به نام خدا

راهنمای انجام و نگارش پروژه پایانی درس داده کاوی

گزارش پروژه باید در قالب CRISP-DM ارائه شود. در ادامه در مورد هر یک از گام‌ها توضیح داده شده است.

۱- شناخت کسب و کار

این گام شامل ۴ بخش است:

جمله اول در مورد هدف و ماهیت کسب و کار مورد مطالعه است. جمله دوم در مورد چالش پیش روی کسب و کار است. جمله بعدی در مورد سؤالی است که قصد دارید آن را با استفاده از داده کاوی پاسخ دهید و جمله آخر در مورد اینکه پاسخ این سئوال چگونه به حل چالش پیش روی کسب و کار مورد مطالعه کمک می‌کند.

۲- شناخت داده‌ها

در این قسمت باید ذکر کنید که داده شما از سازمان واقعی تهیه شده یا از اینترنت دانلود شده است. در صورتی که از سازمان تهیه شده نام سازمان و آدرس وب سایت سازمان و در صورتی که از اینترنت دانلود شده باید لینک دانلود داده را ذکر کنید. علاوه بر این باید یک جدول تهیه کنید که ستون‌های آن شماره ستون داده در فایل داده‌ها، عنوان ستون داده، نوع داده، نقش داده (اینکه داده ویژگی[1] است یا برچسب[2]، بازه مجاز برای مقدار داده و یک توضیح کوتاه درباره ماهیت داده باشد.

۳- آماده سازی داده‌ها

این قسمت یکی از وقتگیرترین بخش‌های پروژه شماست. برای این قسمت باید در صورت لزوم داده‌ها را با استفاده از روش مناسب نرمال‌سازی کنید، داده‌های پرت و غلط و داده‌های ناقص[3] را اصلاح کنید، داده‌های رده‌ای[4] را کد نمایید و ... و در گزارش ذکر نمایید.

٤- مدل‌سازی

در این قسمت باید با استفاده از داده‌های پاکسازی شده مدلسازی را انجام دهید. دقت کنید که برای الگوریتم‌های نظارتی باید حداقل با استفاده از دو الگوریتم مختلف با داده‌های آموزش و تست مدل را ایجاد نمایید (در فایل‌های جداگانه) و خطای هریک را اندازه گیری کنید. در مورد الگوریتم‌های غیر نظارتی باید چندین حالت مختلف را بررسی کنید. مثلا اگر خوشه بندی می‌کنید داده را با تعداد خوشه‌های مختلف خوشه‌بندی نمایید.

---

[1] feature
[2] label
[3] Missing values
[4] categorical

۵- ارزیابی

در این قسمت اگر الگوریتم‌های شما نظارتی است باید خطاهای الگوریتم‌های مختلف را با هم مقایسه کنید (confusion matrix و در صورت نیاز نمودارها را در فایل گزارش ارائه نمایید) و یکی را انتخاب کنید و اگر الگوریتم شما غیر نظارتی است باید طرح به دست آمده از مدل‌سازی را برای حالت‌های مختلف تفسیر کنید (مثلا شبیه تفسیری که در مثال خوشه‌بندی مشتریان فروشگاه داشتیم).

علاوه بر این باید مدل به دست آمده را (در هر دو حالت نظارتی و غیرنظارتی) با پارامترهای مختلف در یک جدول یا نمودار بررسی کنید و پارامترهای با کمترین خطا را انتخاب کنید.

٦- پیاده سازی

در این قسمت باید بررسی کنید که آیا مدل ارائه شده موفق شده است چالش مطرح شده را حل نماید یا خیر.

نکات مهم:

۱- این پروژه ۷ نمره از ارزیابی نهایی شماست.

۲- به پروژه به صورت کلی نمره داده می‌شود و نه بخش بخش.

۳- تاریخ تحویل پروژه ۲ هفته پس از امتحان پایان ترم درس داده‌کاوی است.

۴- در فایل گزارش پروژه توضیحات مربوط به هر گام و در صورت نیاز خروجی‌هایی که در نرم‌افزار به دست آورده‌اید را قرار دهید.

۵- درصفحه اول پروژه نام و نام خانوادگی خود را ذکر کنید.

۶-پروژه به صورت انفرادی می‌باشد.

۷- **فایل گزارش پروژه** (به فرمت pdf) را به همراه فایل/فایل‌های **کد مربوطه** (شامل مراحل مختلف که با کامنت مشخص نموده‌اید: پیش‌پردازش، مدل‌سازی و ارزیابی) در قالب یک فایل فشرده در محل مشخص شده در elearn بارگذاری فرمایید.

۸- پروژه حتماً باید با استفاده از پایتون انجام شود.

برخی از سایت‌هایی که می‌توانید برای دانلود دیتاست استفاده نمایید:

Kaggle.com

KDNugget.com

UCI Data Repository

Google Dataset Search

DataHeart.ir

Data.World

Visualdata.io