به نام خدا

ایمان علیپور

۹۸۱۰۲۰۲۴

تمرین پیاده‌سازی ۲ هوش مصنوعی

استاد: دکتر هجران‌دوست

هدف:
پیاده سازی decision tree.

چالش ها:
اصلی ترین چالشی که با آن روبرو شدم این بود که زمانی که کدی که برای رستوران زده بودم را به دیتاست دوم که برای دیابت بود انتقال دادم، الگوریتمم اجرایش تمام نمیشد و تلاش میکرد تا میشود عمق را زیاد کند، برای همین برای آن مجبور شدم محدودیت عمق قرار دهم که خروجی تولید شود.

چالش دیگر من چیزی بود که نفهمیدم چگونه ایجاد میشود، بعضی وقت ها که تابع محاسبه گر دقت را روی داده های train صدا میزدم، اروری میگرفتم که علت آن را نمیدانم اما احتمالا بخاطر overflow بوده و آنرا پیدا نکردم. این تابع برای داده های test هیچوقت خطا نداد!

نحوه اجرا:
فایل های csv را در کنار فایل های اصلی کد قرار دهید و فایل کد را اجرا کنید.
توجه کنید برای هردو درخت من یک تابع نوشتم که entropy را نیز چاپ کند، اما چون درخت ناخوانا میشد در نهایت از آن استفاده نکردم.

خروجی:
خروجی اینگونه است که هربار در درخت پایین میرویم یک indentationi ایجاد میشود. برای هر دو کد یک تابع نوشتم که درخت را با مقدار آنتروپی تست کند اما آنرا کامنت کردم چون خروجی را زشت میکرد!

نحوه پیاده سازی:
بسیاری از توابع برای بخش رستوران را مشابه اسلاید ها و با توجه به سودوکد های اسلاید نوشتم، از جمله توابع آنتروپی و gain و remainder و تابع training برای همین توضیحی برای آنها اینجا نمی‌آورم.

تابع انتخاب attribute:

```
33  def select_an_attributesibute(data_file, attributes, res_name):
34      attributes_importance = {}
35      entropy = {}
36      remains = {}
37      for tmp_attributes in attributes:
38          attributes_importance[tmp_attributes], entropy[tmp_attributes], remains[tmp_attributes] = gain(data_file, tmp_attributes, res_name)
39      answer = max(attributes_importance.items(), key=operator.itemgetter(1))[0]
40      return answer, attributes_importance[answer], entropy[answer], remains[answer]
41
```

این تابع با بررسی attribute هایی که هنوز استفاده نشده‌اند، با توجه به میزان gain هرکدام، بهترین را انتخاب کرده و آنرا برمیگرداند.

```
16
17  def all_similar(data_file: pd.DataFrame):
18      a = data_file.to_numpy()
19      return (a[0] == a[1:]).all()
```

این تابع چک میکند آیا اعضای یک پنجره داده یا dataframe یکی هستند یا نه یعنی آیا تقسیم بندی ای داریم که نتایج آن همگی برابر باشند و یا نه.

```
def decicion_tree_training(examples: pd.DataFrame, attributesibutes: list, parent_examples, outcome_name, curr_depth
                           ):
    if examples.empty:
        return value_node_of_plural_attributesibute(parent_examples, outcome_name)
    if all_similar(examples[outcome_name]):
        return decicion_tree_leaf(examples[outcome_name].iloc[0], entropy=0)
    if not attributesibutes:
        return value_node_of_plural_attributesibute(examples, outcome_name)
    coloumns = list(examples.columns)
    coloumns.remove(outcome_name)
    attributes, gain_, entropy_, remains_ = select_an_attributesibute(examples, coloumns, outcome_name)
    all_values = examples[attributes].unique()
    tree = decicion_tree_multi_way(attributes, value_node_of_plural_attributesibute(examples, outcome_name), gain_=gain_, entropy=entropy_,
                       remainder_=remains_)
    for vk in all_values:
        new_coloumns = attributesibutes
        if (attributes in new_coloumns):
            new_coloumns.remove(attributes)
        subtree = decicion_tree_training(examples[examples[attributes] == vk], new_coloumns, examples, outcome_name, curr_depth + 1
                                         )
        tree.add(vk, subtree)
    return tree
```

این تابع قلب اصلی کد است و عملیات یادگیری را انجام میدهد، همچنین آنرا مشابه اسلاید ها پیاده سازی کردم.

حال کلاس ها را توضیح میدهم:

