آمار و احتمالات مهندسی

پروژه نهایی - R

حسام و شهرزاد

سؤال ۱.

در این پروژه، ما قصد داریم مجموعه‌ای از داده‌های واقعی را با آن‌چه در این درس آموخته‌اید بررسی و تحلیل کنیم. برای شروع تجزیه و تحلیل یک مجموعه داده، اولین قدم آشنایی با آن است. در اولین قدم می‌توان با مشاهده مواردی مثل ویژگی‌های مجموعه داده و توزیع مقادیر و تجسم داده‌ها برای حدس زدن اولیه در مورد آن آشنایی را انجام داد. در مرحله‌ی بعدی با انجام آزمایشات آماری، اطمینان حاصل می‌کنیم که حدس‌هایمان درست است و ادعاهای خود را با اطمینان بیان می‌کنیم.

برای این سوال از دیتاست AdmissionPredict و زبان R استفاده کنید. (در این مجموعه داده باید از ستون‌های:

- نمره‌ی آزمون GRE (از ۳۴۰ نمره)،
- نمره‌ی آزمون TOEFL (از ۱۲۰ نمره)،
- رتبه‌بندی دانشگاه مبدا (از ۵ ستاره)،
- کیفیت StatementOfPurpose (از ۵ ستاره)،
- کیفیت LetterOfRecommendation (از ۵ ستاره)،
- معدل کل (CGPA) (از ۱۰ نمره)
- و اینکه دانشجو سابقه‌ی کار تحقیقاتی داشته یا خیر

استفاده کنید تا **احتمال پذیرش** هر دانشجو در دانشگاه‌های خارج از کشور را تخمین بزنید.)

سوالات:

ابتدا داده‌ها را در یک $dataframe$ ذخیره کنید. دقت کنید که به طور معمول در داده‌های واقعی بعضی از متغیرها حاوی نمونه‌هایی با مقادیر نامشخص هستند. داده‌ها را به گونه‌ای ذخیره کنید که به جای مقادیر نامشخص $< NA >$ قرار بگیرد. حال ۶ سطر اول موجود در دیتاست را نمایش دهید.

الف) همان‌طور که اشاره شد وجود داده‌های گم‌شده اجتناب‌ناپذیر است، ابتدا تعداد داده‌های ناموجود در هر ستون را نمایش دهید و سپس با جستجو در منابع اینترنتی درباره راهکارهای حل این مشکل جستجو کنید و به صورت خلاصه در گزارش کار شرح دهید. بهترین روش برای پر کردن داده‌های ناموجود را انتخاب کرده و با استفاده از این روش داده‌های ناموجود دیتاست را پر کنید.

ب) برای نمرات تافل دانشجویان یک هیستوگرام با $bin\ size$ مناسب بکشید و ویژگی های بارز توزیع آن را بیان کنید.

ج) **چولگی** یک متغیر تصادفی توصیف مناسبی از رابطه‌ی میانگین و میانه‌ی آن است. رابطه‌ی چولگی یک متغیر تصادفی را یافته و سعی کنید آن را توصیف کنید. حال چولگی را برای این ستون $University\ Rating$ دانشجویان محاسبه کرده و نتیجه را توصیف کنید.

ت) مقادیر میانگین، میانه، چارک‌های اول و سوم و کمترین و بیشترین مقدار هر ستون را برای همه‌ی ستون‌ها نمایش داده و یکی از آن‌ها را توصیف کنید.

ث) با ماتریس کواریانس در درس آشنا شده‌اید، تفاوت عمده ماتریس همبستگی (correlation) با ماتریس کواریانس در این است که درایه‌های آن حاوی همبستگی دوبه‌دوی متغیرها را به جای کواریانس‌شان است. به کمک ابزارهای موجود در R ماتریس همبستگی ویژگی‌های داده‌ها را بدست آورید.

ک) ویژگی‌ای که بیشترین همبستگی با احتمال پذیرش دانشجو دارد را پیدا کنید. نمودار $scatter\ plot$ این ویژگی و احتمال پذیرش را رسم کنید. چه نتیجه‌ای از این نمودار می‌گیرید؟

گ) نمودار $density\ plot$ را برای این متغیر بکشید و خط مربوط به میانگین را به آن اضافه کنید.

ل) در مورد تابع cor.test در R تحقیق کنید. با استفاده از این تابع correlation significance را تست کنید. با استفاده از مقدار p-value همبستگی این دو متغیر را تحلیل کنید.

م) بازه اطمینان ۹۵ درصدی را برای میانگین این متغیر محاسبه کنید. (می‌توانید از کتابخانه‌های R برای محاسبات خود استفاده کنید)

ه) فرض کنید اگر احتمال پذیرش دانشجویی بیشتر از ۸۰ درصد باشد، او تصمیم به اپلای خواهد گرفت و از شما در این باره کمک خواسته است. با توجه به ویژگی‌ای که بیشترین همبستگی با احتمال پذیرش دارد، $threshold$ای برای افراد با احتمال بیشتر از ۸۰ درصد پذیرش استخراج کنید. چگونه می‌توانید با استفاده از این مقدار به فرد مورد نظر کمک کنید؟

ی) یک نمونه تصادفی با سایز ۲۵ از مجموعه داده انتخاب کنید و دو متغیر عددی از ویژگی‌های موجود در مجموعه داده را انتخاب کنید. حال می‌خواهیم از این داده‌ها برای مقایسه مقدار **میانگین** بین دو متغیر استفاده کنیم.

یک آزمون فرضیه طراحی کنید تا ببینید آیا این داده‌ها شواهد قانع‌کننده‌ای از تفاوت بین مقادیر میانگین ارائه می‌دهند یا خیر. آیا نتیجه با فاصله اطمینان ۹۵ درصد مطابقت دارند؟