با سلام.

مجموعه داده ای که پروژه پایانی درس باید روی آن انجام شود را از لینک زیر میتوانید ببینید و دانلود کنید.

https://archive.ics.uci.edu/ml/datasets/Record+Linkage+Comparison+Patterns

لینک مستقیم دانلود داده ها:

https://archive.ics.uci.edu/ml/machine-learning-databases/00210/donation.zip

لینک مستقیم دانلود توضیحات مربوط به داده ها:

https://archive.ics.uci.edu/ml/machine-learning-databases/00210/documentation

فایلی که دانلود میکنید از یک سری فایل csv تشکیل شده که نام آنها شامل "block" هست. برای تحلیلی که انجام میدهید، باید با استفاده از اسپارک این فایل ها رو لود و تجمیع کنید. بعد از تجمیع داده ها، پیش پردازش های لازم را انجام دهید تا مجموعه داده آماده پیاده سازی الگوریتم های داده کاوی/یادگیری ماشین (طبقه بندی) شود. فاز پیش پردازش نیز باید در محیط اسپارک صورت گیرد. بعد از اینکار با استفاده از یک یا چند روش از روش های طبقه بندی که اسپارک ارائه داده، باید عملیات طبقه بندی (Classification) را روی این مجموعه داده پیاده سازی کنید. در نهایت باید ارزیابی مدل های ایجاد شده را با شاخص های accuracy, f-measure (f1) , ROC. انحام دهید.

Dataset Summary

Record Linkage Comparison Patterns

https://archive.ics.uci.edu/ml/datasets/Record+Linkage+Comparison+Patterns

Number of Instances: 5,749,132

Number of Attributes: 12

Size: 250 MB

Attribute Characteristics: Real

Area: Registry Data Patterns

در نهایت برای ارائه گزارش، کل فرایند طی شده در قالب متدولوژی CRISP را به صورت یک گزارش در قالب فازهای این متدولوژی ارائه خواهید داد. فایل گزارش بهمراه فایل نوت بوک که پروژه در ان انجام شده و نتایج قابل مشاهده هست را ارسال کنید. این حداقل کاریست که باید انجام شود. به گروهی که بهترین نتایج را بدست بیاورند نیز پون مثبت اضافه تعلق خواهد گرفت. موضوع Over-fitting را نیز در نظر داشته باشید. میتوانید روش های مختلفی با تنظیم پارامتر های مختلف را برای پیاده سازی طبقه بندی استفاده کنید تا به بهترین نتایج برسید. این اختیار را دارید که مجموعه داده و ویژگی های آن را به هر روشی پیش پردازش و پردازش کنید ولی نباید از چارچوب اسپارک خارج شوید.

مهلت انجام پروژه پایانی تا پایان وقت شنبه ۲۲ بهمن است. پروژه را به صورت گروه های ۴ نفره انجام دهید. مشارکت تمام افراد گروه در انجام پروژه الزامی است.

موفق باشید