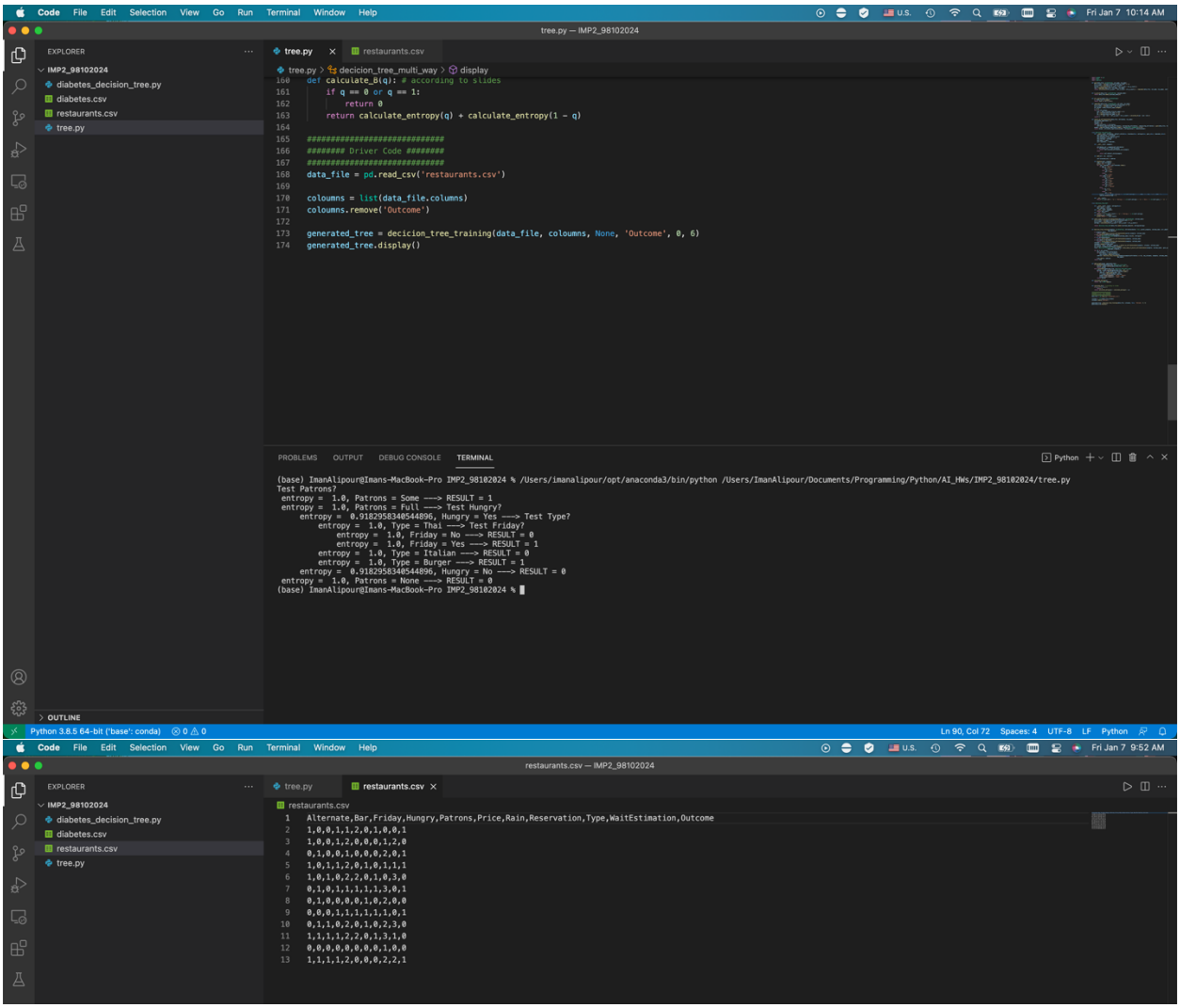
```
class decicion_tree_leaf:
    def __init__(self, result, entropy=None):
        self.result = result
        self.entropy = entropy
    def __call__(self, example):
        return self.result
    def __str__(self):
        string = ' ' + str(self.result) + '\n'
        return string
    def display(self, scope=0):
        print('RESULT =', self.result)
```

کلاس برگ درخت تصمیم که برای آن توابع روتین call و str را نوشته تا اگر خواستم آنرا روی خود نود صدا بزنم ممکن باشد، همچنین بتوان انرا به string تبدیل کرد و چاپ کرد. هر نود برگ مقدار entropy و result خود را نگه

میدارد.

```python
class decicion_tree_multi_way:
    def __init__(self, attributes, default_child=None, branches=None, entropy=None, gain_=None, remainder_=None):
        self.attributes = attributes
        self.default_child = default_child
        self.branches = branches or {}
        self.entropy = entropy
        self.gain_ = gain_
        self.remainder_ = remainder_

    def __call__(self, example):

        attributes_val = example[self.attributes]
        if attributes_val in self.branches:
            return self.branches[attributes_val](example)
        else:

            return self.default_child(example)

    def add(self, val, subtree):

        self.branches[val] = subtree

    def print_tree(self, scope=0):
        name = self.attributes
        print('Test ' + name + '?')
        for (val, subtree) in self.branches.items():
            if name == 'Patrons':
                if val == 0:
                    val = 'None'
                elif val == 1:
                    val = 'Some'
                else:
                    val = 'Full'
            elif name =='Type':
                if val == 0:
                    val = 'French'
                elif val == 1:
                    val = 'Thai'
                elif val == 2:
                    val = 'Burger'
                elif val == 3:
                    val = 'Italian'
            else:
                if val == 0:
                    val = 'No'
                else:
                    val = 'Yes'
            print(' ' * 4 * scope, 'entropy = ', str(self.entropy) + ',', name, '=', val, '--->', end=' ')
            subtree.print_tree(scope + 1)
```

کلاس decision tree که علت multiway بخاطر ایجاد قابلیت چند راهی در آن است، مجددا تابع call را برای آن نوشتم تا بتوان آنرا روی node صدا زد و بازگشتی آنرا پیمایش کرد.
همچنین بجای str برای آن تابع print_tree را نوشتم تا بتوان درخت را چاپ کرد.
در انتها نتیجه خروجی کد برای دیتای رستوران ها که فایل آنرا دستی ساختم اینگونه است:



درخت تولید شده با درخت اسلاید فرق میکند اما درست است، فرمت خروجی هم اینگونه است که هرلایه فرورفتگی نشانه پایین رفتن یک لایه در درخت است، اگر نیاز به تست جدید باشد ابتدا سوال مطرح میشود و زیر درخت چاپ میشود.

