attrs.popitem()
35             self.__cls = {k: l}
36             self.__attrs = attrs.copy()
37             return
38
39         columns = self.__data.columns.to_list()
40         for col in columns[:-1]:
41             self.__attrs[col] = self.__data[col].unique().tolist()
42
43         self.__cls[columns[-1]] = self.__data[columns[-1]].unique().tolist()
44
45     def __entropy_of(self, cls_freq: dict) -> float:
46         total = 0
47         for i in cls_freq.values():
48             total += i
49
50         entropy = 0
51         for i in cls_freq.values():
52             entropy -= (i / total) * math.log2((i / total))
53
54         return entropy
55
56     def __gain_of(self, node_entropy, col, col_cls_name, df):
57         S = defaultdict(dict)
58         d = df.groupby(by=[col, col_cls_name]).count().iloc[:, 0]
```


Prepared by: Abu Naweem Khan & Sayed Muhammad Saifuddin

3/15/22, 1:16 PM

decision_tree.py

```

60     for key in d.keys():
61         attr, cls = key
62         S[attr][cls] = d[key]
63
64     gain = node_entropy
65     for key in S:
66         cls_freq = S[key]
67         attr_total = d.loc[key].sum()
68         total = d.sum()
69         S[key]['entropy'] = self.__entropy_of(cls_freq)
70         gain -= (S[key]['entropy'] * attr_total / total)
71
72     return S, gain
73
74 def __split(self, node: dict, attrs: list):
75     """
76     Splits the node on the attribute with the highest gain value.
77     Uses Entropy for gain calculation.
78     :param node: node to split
79     :param attrs: allowed attributes to split on
80     """
81
82     col_cls_name = list(self.__cls.keys())[0]
83     df = self.__data.loc[node['data']]
84
85     max_gain = -1
86     max_S = None
87     split_col = None
88     for col in attrs:
89         S, gain = self.__gain_of(node['S']['entropy'], col, col_cls_name, df)
90
91         if max_gain < gain:
92             max_gain = gain
93             max_S = S
94             split_col = col
95
96     node['split']['attr'] = split_col
97     node['split']['gain'] = max_gain
98     node['branches'] = dict()
99
100    for key in max_S:
101        cls_freq = max_S[key].copy()
102        cls_freq.pop('entropy')
103        n = {
104            'data': [i for i in df[df[split_col] == key].index],
105            'S': {
106                'cls_freq': cls_freq,
107                'entropy': max_S[key]['entropy']
108            },
109            'split': {'attr': None, 'gain': None},
110            'branches': None,
111            'label': f'{split_col}={key}'
112        }
113        node['branches'][key] = n
114
115 def __make_leaf(self, node: dict):
116     """
117     Marks a node as a leaf, by assigning the majority class as the label.
118     :param node:
119     :return:

```

Prepared by: Abu Naweem Khan & Sayed Muhammad Saifuddin

3/15/22, 1:16 PM

decision_tree.py

