راهنمای پروژه دوره علم داده – مقدماتی

مدرس: دکتر فرزاد مینویی

اهداف پروژه:

انتظار این است دانشجویان با تکمیل پروژه خود، اهداف زیر را محقق کنند:

- رویکرد CRISP-DM را برای انجام روشمند یک پروژه داده در عمل بکار ببرند.
- بتوانند رویکردهای مختلف یادگیری ماشین را که در طول دوره آموختند، بر روی داده های واقعی پیاده سازی کنند.
- یکی از زبان های برنامه نویسی R یا پایتون را برای تحلیل داده‌ها بکار بگیرند.

موضوعات پروژه:

دانشجویان می توانند برای پروژه پایانی خود یکی از موضوعات پیشنهادی زیر را انتخاب کنند:

- پروژه ۱:

فایل college.csv حاوی اطلاعات مربوط به تعداد درخواست های متقاضیان ورود به کالج های آمریکا در سال ۱۹۹۵ است. هدف پیش بینی متغیر Apps براساس سایر متغیرهای موجود است. تعاریف متغیرها در فایل college_data_description آمده است.

در صورت انتخاب این پروژه، آن را تنها می توانید به صورت انفرادی انجام دهید.

- پروژه ۲:

این پروژه مربوط به یکی از مسابقات معروف kaggle است. برای دسترسی به داده های این پروژه، عضو سایت kaggle شوید و به فایل های موجود دسترسی پیدا کنید. لینک مسابقه در زیر آمده است:

https://www.kaggle.com/c/house-prices-advanced-regression-techniques

در صورت انتخاب این پروژه، می‌توانید آن را به صورت انفرادی یا گروهی (۲ نفره) انجام دهید.

- پروژه ۳:

فایل weekly.csv حاوی اطلاعات مربوط نرخ بازدهی هفتگی S&P 500 بین سال های ۱۹۹۰ و ۲۰۱۰ است. هدف پیش بینی متغیر Direction براساس سایر متغیرهای موجود است. تعاریف متغیرها در فایل weekly_data_description آمده است.

در صورت انتخاب این پروژه، آن را تنها می توانید به صورت انفرادی انجام دهید.

خروجی مورد انتظار:

انتظار این است دانشجویان خروجی پروژه را در قالب فایل html تحویل دهند که در آن کدها به همراه خروجی آنها، نمودارها، توضیحات و نتیجه‌گیری‌ها آمده باشد. برای این منظور می توانید از امکانات RMarkdown و یا Jupyter Notebook استفاده کنید.

انتظار این است براساس مدل CRISP-DM خروجی پروژه شما از بخش‌های زیر تشکیل شده باشد و به این سوالات پاسخ داده شده باشد:

فهم مسئله:

۱. انگیزه اصلی چنین پروژه ای چیست؟
۲. خروجی چنین پروژه ای برای چه مواردی ممکن است کاربرد داشته باشد؟
۳. چه کسانی ممکن است به نتایج این پروژه علاقمند باشند؟ چرا؟

درک داده:

۱. داده ها از کجا بدست آمده اند و چگونه جمع آوری شده اند؟
۲. هر یک از متغیرها چه چیزی را اندازه گیری می کنند؟
۳. آیا ابهامی در تعاریف داده ها وجود دارد؟
۴. آیا ممکن است در اندازه گیری متغیرها و یا ثبت داده‌ها خطایی وجود داشته باشد؟
۵. چه متغیرهای دیگری اگر وجود داشتند، می توانست به حل مسئله کمک کند؟
۶. متغیرهای موجود از کدام نوعند (رسته ای – عددی)؟
۷. خلاصه آماری متغیرهای موجود چیست؟

آماده‌سازی داده

۱. آیا نیاز به درآمیختن داده‌ها است؟ اقدامات و نتایج گزارش شود.
۲. آیا نیاز به پاک سازی داده است؟ اقدامات و نتایج گزارش شود.
۳. آیا نیاز به تبدیل داده است؟ اقدامات و نتایج گزارش شود.
۴. آیا نیاز به کاهش داده است؟ اقدامات و نتایج گزارش شود.

مدل‌سازی:

۱. روی داده‌های آموزش، حداقل ۶ مدل برای پیش بینی متغیر پاسخ ساخته شود. اقدامات و نتایج گزارش شود.
۲. با استفاده از شاخص‌های مطرح شده در کلاس، خروجی‌های مدل ها باهم مقایسه گردد. اقدامات و نتایج گزارش شود.

ارزیابی:

۱. مدل‌های ارائه شده، روی داده های آزمایش با استفاده از شاخص‌های متداول در یادگیری ماشین ارزیابی شوند. اقدامات و نتایج گزارش شود.
۲. چه پیشنهاداتی دارید تا نتایج در محیط واقعی، آزمایش گردد؟

استقرار:

حال اگر بخواهید چنین الگوریتمی را در مقیاس صنعتی توسعه دهید، به این فکر کنید با چه چالش هایی مواجه خواهید شد و برای آن چه راهکارهایی دارید. موارد زیر را گزارش کنید:

۱. چالش‌های توسعه الگوریتم را بررسی کنید.
۲. چه راهکارهایی برای حل آن ها دارید؟
۳. چه ملزوماتی برای ارائه آن راهکارها نیاز دارید؟

نتیجه گیری:

۱. انجام این پروژه چه یادگیری برای شما داشت؟
۲. با چه چالش هایی مواجه شدید؟ چگونه آن ها را حل کردید؟

ارزیابی پروژه:

کسانی که به بیش از ۷۰ درصد موارد بالا پاسخ صحیح داده باشند، نمره قبولی پروژه را دریافت خواهند کرد.

موفق باشید.

فرزاد مینویی