**به نام خدا**

پروژه درس داده‌کاوی: طبقه‌بندی (مقدماتی و پیشرفته)

استاد درس: دکتر رضا رمضانی

مسئول پروژه: علی عابدزاده (@abedzadeh\_ali)

مهلت تحویل: چهارشنبه – 11 خرداد

در این پروژه، یک دیتاست با حدود 700 سطر داده، در اختیار شما قرار داده شده است. در این دیتاست، ویژگی‌های مختلفی از ساختمان‌های مختلف ارائه شده است که شما باید از این ویژگی‌ها برای پیش‌بینی و طبقه‌بندی دو ستون آخر (heating/cooling load) استفاده کنید. مراحل خواسته شده‌ای که در ادامه ذکر خواهد شد را انجام دهید و مراحل انجام کار و نتایج آن و همینطور توضیحات لازمه را ثبت کنید. لازم است تا کدها به همراه یک فایل pdf (شامل تحلیل داده‌ها، نتایج و …) ارسال شود و یا می‌تواند به جای آن، یک فایل Notebook پایتون ارسال کنید که نتایج و توضیحات شما در آن وجود دارد. استفاده از کتابخانه‌های موجود مجاز است. پروژه به صورت انفرادی است.

- در ابتدا لازم است تا اطلاعات کلی در مورد ستون‌ها ارائه کنید. در صورت لزوم، داده‌ها را تغییر دهید (گسسته‌سازی، نرمال‌سازی و …). همینطور شما می‌توانید با ارائه‌ی دلایل خود، در هر یک از مراحل پیش رو، یک یا چند ستون را حذف یا ادغام کنید تا بتوانید ویژگی‌های مرتبط‌تری را استفاده کنید.

از دو ستون آخر به عنوان صفت‌های اولیه برای آموزش مدل استفاده نکنید. این دو صفت کلاس برچسب هستند.

**- پیش‌بینی heating load**

* برای پیش‌بینی ستون heating load، لازم است تا یک Decision Tree و یک Random Forest را آموزش دهید. انتخاب پارامتر‌های این طبقه‌بندها به عهده‌ی شما است. این طبقه‌بندها موظف هستند تا بالا یا پایین بودن میزان بار گرمایش را پیش‌بینی کنند. مقدار بار بالاتر یا برابر 18 را «بار بالا» و مقادیر کمتر از 18 را «بار پایین» در نظر بگیرید. پس از آموزش، Confusion matrix پیش‌بینی‌های heating load دو مدل را برای دیتاست test بدست آورید.
* با توجه به Confusion matrix به دست آمده، مقادیر زیر را محاسبه کنید:

accuracy, error rate, sensitivity, specificity, precision, recall, F-score

* کدام‌یک از این دو مدل، پیش‌بینی‌های بهتر و دقیق‌تری کرده است؟

**- پیش‌بینی cooling load**

* برای پیش‌بینی ستون cooling load، لازم است تا یک SVM و یک شبکه‌ی عصبی چندلایه با معماری دلخواه را آموزش دهید. انتخاب پارامتر‌های این طبقه‌بندها به عهده‌ی شما است. این طبقه‌بندها موظف هستند تا میزان بار سرمایش را در ۵ سطح پیش‌بینی کنند. برای تبدیل کردن ستون cooling load به 5 سطح، مینیمم و ماکسیمم این ستون را بدست آورید و این بازه را به 5 بازه‌ی جدید و هم‌اندازه تبدیل کنید. سپس هر مقدار از این ستون را در یکی از این بازه‌های جدید قرار دهید.
* مقادیر زیر را برای پیش‌بینی‌های cooling load دو مدل برای دیتاست test به دست آورید:

accuracy, precision, recall, F-score

* کدام‌یک از این دو مدل، پیش‌بینی‌های بهتر و دقیق‌تری کرده است؟