**پروژه پایانی درس آمار و احتمالات**

**موضوع :**تحلیل آماری داده‌های بیماران مبتلا به دیابت

**سطح :**کارشناسی مهندسی صنایع

**تعداد نفرات :**انفرادی یا تیم‌های دو نفره

**موعد تحویل :** 7 / 4 / 1404

**هدف پروژه:**

در این پروژه، دانشجویان با استفاده از داده‌های واقعی بیماران (مجموعه داده دیابت)، مفاهیم آمار توصیفی، برآوردهای آماری و آزمون‌های فرض را به‌صورت کاربردی تمرین کرده و توانایی تحلیل آماری داده‌های دنیای واقعی را تقویت می‌کنند.

**داده مورد استفاده:**

مجموعه دادهPima Indians Diabetes

منبع UCI Machine Learning Repository یا Kaggle[1]

توضیح: این داده‌ها شامل 768 بیمار زن با متغیرهایی مانند سن، تعداد بارداری، شاخص توده بدنی(BMI) ، سطح قند خون(Glucose) ، و... به‌همراه وضعیت ابتلا به دیابت (0 یا 1) هستند.

**مراحل انجام پروژه:**

**۱ .تحلیل اکتشافی داده‌ها:(EDA)**

در این بخش، داده‌ها را از نظر آماری اولیه بررسی کرده و دید کلی نسبت به توزیع و ویژگی‌های آن‌ها به‌دست آورید:

- محاسبه شاخص‌هایی مانند میانگین، میانه، واریانس، انحراف معیار، بیشینه، کمینه و چارک‌ها برای متغیرهای عددی.

[1] https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

- بررسی وجود مقادیر پرت (Outliers) با نمودار Boxplot.
- رسم هیستوگرام و توصیف شکل توزیع برای متغیرهای اصلی Glucose ، BMI، Age،...
- رسم Heatmap یا Pairplot برای مشاهده همبستگی بین ویژگی‌ها.
- بررسی وجود داده‌های گمشده و مستندسازی نحوه برخورد با آن‌ها.

**۲ .برآوردهای آماری:**

- **برآورد نقطه‌ای :**برای هر متغیر کلیدی مثلاً Glucose ، یک مقدار نقطه‌ای برای جامعه ارائه دهید (مثلاً میانگین سطح قند خون برای گروه دیابتی).
- **برآورد فاصله‌ای :**محاسبه فاصله اطمینان ٪۹۵ برای میانگین BMI و Glucose به تفکیک دو گروه Outcome=0و Outcome=1 با استفاده از توزیع t ( تفسیر فاصله اطمینان به زبان ساده آماری مثلاً: با احتمال ٪۹۵، میانگین سطح گلوکز افراد دیابتی بین X و Y قرار دارد. )

**۳ .آزمون فرض:**

فرموله‌سازی و اجرای حداقل دو آزمون فرض آماری مانند:

- آزمون t برای بررسی تفاوت معنادار میانگین Glucose بین دو گروه دیابتی و غیردیابتی.
- آزمون برابری واریانس‌ها یا آزمون نسبت برای متغیرهای باینری (در صورت لزوم).
- محاسبه p-value ، تصمیم‌گیری درباره رد یا عدم رد $H_0$ و تفسیر نتیجه.

**۴ .مدل‌سازی (اختیاری ولی امتیازی):**

اجرای یک مدل ساده برای پیش‌بینی دیابت بر اساس متغیرهای انتخاب‌شده.

- تحلیل ضرایب مدل و تعیین متغیرهای معنادار.
- محاسبه دقت مدل (accuracy) و گزارش معیارهای طبقه‌بندی مانند confusion matrix یا ROC.

**۵ .تهیه گزارش نهایی:**

تهیه یک فایل PDF شامل:

- مقدمه، هدف پروژه و معرفی داده‌ها
- تحلیل‌های آماری و گرافیکی همراه با تفسیر
- نتایج آزمون‌های فرض و برآوردها
- تحلیل مدل (در صورت اجرا)
- نتیجه‌گیری کلی از یافته‌ها
- پیوست نمودارها و کدها