کد بخش دیابت:
حال با استفاده از کد قسمت قبل و ایجاد اندکی تغییر در ان، برای دیتاست دیابت کد را توضیح میدهم.

اولا کد نیاز به کمی تغییر داشت، همانطور که در چالش ها گفتم، الگوریتم من با این دیتاست هی تلاش میکرد عمق را زیاد کند پس من با اضافه کردن یک limit برای عمق این چالش را حل کردم که نمیدانم ایده خوبی بود یا نه. بقیه توابع اضافه شده اینها هستند:

```
179    def accuracy_calculator(test_data, diabetes_decision_tree):
180        number_of_correct_guesses = 0
181        for i in range(len(test_data)):
182            test_res = diabetes_decision_tree(test_data.iloc[i])
183            if test_res == test_data.iloc[i].Outcome:
184                number_of_correct_guesses += 1
185
186        return (number_of_correct_guesses / len(test_data) * 100)
187
```

که دقت را برای داده های داده شده محاسبه میکند.

```
93    def discrete_find_clusters(data_file: pd.DataFrame, column_name, number_of_clusters):
94        max_i_value = data_file[column_name].max()
95        min_i_value = data_file[column_name].min()
96        difference = (max_i_value - min_i_value) / number_of_clusters
97        clusters = []
98        clusters.append(round(min_i_value - difference, 2))
99        for i in range(number_of_clusters):
100            clusters.append(round(min_i_value + difference * i, 2))
101        clusters.append(round(max_i_value, 2))
102        clusters.append(round(max_i_value + difference, 2))
103        return clusters
```

که داده هارا کلاستر بندی میکند و همانطور که در متن تمرین آمده آنها را گسسته میکند، درواقع مرز دسته هارا با min و max مشخص میکند و یک دسته برای مقادیر بیشتر و یا کمتر هم درنظر میگیرد(همانطور که در متن تمرین

آمده بود)

```python
def discretify_data(training_data: pd.DataFrame, test_data: pd.DataFrame):
    coloumn_name = list(training_data.columns)

    Pregnancies_clusters = discrete_find_clusters(training_data, 'Pregnancies', 5)
    training_data = make_discrete(training_data, 'Pregnancies', Pregnancies_clusters)
    test_data = make_discrete(test_data, 'Pregnancies', Pregnancies_clusters)

    Glucose_clusters = discrete_find_clusters(training_data, 'Glucose', 5)
    training_data = make_discrete(training_data, 'Glucose', Glucose_clusters)
    test_data = make_discrete(test_data, 'Glucose', Glucose_clusters)

    BloodPressure_clusters = discrete_find_clusters(training_data, 'BloodPressure', 5)
    training_data = make_discrete(training_data, 'BloodPressure', BloodPressure_clusters)
    test_data = make_discrete(test_data, 'BloodPressure', BloodPressure_clusters)

    SkinThickness_clusters = discrete_find_clusters(training_data, 'SkinThickness', 5)
    training_data = make_discrete(training_data, 'SkinThickness', SkinThickness_clusters)
    test_data = make_discrete(test_data, 'SkinThickness', SkinThickness_clusters)

    Insulin_clusters = discrete_find_clusters(training_data, 'Insulin', 5)
    training_data = make_discrete(training_data, 'Insulin', Insulin_clusters)
    test_data = make_discrete(test_data, 'Insulin', Insulin_clusters)

    BMI_clusters = discrete_find_clusters(training_data, 'BMI', 5)
    training_data = make_discrete(training_data, 'BMI', BMI_clusters)
    test_data = make_discrete(test_data, 'BMI', BMI_clusters)

    DiabetesPedigreeFunction_clusters = discrete_find_clusters(data_file, 'DiabetesPedigreeFunction', 5)
    training_data = make_discrete(training_data, 'DiabetesPedigreeFunction', DiabetesPedigreeFunction_clusters)
    test_data = make_discrete(test_data, 'DiabetesPedigreeFunction', DiabetesPedigreeFunction_clusters)

    Age_clusters = discrete_find_clusters(data_file, 'Age', 5)
    training_data = make_discrete(training_data, 'Age', Age_clusters)
    test_data = make_discrete(test_data, 'Age', Age_clusters)
```

