تکالیف درس داده کاوی

۱- **مسئله دسته بندی نودها در گراف دیتاست obgn-product**:

- ساختار گراف :یک گراف بدون جهت و بدون وزن که نمایانگر شبکهٔ خرید مشترک محصولات در Amazon است. شامل حدود ۲ میلیون گره و ۶۰ میلیون یال
- گره‌ها :(Nodes) هر گره نمایانگر یک محصول است.
- یال‌ها :(Edges) وجود یال بین دو گره نشان‌دهندهٔ خرید همزمان آن دو محصول توسط کاربران است.
- ویژگی‌های گره‌ها :برای هر گره، یک بردار ویژگی ۱۰۰بعدی استخراج‌شده از توضیحات متنی محصول با استفاده از روش bag-of-words و کاهش ابعاد با PCA فراهم شده است.
- برچسب‌های هر نود :(Labels) هر محصول به یکی از ۴۷ دسته‌بندی سطح بالا اختصاص داده شده است.
- لینک توضیحات دیتاست:

https://ogb.stanford.edu/docs/nodeprop/#ogbn-products

۱.آماده‌سازی داده‌ها

- بارگذاری دیتاست با استفاده از PygNodePropPredDatasetاز کتابخانهٔ ogb.

۲.مدل‌سازی باGNN

- پیاده‌سازی مدل‌های مختلف مانند GCN و GraphSAGE برای طبقه‌بندی گره‌ها.

۳.ارزیابی مدل

- استفاده از معیارهایی مانند دقت (Accuracy) و F1-Score برای ارزیابی عملکرد مدل‌ها.
- مقایسه نتایج مدل‌های مختلف و تحلیل عملکرد آن‌ها در دسته‌بندی‌های مختلف.
- + موارد اضافی: حل مسئله edge prediction روی همین دیتاست.

۲- **مسئله دسته بندی با پکیج spark** :

حل مسئله دسته‌بندی به صورت لوکال توسط پکیج spark و ارائه گزارش کامل نتایج معیارهای کارآیی روی دیتاست زیر:

https://www.kaggle.com/datasets/ronitf/heart-disease-uci

+ موارد اضافی: در نظر گرفتن چند ماشین مجازی روی یک سیستم و اجرای کد در مد کلاستر

نکات:

- تکالیف به صورت تک نفره انجام شوند، در صورتیکه می‌خواهید پروژه را به صورت گروهی انجام دهید، باید موارد اضافی هم پیاده‌سازی شوند.
- یک گزارش مختصر یک صفحه‌ای از مشخصات سخت‌افزار اجرایی، خلاصه معماری نهایی مدل و دقت‌های نهایی به همراه اصل کدها می‌بایست ارائه شوند.
- گزارش خود کد یا به صورت یک ویدیوی نهایتا ۱۰ دقیقه‌ای و یا به صورت یک فایل pdf مفصل حداکثر ۱۰ صفحه‌ای ارائه گردد.
- ارسال تکلیف‌ها تا ساعت ۲۴ روز ۶ تیرماه فرصت دارد. بعد از این زمان، هیچ تکلیفی تحویل گرفته نخواهد شد.
- برای پیشگیری از مشکلات دریافت ایمیل، هر دو تکلیف در قالب یک ایمیل به ایمیل drmohamadkiani@gmail.com ارسال شوند. الزاما عنوان ایمیل (تکالیف نهایی درس داده کاوی+ نام و نام خانوادگی + شماره دانشجویی) باشد. ایمیل های جداگانه یا با عناوین دیگر، تحویل گرفته نمی شود.