

## 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 داده‌های واقعی پیاده‌سازی کنند.
- زبان برنامه نویسی پایتون را برای تحلیل داده و بکارگیری الگوریتم‌های یادگیری ماشین بکار بگیرند.

### موضوعات پروژه:

فایل داده‌ها و توضیحات مربوطه در پورشه SProject در گوگل درایو کلاس موجود است. دانشجویان می‌توانند برای پروژه پایانی خود یکی از موضوعات پیشنهادی زیر را انتخاب کنند:

- پروژه ۱:

فایل college.csv حاوی اطلاعات مربوط به تعداد درخواست‌های متقاضیان ورود به کالج‌های آمریکا در سال ۱۹۹۵ است. هدف پیش‌بینی متغیر Apps براساس سایر متغیرهای موجود است. تعاریف متغیرها در فایل college\_data\_description آمده است. در صورت انتخاب این پروژه، آن را تنها می‌توانید به صورت انفرادی انجام دهید.

- پروژه ۲:

این پروژه مربوط به یکی از مسابقات معروف kaggle است. لینک مسابقه در زیر آمده است:

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

فایل kaggle\_dataset.zip حاوی داده‌های موردنیاز برای این پروژه است.

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 به نرخ بازدهی هفتگی شاخص S&P 500 بین سال‌های ۱۹۹۰ و ۲۰۱۰ است. هدف پیش‌بینی متغیر Direction براساس سایر متغیرهای موجود است. تعاریف متغیرها در فایل weekly\_data\_description آمده است.

در صورت انتخاب این پروژه، آن را تنها می‌توانید 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 نتیجه گیری ها آمده باشد. برای این منظور می توانید از امکانات Jupyter Notebook استفاده کنید.

انتظار این است براساس [رویکرد CRISP-DM](#) خروجی پروژه شما از بخش های زیر تشکیل 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 داده ها 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؟

### آماده سازی داده

۱. آیا نیاز به درآمیختن داده ها 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.

### مدل سازی:

۱. روی داده های آموزش، حداقل ۶ مدل برای پیش بینی متغیر پاسخ ساخته شود. اقدامات و نتایج گزارش شود.
۲. در صورت نیاز فرآیندهای مدل را تنظیم و نتایج اقدامات را گزارش کنید.

## ارزیابی:

۱. مدل‌های ارائه شده، روی داده‌های 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 چالش‌هایی 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 آن‌ها دارید؟
۳. چه ملزوماتی برای ارائه این راهکارها نیاز دارید؟
۴. (اختیاری): پروژه خود را در قالب یک Web Application توسعه دهید.

## نتیجه‌گیری:

۱. انجام این پروژه چه یادگیری برای شما داشت؟
۲. با چه چالش‌هایی مواجه شدید؟ چگونه آن‌ها 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