این تابع با خروجی تابع قبل ستون هارا گسسته سازی میکند، همچنین من مقادیر مختلفی را برای تعداد کلاستر ها امتحان کردم که ۵ و ۶ بهتر بودند و ۵ را نگه داشتم چون نسبت به ۶ عملکرد بهتری داشت.

```python
data_file = pd.read_csv('diabetes.csv')
training_data, test_data = train_test_split(data_file, test_size=0.2)
training_data, test_data = discretify_data(training_data, test_data)
```

اینگونه داده هارا از فایل خواندم و با تابع آماده sklearn دیتاست را به یک بخش تست و یک بخش train تقسیم کردم که بخش تست ۰.۲ کل داده ها باشد، سپس با تابع discritify data داده هارا گسسته کردم، حال کافیست الگوریتم را اجرا کنیم

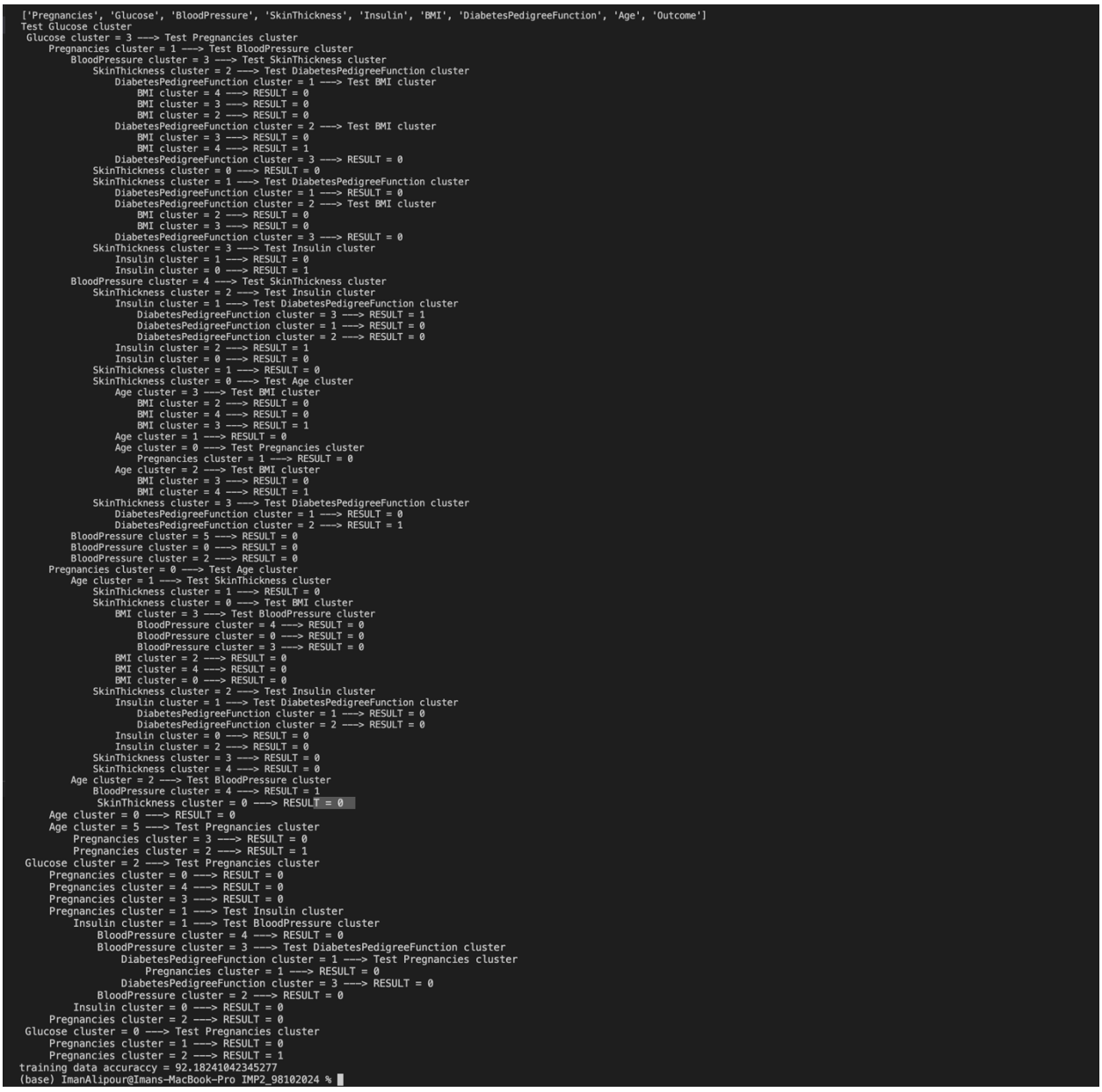
خروجی اینگونه است:

```
            BloodPressure cluster = 5 ---> RESULT = 0
            BloodPressure cluster = 0 ---> RESULT = 0
        Age cluster = 4 ---> RESULT = 1
        Age cluster = 0 ---> RESULT = 0
        Age cluster = 3 ---> RESULT = 0
    Pregnancies cluster = 2 ---> Test DiabetesPedigreeFunction cluster
        DiabetesPedigreeFunction cluster = 1 ---> Test BMI cluster
            BMI cluster = 3 ---> Test Age cluster
                Age cluster = 0 ---> RESULT = 0
                Age cluster = 1 ---> Test SkinThickness cluster
                    SkinThickness cluster = 2 ---> RESULT = 0
                    SkinThickness cluster = 1 ---> RESULT = 0
                    SkinThickness cluster = 0 ---> RESULT = 0
                Age cluster = 2 ---> Test SkinThickness cluster
                    SkinThickness cluster = 2 ---> RESULT = 0
                    SkinThickness cluster = 1 ---> RESULT = 0
                    SkinThickness cluster = 0 ---> RESULT = 0
                Age cluster = 3 ---> Test Insulin cluster
                    Insulin cluster = 0 ---> RESULT = 1
                    Insulin cluster = 1 ---> RESULT = 0
                Age cluster = 4 ---> RESULT = 0
            BMI cluster = 2 ---> RESULT = 0
            BMI cluster = 4 ---> Test SkinThickness cluster
                SkinThickness cluster = 2 ---> RESULT = 0
                SkinThickness cluster = 0 ---> Test BloodPressure cluster
                    BloodPressure cluster = 4 ---> RESULT = 0
                    BloodPressure cluster = 3 ---> RESULT = 1
                SkinThickness cluster = 3 ---> RESULT = 0
            BMI cluster = 5 ---> RESULT = 1
        DiabetesPedigreeFunction cluster = 3 ---> RESULT = 1
        DiabetesPedigreeFunction cluster = 2 ---> Test SkinThickness cluster
            SkinThickness cluster = 1 ---> Test Pregnancies cluster
                Pregnancies cluster = 2 ---> Test Pregnancies cluster
                    Pregnancies cluster = 2 ---> RESULT = 0
            SkinThickness cluster = 0 ---> RESULT = 0
            SkinThickness cluster = 3 ---> RESULT = 1
            SkinThickness cluster = 2 ---> Test BloodPressure cluster
                BloodPressure cluster = 2 ---> RESULT = 0
                BloodPressure cluster = 4 ---> Test Age cluster
                    Age cluster = 2 ---> RESULT = 1
                    Age cluster = 1 ---> RESULT = 0
                BloodPressure cluster = 3 ---> RESULT = 0
    Pregnancies cluster = 3 ---> Test SkinThickness cluster
        SkinThickness cluster = 0 ---> Test Age cluster
            Age cluster = 2 ---> Test BloodPressure cluster
                BloodPressure cluster = 3 ---> Test DiabetesPedigreeFunction cluster
                    DiabetesPedigreeFunction cluster = 1 ---> RESULT = 0
                    DiabetesPedigreeFunction cluster = 2 ---> RESULT = 1
                BloodPressure cluster = 0 ---> RESULT = 0
                BloodPressure cluster = 4 ---> RESULT = 0
                BloodPressure cluster = 5 ---> RESULT = 0
            Age cluster = 1 ---> Test BMI cluster
                BMI cluster = 3 ---> Test BloodPressure cluster
                    BloodPressure cluster = 0 ---> RESULT = 0
                    BloodPressure cluster = 3 ---> RESULT = 0
                    BloodPressure cluster = 4 ---> RESULT = 1
                BMI cluster = 2 ---> RESULT = 0
                BMI cluster = 0 ---> RESULT = 1
            Age cluster = 4 ---> RESULT = 0
            Age cluster = 3 ---> RESULT = 0
        SkinThickness cluster = 3 ---> RESULT = 0
        SkinThickness cluster = 2 ---> Test Age cluster
            Age cluster = 2 ---> Test Insulin cluster
