قسمت دوم پروژه پایانی :

بخش اول) سوالات تئوری (۱ نمره)

1 - انواع روش‌های خوشه‌بندی را نام ببرید و هر یک را به صورت مختصر توضیح دهید.

2 - توضیح دهید منظور از dendrogram چیست و چه کاربردی دارد؟

3 - انواع معیارهای شباهت برای ادغام دو خوشه را نام برید و هر یک را مختصراً توضیح دهید.

4 - دو مورد از روش‌های ارزیابی خوشه‌بندی را نام ببرید و هر یک را مختصراً توضیح دهید.

بخش دوم) تمرین عملی (حداکثر ۳ نمره)

یکی از مشکلاتی که در کشورهای پیشرفته وجود دارد، معضل چاقی بیش از حد است. افراد بسیاری وجود دارند که از نظر تغذیه روزانه، مراقبت کافی را انجام می‌دهند ولی باز هم به دلیل فاکتورهای ژنتیکی به این معضل مبتلا می‌شوند. در یک پژوهش تلاش شده است که ژن‌های مؤثر در چاقی شناسایی شوند. از آنجا که انجام آزمایشات کنترل شده، روی مدل حیوانی بسیار ساده‌تر از انسان است (زیرا حیوان در آزمایشگاه تحت نظر و کنترل قرار می‌گیرد و تعداد نمونه‌های مورد آزمایش در شرایط کنترل شده را در این صورت می‌توان افزایش داد)، و همچنین به دلیل قرابت زیستی موش و انسان، این پژوهش بر روی گروهی از موش‌ها انجام شده است.

داده‌های این پژوهش، شامل دو گروه از موش‌های نر (۱۲۴ نمونه) ذخیره شده در فایل male_data.csv، و ماده (۱۳۵ نمونه) ذخیره شده در فایل female_data.csv است. در هر یک از نمونه‌ها مقادیر ۳۶۰۰ ژن مختلف گزارش شده است. در فایل، شناسه موش‌ها با عبارت F۲ و شناسه ژن‌ها با حروف MMT شروع می‌شود.

(به دلایل زیستی، شما باید دو فایل را جداگانه تحلیل کنید و داده‌ها نباید ترکیب شوند.)

* از شما بعنوان یک تحلیل‌گر داده خواسته شده است که در ابتدا این داده‌ها را پیش‌پردازش کنید. با ذکر دلیل، توضیح دهید چه کارهایی برای پیش‌پردازش انجام داده‌اید و کد را به همراه نتیجه گزارش کنید. (هر مرحله از پیش‌پردازش باید هدفمند و با ذکر دلیل باشد. به صورت تصادفی از همه روش‌ها استفاده نکنید).

* هر مجموعه داده را به صورت مستقل، خوشه بندی کنید. چند خوشه در این داده‌ها تشخیص می‌دهید؟ برای هر مجموعه داده به صورت جداگانه عملیات را انجام دهید و نتیجه را گزارش دهید.

* خوشه بندی انجام شده برای هر مجموعه داده را نمایش دهید.

* برای محققان مهم است که تمام داده‌ها قابل اعتماد باشند. معمولاً در این نوع داده، اگر خطایی در آزمایشگاه رخ داده باشد، آن داده به صورت داده‌ی پرت (outlier) قابل تشخیص است. در هر یک از دو مجموعه داده بررسی کنید که آیا داده پرت وجود دارد یا خیر؟ در صورت وجود آن را حذف کنید.

(نکته بسیار مهم)

** ابتدا صورت مساله را به طور کامل مطالعه کنید، و درباره راهکار تمامی قسمت‌ها فکر کنید. در این صورت خواهید توانست به صورت بهینه (و بدون اضافه کاری) و با صرف کمترین زمان، پاسخ سوالات را بدهید.

** پیاده سازی شما باید حتماً به زبان پایتون باشد. اگر به‌جای آن از weka استفاده نمایید در صورت کامل بودن هم فقط نیمی از نمره را خواهید گرفت. پس در صورتی که به جواب نرسیدید از weka استفاده نمایید.

(به طور خاص پکیج scikitlearn میتوانید استفاده کنید).