```

120     """
121     max_freq = -1
122     label = ''
123     for key in node['S']['cls_freq']:
124         freq = node['S']['cls_freq'][key]
125         if max_freq < freq:
126             max_freq = freq
127             label = key
128
129     node['label'] = f"{node['label']} || {label}"
130
131 def __tree_at(self, node: dict, attrs: list):
132     """
133     Uses Depth-first-search approach to build the tree
134     recursively using ID3 algorithm.
135     :param node: the current node to split
136     :return:
137     """
138     if node['S']['entropy'] == 0 or len(attrs) == 0:
139         self.__make_leaf(node)
140         return
141
142     self.__split(node, attrs)
143
144     # TODO: remove split attr from list and recursive call
145     new_attrs = attrs.copy()
146     new_attrs.remove(node['split']['attr'])
147
148     if node['branches'] is not None:
149         for key in node['branches']:
150             self.__tree_at(node['branches'][key], new_attrs)
151
152 def train(self):
153     cls_col_name = list(self.__cls.keys())[0]
154     attr_names = list(self.__attrs.keys())
155
156     self.__tree = {
157         'data': [i for i in range(len(self.__data))],
158         'S': {
159             'cls_freq': self.__data.groupby(cls_col_name).count()[
160                 attr_names[0]].to_dict()
161         },
162         'split': {'attr': None, 'gain': None},
163         'branches': None,
164         'label': 'Root'
165     }
166
167     self.__tree['S']['entropy'] = self.__entropy_of(self.__tree['S']
168 ['cls_freq'])
169
170     self.__tree_at(node=self.__tree, attrs=attr_names)
171
172 def predict(self, attr_vals: dict):
173     node = self.__tree
174
175     predict_vals = dict()
176     for key in attr_vals:
177         predict_vals[key.upper()] = attr_vals[key].upper()
178
179     while True:

```

Prepared by: Abu Naweem Khan & Sayed Muhammad Saifuddin

3/15/22, 1:16 PM

decision_tree.py

```

179         if node['branches'] is None:
180             print(f'{node["label"]}')
181             return
182
183         print(f"{node['label']} -> ", end='')
184
185         split = node['split']['attr']
186         if split not in predict_vals:
187             print(f'{split} not in given values {attr_vals.keys()}')
188             return
189
190         attr_val = predict_vals[split]
191
192         for key in node['branches']:
193             n = node['branches'][key]
194             if n['label'].startswith(f"{split}={attr_val}"):
195                 node = n
196                 break
197         else:
198             total = len(node['data'])
199             cls_col = list(self.__cls.keys())[0]
200             counts = self.__data.loc[node['data']].groupby(cls_col).count()[
201                 self.__data.columns[0]].to_dict()
202             for key in counts:
203                 counts[key] = f"Prob: {counts[key] / total:.3f}"
204
205             print(f'Not predictable: {counts}')
206             return
207
208     def plot_tree(self):
209         # print(self.__tree)
210         net = Network()
211         net.height = '100%'
212         net.width = '100%'
213
214         node = self.__tree
215         node['id'] = 1
216         title = f"{node['S']['cls_freq']} || " \
217             f"{node['S']['entropy']:.3f} || " \
218             f"data: {[i + 1 for i in node['data']]}"
219         net.add_node(1, label=node['label'], title=title)
220
221         q = Queue()
222         q.put(node)
223
224         self.__bfs_plot(q, net)
225
226         net.show('tree.html')
227
228     def __bfs_plot(self, q: Queue, net: Network):
229         i = 1
230         while not q.empty():
231             node = q.get()
232             uid = node['id']
233
234             if node['branches'] is not None:
235                 for key in node['branches'].keys():
236                     n = node['branches'][key]
237                     n['id'] = uid + i
238                     title = f"{n['S']['cls_freq']} || " \

```

Prepared by: Abu Naweem Khan & Sayed Muhammad Saifuddin

3/15/22, 1:16 PM

decision_tree.py

```
239         f"{n['S']['entropy']:.3f} || " \
240         f"data: {[i + 1 for i in n['data']]}"
241     net.add_node(uid + i, label=n['label'], title=title)
242     net.add_edge(uid, uid + i)
243     i += 1
244     q.put(n)
245
246
247 def pre_processing(filepath: str) -> tuple[pd.DataFrame, dict]:
248     data = []
249     attrs = defaultdict(list)
250     with open(filepath, mode='r') as fin:
251         for line in fin:
252             line = line.strip()
253
254             if line.startswith("#attributes") or line.startswith("#target"):
255                 for l in fin:
256                     l = l.strip().upper()
257                     if len(l) == 0:
258                         break
259                     parts = l.split(":")
260                     attrs[parts[0][1:]] = parts[1].strip().split(",")
261
262             elif line.startswith("#data"):
263                 for l in fin:
264                     l = l.strip().upper()
265                     if len(l) == 0:
266                         break
267                     parts = l.split(",")
268                     data.append(parts)
269
270     df = pd.DataFrame(data)
271     df.columns = list(attrs.keys())
272     print(df.info())
273
274     return df, attrs
275
```

Note: Please turn next page for rest of the codes:

Prepared by: Abu Naweem Khan & Sayed Muhammad Saifuddin

3/15/22, 1:18 PM

project.py