                Insulin cluster = 2 ---> RESULT = 1
                Insulin cluster = 0 ---> Test BloodPressure cluster
                    BloodPressure cluster = 4 ---> RESULT = 0
                    BloodPressure cluster = 3 ---> RESULT = 0
                    BloodPressure cluster = 5 ---> RESULT = 1
                Insulin cluster = 1 ---> RESULT = 0
            Age cluster = 3 ---> RESULT = 1
            Age cluster = 1 ---> RESULT = 1
```

```
        SkinThickness cluster = 1 ---> RESULT = 0
    Pregnancies cluster = 4 ---> Test BloodPressure cluster
        BloodPressure cluster = 3 ---> Test SkinThickness cluster
            SkinThickness cluster = 3 ---> RESULT = 0
            SkinThickness cluster = 0 ---> Test Age cluster
                Age cluster = 2 ---> RESULT = 1
                Age cluster = 3 ---> RESULT = 0
        BloodPressure cluster = 4 ---> RESULT = 0
    Pregnancies cluster = 5 ---> RESULT = 1
Glucose cluster = 5 ---> Test BMI cluster
    BMI cluster = 3 ---> Test Insulin cluster
        Insulin cluster = 0 ---> Test DiabetesPedigreeFunction cluster
            DiabetesPedigreeFunction cluster = 1 ---> Test Pregnancies cluster
                Pregnancies cluster = 2 ---> Test BloodPressure cluster
                    BloodPressure cluster = 3 ---> RESULT = 1
                    BloodPressure cluster = 4 ---> RESULT = 0
                    BloodPressure cluster = 0 ---> RESULT = 1
                Pregnancies cluster = 3 ---> Test Age cluster
                    Age cluster = 3 ---> RESULT = 1
                    Age cluster = 4 ---> RESULT = 0
                    Age cluster = 2 ---> RESULT = 0
                Pregnancies cluster = 0 ---> RESULT = 1
                Pregnancies cluster = 5 ---> RESULT = 1
            DiabetesPedigreeFunction cluster = 3 ---> RESULT = 1
            DiabetesPedigreeFunction cluster = 2 ---> Test Pregnancies cluster
                Pregnancies cluster = 3 ---> RESULT = 0
                Pregnancies cluster = 1 ---> RESULT = 1
                Pregnancies cluster = 2 ---> RESULT = 0
        Insulin cluster = 1 ---> Test BloodPressure cluster
            BloodPressure cluster = 3 ---> Test Pregnancies cluster
                Pregnancies cluster = 2 ---> RESULT = 1
                Pregnancies cluster = 1 ---> Test SkinThickness cluster
                    SkinThickness cluster = 2 ---> RESULT = 0
                    SkinThickness cluster = 1 ---> RESULT = 1
                Pregnancies cluster = 3 ---> RESULT = 1
            BloodPressure cluster = 4 ---> RESULT = 0
        Insulin cluster = 2 ---> RESULT = 1
        Insulin cluster = 3 ---> Test Pregnancies cluster
            Pregnancies cluster = 3 ---> RESULT = 1
            Pregnancies cluster = 0 ---> RESULT = 1
            Pregnancies cluster = 1 ---> RESULT = 0
        Insulin cluster = 4 ---> RESULT = 1
        Insulin cluster = 5 ---> RESULT = 1
    BMI cluster = 4 ---> Test Age cluster
        Age cluster = 1 ---> RESULT = 1
        Age cluster = 2 ---> Test Pregnancies cluster
            Pregnancies cluster = 3 ---> RESULT = 1
            Pregnancies cluster = 1 ---> RESULT = 1
            Pregnancies cluster = 0 ---> RESULT = 1
            Pregnancies cluster = 2 ---> RESULT = 0
        Age cluster = 3 ---> RESULT = 1
        Age cluster = 4 ---> RESULT = 0
    BMI cluster = 2 ---> RESULT = 1
    BMI cluster = 5 ---> Test Insulin cluster
        Insulin cluster = 1 ---> RESULT = 1
        Insulin cluster = 2 ---> RESULT = 0
        Insulin cluster = 5 ---> RESULT = 0
Glucose cluster = 4 ---> Test Age cluster
    Age cluster = 4 ---> Test BMI cluster
        BMI cluster = 3 ---> Test Pregnancies cluster
            Pregnancies cluster = 2 ---> Test Insulin cluster
                Insulin cluster = 0 ---> Test SkinThickness cluster
                    SkinThickness cluster = 2 ---> RESULT = 0
                    SkinThickness cluster = 0 ---> RESULT = 0
                Insulin cluster = 2 ---> RESULT = 1
            Pregnancies cluster = 3 ---> RESULT = 0
            Pregnancies cluster = 4 ---> RESULT = 0
            Pregnancies cluster = 1 ---> Test SkinThickness cluster
                SkinThickness cluster = 1 ---> RESULT = 0
                SkinThickness cluster = 0 ---> RESULT = 1
            Pregnancies cluster = 0 ---> RESULT = 0
