سلام بر تمام دانشجویان عزیز، چند نکته مهم:

1. حجم گزارش به هیچ عنوان معیار نمره‌دهی نیست، در حد نیاز توضیح دهید.
2. نتایج و نمودار های مربوطه به صورت واضح قرار داده شود، این موارد را حتما استدلال کنید.
3. نکته‌ی مهم در گزارش نویسی، روشن بودن پاسخ ها می باشد، اگر فرضی برای حل سوال استفاده می‌کنید حتما آن را ذکر کنید، اگر جواب نهایی عددی است به صورت واضح آن را بیان کنید.
4. مهلت تحویل پروژه تمدید نخواهد شد.
5. شباهت در کد و گزارش مربوطه, به منزله تقلب می باشد و افراد مذکور نمره کل پروژه را از دست خواهند داد.
6. برای دیتاست مربوطه از فایل diabetes.csv استفاده کنید.
7. در صورت سوال داشتن با ایمیل salar.nouri@ut.ac.ir در ارتباط باشید.

بخش اول) Introduction:

بیماری دیابت، یک مشکل اساسی در ایالات متحده آمریکا می‌باشد. در سال ۲۰۱۰، به طور تقریبی از کل جمعیت ایالات متحده، حدود ۲۹,۱ میلیون نفر (۹,۳ درصد جمعیت کل) بیماری دیابت داشتند.

فایل diabetes.csv شامل ۳۷۵ نمونه از آفریقایی های ساکن در ایالت ویرجینیا تک مرکزی که به بیماری قلبی-عروقی مبتلا هستند، می باشد، که برای مطالعه جهت جلوگیری از این بیماری به صورت تصادفی جمع آوری شده است. بیماری دیابت نوع ۲ بیماری است که فرد از بدو تولد به آن مبتلا نمی باشد، بلکه در طول زندگی فرد این بیماری توسعه می‌یابد. تلاش برای ارزیابی فاکتورها تاثیر گذار بر دیابت نوع ۲ جهت پیشگیری آن در دهه های اخیر بیشتر شده است.

در افرادی که وضعیت دیابت خود را بررسی نمی کنند، مقدار هموگلوبین گلیکوزیله در مقدار خون آن ها در حال افزایش می باشد. از آن جهت که گلوکوز به هموگلوبین در حدود ۱۲۰ روز در خون می چسبد، در نتیجه مقدار هموگلوبین گلیکوزیله در خون، نشان دهنده ی مقدار متوسط گلوکوز در حدود ۳ ماه می باشد. مقدار سطح ۷ و یا بالاتر برای مقدار هموگلوبین گلیکوزیله نشان‌دهنده دیابت نوع ۲ می‌باشد. فاکتورهای که می توانند نشان دهنده دیابت باشند:

Chol: *cholesterol*	**stab.glu:** *stabilized glucose levels*
hdl: *high density lipoprotein*	**ratio**: *chol/hdl*
age	**height**: *is in inches*
weight: *is in pounds*	**bp.1s**: *_rst systolic blood pressure*
bp.1d: *_rst diastolic blood pressure*	**waist**: *given in inches*
hip: *circumference given in inches*	**frame**: *complexity*
time.ppn: *minutes after eating that their glucose levels were measured (postprandial time)*	

هدف شما تحلیل آماری این متغیرها می‌باشد تا بینشی جهت تاثیر فاکتورها مختلف برای وجود دیابت به دست آورید.

بخش دوم) Accessing Data, Visualization and Summarization :

در ابتدا با visualization و summarization به بررسی و آشنایی با دیتاست مورد نظر می پردازیم:

1. برای فاکتور glyhb، هیستوگرام، نمودار جعبه‌ای و qq-plot را جهت بررسی نرمال بودن بررسی کنید. ویژگی هایی که مشاهده می کنید را گزارش کنید.
2. موارد قسمت اول را برای فاکتور chol نیز تکرار کنید. کدام یک از این فاکتورها تقریب بهتری از توزیع گاوسی می باشد؟
3. مستقل و وابسته بودن فاکتورها (ویژگی ها) را به یکدیگر بررسی کنید.
4. با توجه به موارد بیان شده در بخش اول، برای هر کدام از فاکتورها، جهت به دست آوردن وابستگی بین دو فاکتور، نمودار جعبه‌ای و مقدار glyhb را مقایسه کنید.
5. برای متغیر های BMI (Body Mass Index) و WHR (waist-to-hip ratio) که به ترتیب زیر تعریف می‌شوند، توزیع مربوطه را به دست آورید، همچنین نمودارهای جعبه‌ای شرطی را رسم کنید.

$$BMI = \frac{weight}{(height)^2}; \quad WHR = \frac{waist}{hip}$$

6. فاکتورها مربوطه را به ترتیب بیشترین تا کمترین وابستگی به دیابت نوع ۲ مرتب کنید و گزارش دهید.

بخش سوم) Parametric Inference:

در این بخش بر روی تعدادی از فاکتورها مشخص، استنباط پارامتری انجام می‌دهیم.

7. با استفاده از method of moments، یک توزیع گاما به BMI برازش کنید. با استفاده از bootstrap غیر-پارامتری، بازه اطمینان ۹۵٪ حول پارامترهای تخمین زده شده به دست آورید. نیکویی برازش را برای برازش انجام شده بررسی کنید.
8. با استفاده از Maximum Likelihood، یک توزیع نرمال به WHR برازش کنید. و بازه اطمینان ۹۵٪ را به دست آورید. نیکویی برازش را برای برازش انجام شده بررسی کنید.
9. نسبت های مطرح شده در بالا، با توجه به سن افراد رفتار متفاوتی دارند. هدف ما فهمیدن تاثیر آن ها بر مقدار glyhb می باشد. برای این هدف، برازشی را طراحی کنید، همچنین بازه اطمینان را بر حسب میانگین و واریانس گزارش دهید. نتیجه خود را بیان کنید.