```
1 from collections import defaultdict
2
3 import graphviz
4 import pandas as pd
5 from sklearn import tree
6
7 from project3.decision_tree import DecisionTree, pre_processing
8
9
10 def custom_dtree(filename: str):
11     print('*****')
12     print('\tCustom implementation of Decision Tree.')
13     print('*****')
14
15     print('\nStep 1: Pre-processing and Initialization.')
16     df, attr = pre_processing(filename)
17     dt = DecisionTree(df, attr)
18
19     print('\nStep 2: Training...', end='')
20     dt.train()
21     print('Done')
22
23     print('\nStep 2: Plotting Tree...', end='')
24     dt.plot_tree()
25     print('Done')
26
27     print('\nStep 3: Prediction.')
28     q = 'a'
29     inp = defaultdict(str)
30     for item in attr:
31         inp[item] = ''
32     while q != 'q':
33         for key in inp:
34             val = input(f"{key}: ")
35             inp[key] = val
36
37     # dt.predict({'WIND': 'WEAK', 'WATER': 'MODERATE', 'AIR': 'WARMAIR', 'SKY':
    'RAINY'})
38     dt.predict(inp)
39     q = input('Press "q" to quit, any key to continue: ')
40
41
42 def load_and_encode(filepath: str) -> tuple[pd.DataFrame, dict]:
43     df, attrs = pre_processing(filepath)
44
45     for col in attrs.keys():
46         series = df[col].tolist()
47         for i in range(len(series)):
48             series[i] = attrs[col].index(series[i])
49         df[f'{col}_encoded'] = series
50
51     return df, attrs
52
53
54 def sklearn_dtree(filename: str):
55     print('*****')
56     print('\tsklearn implementation of Decision Tree.')
57     print('*****')
58
```


Prepared by: Abu Naweem Khan & Sayed Muhammad Saifuddin

3/15/22, 1:18 PM

project.py

```
59 print('\nStep 1: Pre-processing and Initialization.')
60 df, attrs = load_and_encode(filename)
61 cols = list(attrs.keys())
62 print(df.head())
63
64 print('\nStep 2: Training...', end='')
65 x = []
66 for i in df.index:
67     x.append(df.iloc[i, len(cols):2*len(cols)-1].tolist())
68
69 y = df[df.columns[-1]].tolist()
70
71 dt = tree.DecisionTreeClassifier()
72 dt = dt.fit(x, y)
73 print('Done')
74
75 print('\nStep 2: Plotting Tree...', end='')
76 dot_tree = tree.export_graphviz(dt, out_file=None,
77                                 feature_names=cols[:-1],
78                                 class_names=cols[-1],
79                                 filled=True, rounded=True,
80                                 special_characters=True)
81 graph = graphviz.Source(dot_tree)
82 graph.render('decision_tree')
83 print('Done')
84
85 print('\nStep 3: Prediction.')
86 q = 'a'
87 inp = []
88 while q != 'q':
89     for key in cols[:-1]:
90         val = input(f"{key}: ")
91         inp.append(attrs[key].index(val.upper()))
92
93     # dt.predict({'WIND': 'WEAK', 'WATER': 'MODERATE', 'AIR': 'WARMAIR', 'SKY':
'RAINY'})
94     cls = dt.predict([inp])[0]
95     print(f"Prediction: {attrs[cols[-1]][cls]}")
96     q = input('Press "q" to quit, any key to continue: ')
97
98
99 def run():
100     filename = input('Enter dataset name: ')
101     custom_dtree(filename)
102     sklearn_dtree(filename)
103
```

Thank You!!!