```

```
        BMI cluster = 2 ---> RESULT = 0
        BMI cluster = 4 ---> RESULT = 1
        BMI cluster = 0 ---> RESULT = 0
    Age cluster = 1 ---> Test BloodPressure cluster
        BloodPressure cluster = 4 ---> Test BMI cluster
            BMI cluster = 4 ---> Test Pregnancies cluster
                Pregnancies cluster = 1 ---> Test DiabetesPedigreeFunction cluster
                    DiabetesPedigreeFunction cluster = 2 ---> RESULT = 0.0
                    DiabetesPedigreeFunction cluster = 1 ---> RESULT = 0
                    DiabetesPedigreeFunction cluster = 3 ---> RESULT = 0
                Pregnancies cluster = 2 ---> Test Insulin cluster
                    Insulin cluster = 0 ---> RESULT = 1
                    Insulin cluster = 1 ---> RESULT = 0
                    Insulin cluster = 2 ---> RESULT = 1
                Pregnancies cluster = 0 ---> RESULT = 0
            BMI cluster = 3 ---> Test Pregnancies cluster
                Pregnancies cluster = 2 ---> Test DiabetesPedigreeFunction cluster
                    DiabetesPedigreeFunction cluster = 1 ---> RESULT = 0
                    DiabetesPedigreeFunction cluster = 2 ---> RESULT = 1
                Pregnancies cluster = 1 ---> Test DiabetesPedigreeFunction cluster
                    DiabetesPedigreeFunction cluster = 2 ---> RESULT = 0
                    DiabetesPedigreeFunction cluster = 1 ---> RESULT = 0
                Pregnancies cluster = 0 ---> Test SkinThickness cluster
                    SkinThickness cluster = 1 ---> RESULT = 0
                    SkinThickness cluster = 0 ---> RESULT = 0
                    SkinThickness cluster = 2 ---> RESULT = 0
                Pregnancies cluster = 3 ---> RESULT = 1
            BMI cluster = 5 ---> RESULT = 1
        BloodPressure cluster = 3 ---> Test Pregnancies cluster
            Pregnancies cluster = 1 ---> Test Insulin cluster
                Insulin cluster = 0 ---> Test SkinThickness cluster
                    SkinThickness cluster = 2 ---> RESULT = 0
                    SkinThickness cluster = 0 ---> RESULT = 0
                    SkinThickness cluster = 3 ---> RESULT = 0
                Insulin cluster = 4 ---> RESULT = 1
                Insulin cluster = 1 ---> Test BMI cluster
                    BMI cluster = 3 ---> RESULT = 0
                    BMI cluster = 4 ---> RESULT = 0
                Insulin cluster = 2 ---> Test DiabetesPedigreeFunction cluster
                    DiabetesPedigreeFunction cluster = 1 ---> RESULT = 1
                    DiabetesPedigreeFunction cluster = 4 ---> RESULT = 0
                Insulin cluster = 3 ---> Test SkinThickness cluster
                    SkinThickness cluster = 2 ---> RESULT = 1
                    SkinThickness cluster = 1 ---> RESULT = 0
            Pregnancies cluster = 0 ---> RESULT = 1
            Pregnancies cluster = 2 ---> Test DiabetesPedigreeFunction cluster
                DiabetesPedigreeFunction cluster = 1 ---> Test SkinThickness cluster
                    SkinThickness cluster = 2 ---> RESULT = 0
                    SkinThickness cluster = 0 ---> RESULT = 0
                    SkinThickness cluster = 1 ---> RESULT = 0
                DiabetesPedigreeFunction cluster = 3 ---> RESULT = 1
                DiabetesPedigreeFunction cluster = 2 ---> RESULT = 0
            Pregnancies cluster = 3 ---> RESULT = 1
        BloodPressure cluster = 0 ---> RESULT = 1
        BloodPressure cluster = 5 ---> Test SkinThickness cluster
            SkinThickness cluster = 3 ---> RESULT = 1
            SkinThickness cluster = 2 ---> RESULT = 0
        BloodPressure cluster = 2 ---> Test SkinThickness cluster
            SkinThickness cluster = 2 ---> RESULT = 0
            SkinThickness cluster = 3 ---> RESULT = 1
            SkinThickness cluster = 1 ---> RESULT = 0
    Age cluster = 3 ---> Test SkinThickness cluster
        SkinThickness cluster = 2 ---> Test BMI cluster
            BMI cluster = 3 ---> RESULT = 1
            BMI cluster = 4 ---> Test Pregnancies cluster
                Pregnancies cluster = 4 ---> RESULT = 1
                Pregnancies cluster = 2 ---> RESULT = 1
                Pregnancies cluster = 3 ---> RESULT = 0
            BMI cluster = 2 ---> RESULT = 0
        SkinThickness cluster = 0 ---> Test BMI cluster
