به نام خدا

پروژه درس داده‌کاوی: فصل ۸، خوشه‌بندی

استاد درس: دکتر رضا رمضانی

مهلت تحویل: ۹ تیر ۱۴۰۰

پروژه بصورت ۱ نفره یا ۲ نفره است

در این پروژه، یک دیتاست شامل اطلاعات حدود ۲۰۰ ستاره و ویژگی‌های آن‌ها در اختیار شما قرار گرفته است. هدف، خوشه‌بندی کردن این ستاره‌ها در ۶ دسته است. همچنین کلاس و دسته‌ی صحیح هر ستاره در ستون type با اعداد ۰ تا ۵ مشخص شده است که فعلا برای فرآیند خوشه‌بندی با این ستون کاری نداریم اما هنگام ارزیابی برای ما مهم است. ۴ ستون اول مقادیر عددی پیوسته هستند که به ترتیب دما، درخشش، شعاع و قدر مطلق ستاره است. دو ستون بعدی، رنگ و نوع خاص ستاره است که هر دو کمیت‌هایی nominal هستند. حال می‌خواهیم این داده‌ها را بدون استفاده و توجه به ستون type خوشه‌بندی نماییم. مراحلی که در ادامه مشخص شده‌اند را انجام دهید و نتایج کار را به صورت کامل مستند کنید. استفاده از کتابخانه‌هایی مانند scikit-learn مجاز است.

۱. ابتدا داده‌ها را بدون هیچگونه پیش پردازشی و بدون استفاده از داده‌های ستون type خوشه‌بندی کنید. خوشه‌بندی را به کمک یک الگوریتم Partitional و یک الگوریتم Hierarchical انجام داده و تعداد خوشه را ۶ در نظر بگیرید. در نهایت دقت خوشه‌بندی خود را حساب کنید.

> **نحوه محاسبه دقت خوشه‌بندی:** راه‌های زیادی برای ارزیابی دقت خوشه‌بندی وجود دارد. برای این مساله، بعد از خوشه‌بندی، ببینید برچسب (type) اکثریت داده‌های موجود در یک خوشه چیست. سپس برچسب تمام داده‌های آن خوشه را همین برچسب اکثریت در نظر بگیرید. این کار را برای تمام خوشه‌ها انجام دهید. لذا تمام داده‌های خوشه‌بندی شده، یک برچسب خوشه دارد. در نهایت ببینید چند درصد از کل داده‌ها برچسب صحیح دریافت نموده‌اند.

۲. در گام بعدی داده‌ها را پیش پردازش کنید (مانند گسسته سازی، مقیاس‌بندی و ... - با مکتوب کردن دقیق پیش پردازش) و مجددا خوشه‌بندی بیان شده در مرحله ۱ را انجام دهید و دقت را حساب کنید. بایستی سعی کنید پیش پردازش را به نحوی انجام دهید که دقت خوشه‌بندی افزایش یابد.