# فاز اول پروژه آکادمی لوپ

**اساتید آکادمی لوپ**
آقایان علی شریفی ، آرشام غلام زاده ،
احمدرضا شریفیان زاده، علیرضا جواهری، ابولفضل صفری

۱۲ فروردین ۱۴۰۱

# ۱ تشریح مسئله

یک پروژه در حوزه علم داده ممکن است از بخش های مختلفی تشکیل شده است. مرسوم است که در پروژه های علم داده، فاز صفر(یا فاز۱ ) پروژه را به عنوان **کیفیت داده** [۱] یاد می‌شود.
فاز کیفیت‌داده را می‌توان یکی از فازهای مهم در نظر گرفت چرا که درک درستی از ماهیت داده های موجود به تیم داده و کارفرما میدهد. گاها به خصوص زمانی که کارفرما دید فنی مناسبی نسبت به علم داده نداشته باشد ، خواسته هایی را دارد که با توجه به داده ها موجود و فیچرهای موجود شدنی نیست. فاز کیفیت داده ها معمولا به بررسی فیچرهای موجود در داده ها و تهیه گزارش از فیچرهای موجود به همراه مصورسازی هایی پرداخته می شود. درکنار این بررسی، یافتن داده های missing و داده های پرت ، داده های تکراری پرداخته میشود. معمولا پس از تهیه گزارش از این دست داده ها ، گزارش ابتدایی به کارفرما ارایه میشود که داده های موجود مثل داده های missing و داده های پرت بازبینی شوند و ممکن است در این مرحله کارفرما اطلاع دهد برخی از این فیچرها مهم نیستند و قابل صرف نظر هستند.
از شما خواسته میشود یک گزارش از کیفیت داده بر روی دیتاست موجود تهیه کنید. پس از انجام این مرحله در فاز بعدی از شما خواسته میشود که به بررسی داده های missing و اتخاذ سیاست هایی برای آنها بپردازید که میتوانید در این فاز نیز آنها را انجام دهید.
در فاز بعدی پس از بررسی کارهای این فاز و روند اصلاح داده ها به بحث های پردازش داده ها و مهندسی ویژگی خواهیم پرداخت و در فاز نهایی به مدل سازی بر روی داده ها می پردازیم.
توجه داشته باشید که فاز دو و سه به همدیگر وابسته می باشند و قطعا نیاز است که روند رفت و برگشتی میان فاز ۲ و فاز ۳ طی شود.

# ۲ دیتاست

توجه کنید شما میتوانید بر روی ککل یا کولب و یا کامپیوتر های شخصی خود کار کنید .
به جای دانلود و آپلود دیتاست در گوگل درایو برای استفاده در کولب میتوانید به شیوه زیر عمل کنید .

چگونه از دیتاست های ککل در کولب استفاده کنیم ؟

دیتاست مورد استفاده در این چالش ، یکی از دیتاست های معروف سایت kaggle می باشد که تغییرات در آنها صورت گرفته است و دیگر دیتاست اصلی نمی باشد.
شما میتوانید از طریق آدرس زیر به توضیحات دیتاست از قبیل فیچرها دسترسی داشته باشید.
(توجه داشته باشید در پروژه های دنیای واقعی خیلی اتفاق می افتد که شما هیچ گونه اطلاعاتی از فیچرها ندارید یعنی هیچ گونه داکیومنتی در اختیار شما قرار نمیگیرد و شما خود باید ارتباطات

[1] **Data Quality**

را پیدا کنید و برای فیچرها فرضیه سازی کنید و در نهایت فرضیات خود در خصوص هر فیچر را از کارفرما بپرسید که گاها به نتیجه ای هم نخواهید رسید.)

توضیحات دیتاست

لینک زیر دیتاستی است که شما باید بر روی آن کار کنید.

لینک دیتاست پروژه

# ۳ نحوه ارسال ها

جهت دریافت کارها در هر فاز لازم است که کدها به همراه گزارش هر فاز در ریپازیتوری گفته شده قرار گیرد.

ابتدا ریپازیتوری اصلی را fork کنید سپس بر روی ریپازیتوری fork شده کارهای خود را انجام دهید و در آخر جهت ادغام pull request کنید.

برای این فاز زمان پیشنهادی آکادمی لوپ ۴ روز می باشد. توجه داشته باشید که تمامی کارها در آخر از شما تحویل گرفته میشود و ورژن نهایی کارها در آخر مورد ارزیابی قرار گرفته میشود اما در هرفاز منتورها شما را راهنمایی و کدها را بررسی و نکات لازم را به شما اطلاع میدهند.

مدیریت زمان کل پروژه بر عهده شما می باشد اما منتورها زمان پیشنهادی را ارایه میدهند که در صورت رعایت این زمان بندی انتظار میرود نتیجه نهایی مطلوب تر باشد.