            BMI cluster = 3 ---> Test Pregnancies cluster
```

```
                Pregnancies cluster = 3 ---> Test DiabetesPedigreeFunction cluster
                    DiabetesPedigreeFunction cluster = 1 ---> RESULT = 0
                    DiabetesPedigreeFunction cluster = 3 ---> RESULT = 0
                Pregnancies cluster = 1 ---> RESULT = 1
                Pregnancies cluster = 2 ---> Test BloodPressure cluster
                    BloodPressure cluster = 4 ---> RESULT = 1
                    BloodPressure cluster = 3 ---> RESULT = 0
            BMI cluster = 0 ---> RESULT = 1
            BMI cluster = 4 ---> Test Pregnancies cluster
                Pregnancies cluster = 4 ---> RESULT = 0
                Pregnancies cluster = 3 ---> RESULT = 0
                Pregnancies cluster = 2 ---> RESULT = 0
                Pregnancies cluster = 0 ---> RESULT = 1
            BMI cluster = 2 ---> RESULT = 0
        SkinThickness cluster = 3 ---> RESULT = 1
        SkinThickness cluster = 1 ---> RESULT = 0
    Age cluster = 2 ---> Test Pregnancies cluster
        Pregnancies cluster = 3 ---> Test SkinThickness cluster
            SkinThickness cluster = 0 ---> Test BloodPressure cluster
                BloodPressure cluster = 3 ---> Test DiabetesPedigreeFunction cluster
                    DiabetesPedigreeFunction cluster = 1 ---> RESULT = 0
                    DiabetesPedigreeFunction cluster = 2 ---> RESULT = 1
                BloodPressure cluster = 4 ---> Test BMI cluster
                    BMI cluster = 3 ---> RESULT = 1
                    BMI cluster = 4 ---> RESULT = 0
                BloodPressure cluster = 0 ---> RESULT = 1
            SkinThickness cluster = 2 ---> Test Insulin cluster
                Insulin cluster = 3 ---> Test BloodPressure cluster
                    BloodPressure cluster = 3 ---> RESULT = 0
                    BloodPressure cluster = 4 ---> RESULT = 1
                Insulin cluster = 0 ---> RESULT = 1
                Insulin cluster = 1 ---> Test BloodPressure cluster
                    BloodPressure cluster = 3 ---> RESULT = 1
                    BloodPressure cluster = 4 ---> RESULT = 0
                Insulin cluster = 2 ---> Test BloodPressure cluster
                    BloodPressure cluster = 3 ---> RESULT = 0
                    BloodPressure cluster = 4 ---> RESULT = 1
            SkinThickness cluster = 1 ---> RESULT = 0
            SkinThickness cluster = 3 ---> Test Insulin cluster
                Insulin cluster = 0 ---> RESULT = 0
                Insulin cluster = 1 ---> RESULT = 1
                Insulin cluster = 3 ---> RESULT = 0
        Pregnancies cluster = 4 ---> Test BloodPressure cluster
            BloodPressure cluster = 0 ---> RESULT = 1
            BloodPressure cluster = 3 ---> RESULT = 1
            BloodPressure cluster = 4 ---> Test Insulin cluster
                Insulin cluster = 1 ---> RESULT = 1
                Insulin cluster = 0 ---> Test DiabetesPedigreeFunction cluster
                    DiabetesPedigreeFunction cluster = 1 ---> RESULT = 0
                    DiabetesPedigreeFunction cluster = 2 ---> RESULT = 1
            BloodPressure cluster = 5 ---> RESULT = 1
        Pregnancies cluster = 0 ---> Test BloodPressure cluster
            BloodPressure cluster = 4 ---> RESULT = 0
            BloodPressure cluster = 3 ---> RESULT = 1
        Pregnancies cluster = 5 ---> RESULT = 1
        Pregnancies cluster = 2 ---> Test Insulin cluster
            Insulin cluster = 2 ---> RESULT = 0
            Insulin cluster = 1 ---> Test SkinThickness cluster
                SkinThickness cluster = 1 ---> RESULT = 1
                SkinThickness cluster = 2 ---> RESULT = 0
                SkinThickness cluster = 3 ---> RESULT = 1
            Insulin cluster = 0 ---> Test BMI cluster
                BMI cluster = 3 ---> Test BloodPressure cluster
                    BloodPressure cluster = 4 ---> RESULT = 0
                    BloodPressure cluster = 3 ---> RESULT = 0
                BMI cluster = 4 ---> RESULT = 1
            Insulin cluster = 4 ---> RESULT = 0
        Pregnancies cluster = 1 ---> Test SkinThickness cluster
            SkinThickness cluster = 2 ---> RESULT = 1
            SkinThickness cluster = 3 ---> RESULT = 0
            SkinThickness cluster = 0 ---> RESULT = 0
```



همانطور که دیده میشود دقت روی داده های train ۹۲ درصد است و روی داده های تست 65 درصد، با کاهش داده های test و افزایش داده های تست به ۰.۵ و ۰.۵ دقت زیاد شد که خلاف انتظار من بود، من انتظار داشتم با کاهش داده train عملکرد در داده های test پایین بیاید که این اتفاق رخ نداد و دقت برای train به ۹۵ رسید و برای تست به 70! که نشان از overfitting است.
همچنین با افزایش محدودیت عمق overfitting کمتر رخ میداد اما زمان اجرای کد طولانی تر میشد.

آموخته های من از این تمرین:
اولا با پیاده سازی شبه کد های آنتروپی و محاسبه B و remainder کمی نسبت به آنها احساس ترس کمتری پیدا کردم، متاسفانه خیلی باگ های زیادی خوردم و بیشتر آن هم به این دلیل بود که سعی داشتم کدم را کوتاه بنویسم پ با خیلی از روش هایی که استفاده کردم دوستی زیادی نداشتم و این منجر به طولانی شدن فرایند کد زدن شد.

در کل نسبت به سختی این مسئله و دلیل اینکه بجای این روش از روش های دیگر استفاده میشود و عدم کارایی زیاد آن آگاهی پیدا کردم.
برای افزایش دقت حس میکنم افزایش یا برداشتن محدودیت ارتفاع درخت ایده خوبی باشد، همچنین روش های هوشمندانه تر برای گسسته سازی داده قطعا راهگشا خواهند بود. در این لینک اینطور ایده ها آمده است که من آنها البته پیاده سازی نکردم اما برای بسیار جذاب بودند.

حس میکنم با کمی fine tuning میشد دقت را بیشتر کرد اما خب این به تست های زیاد نیاز دارد.
در کل با انجام تمرین حس میکنم فهم بهتری نسبت به گسسته سازی داده ها و محاسبه آنتروپی و ... پیدا کردم.
در انتها باید بگویم که بسیاری از ایده های پیاده سازی را از این لینک گرفتم، در واقع در ابتدا خودم بخشی از کد را زدم اما چون نمیدانستم مسیر را چگونه باید ادامه دهم کمی گشتم و یک مثال دیدم و سپس آنرا سعی کردم خودم پیاده سازی کنم که به باگ های بسیار بسیار زیادی برخوردم، خواهشا این را درنظر بگیرید کدی کپی نکردم و فقط ایده های آنها را دیدم تا بفهمم باید چگونه کلاستر بندی و ... را انجام دهم.
همچنین از sklearn فقط برای تقسیم کردن دیتاست استفاده کردم و از آن استفاده دیگری نکردم!