بخش چهارم) Testing:

هدف از این بخش، یافتن فاکتورهای است که بر دیابت نوع ۲ تاثیر می‌گذارد.

1. آیا مردان و زنان به طور برابر در معرض دیابت نوع ۲ قرار دارند؟
2. به دلخواه ۵ فاکتور را انتخاب کنید، برای هر فاکتور تست برابری میانگین برای کسانی که به دیابت نوع ۲ مبتلا می باشند و نمی باشند را انجام دهید. (مقدار سطح معناداری ۰,۰۵ می باشد.).
   آماره تست (پارامتری ویا غیر-پارامتری) را با آرگومان های مربوطه توجیه کنید.
3. احتمال اینکه مرد دیابتی، مقدار BMI بیشتری نسبت به مرد غیردیابتی داشته باشد، را تخمین بزنید.
   این عمل را برای WHR نیز تکرار کنید. بازه اطمینان ۹۵٪ را برای هر دو حالت تشکیل دهید.
4. در صورتی که شما یک بیمار مرد جدیدی از همان جامعه آماری داشته باشید، و اگر فقط به مقدار WHR آن شخص دسترسی داشته باشید، با استفاده از توزیع empirical، یک تست برای دیابت نوع ۲ تشکیل دهید. مقدار سطح معناداری را ۰,۰۵ در نظر بگیرید، و مقدار توان تست را به دست آورید.
5. با استفاده از مقادیر جدول برای BMI، آیا جامعه مرد و زن توزیع همگن دارند؟ (استدلال کنید)
   این کار را برای مقادیر مربوط به WHR نیز انجام دهید.

6. با استفاده از مقادیر مربوط به دو جدول، آزمایش کنید که این دو مقدار برای شناسایی مقدار glyhb، چگونه تعامل دارند. (مقدار سطح معناداری ۰،۰۵ می باشد.) کدام یک از این نسبت‌ها به مقدار glyhb حساس‌تر می باشد؟ (استدلال کنید)
7. چهار تا فاکتور chol, hdl, age, hip را در نظر بگیرید، از این فاکتورها ۲ مورد را به دلخواه انتخاب کنید، با استفاده مقادیر مربوطه، بررسی کنید که این فاکتورها چه رابطه ای با شناسایی glyhb دارند، نحوه تعامل این فاکتورها را بررسی کنید. آیا می توان بر روی میانگین فاکتورها نتیجه ای گرفت؟ تستی را جهت بررسی تفاوت میانگین این فاکتورها انجام دهید.
8. معیار LLC را به عنوان فاکتور سلامت در نظر بگیرید، موارد زیر را بررسی کنید.

$$LLC = \frac{(BMI)^2}{Chol} \times \frac{WHR}{age}$$

الف) به کمک رگرسیون رابطه فاکتور سلامت (به عنوان متغیر مستقل) و مقدار glyhb را به دست آورده و بازه اطمینان 95٪ پارامتر آن را به کمک ttest و Fisher بیان کنید.

ب) به روش bootstrap بازه اطمینان 95٪ جهت تخمین فاکتور سلامت جامعه را به کمک سه روش بحث شده در کلاس به دست آورید (روش‌ها: تقریب گوسی، کوانتایل، و روش پایه)

ج) با استفاده از تست فرض مناسب و موارد بیان شده در موارد الف و ب سلامت جامعه را تعیین کنید.

9. الف) با مقایسه دو به دوی فاکتورهای ذکر شده ماتریس p_value را بسازید و به کمک نقشه رنگی نمایش دهید. درایه ijام این ماتریس p_value مقایسه فاکتور iام و فاکتور jم است. ب) این ماتریس را به کمک Bonferroni و FDR تصحیح کرده و مجددا نمایش دهید. ج) دو روش سوال قبل را از منظر کارایی مقایسه کنید.
10. فرض کنید افراد شرکت کننده در این مطالعه را بر اساس مقدار glyhb به چهار دسته تقسیم کنیم (از چارک‌های مختلف glyhb برای دسته بندی استفاده کنید). به کمک MANOVA تحقیق کنید که سایر فاکتورهای موجود (همه 12 فاکتور به صورت همزمان) کدام یک از دسته‌های ایجاد شده بر اساس glyhb را بهتر توصیف می‌کنند. (اختیاری)

بخش پنجم) Regression:

هدف از این بخش، پیش بینی دیابت نوع ۲ با استفاده از فاکتورها اشاره شده، و نتایج به دست آمده از قسمت‌های قبل می‌باشد.

حال فرض کنید که تابع زیر را به عنوان تابع آستانه در نظر بگیریم:

$$\rho(y,\lambda)=\begin{Bmatrix}1 & \text{if } y>\lambda\\ 0 & \text{o.w.}\end{Bmatrix}$$

11. با استفاده از رگرسیون خطی بر glyhb، و فاکتورهای که به بیشترین وابستگی را دارند، جهت پیش بینی دیابت نوع ۲ به کارگیرید. مقدار خطا را برای این پیش بینی خود محاسبه کنید.
12. مقدار False Positive Rate (کسانی که دیابت دارند، در حالی که غیر دیابتی پیش بینی شده اند) و False Negative Rate را به دست آورید. حال در نظر بگیرید که مقدار نسبت دومی حدود ۱۰٪ از positive بیشتر باشد. مقدار lambda در تابع آستانه چگونه باید باشد؟ مقدار False Positive Rate که در این حالت به دست می آید، چقدر می‌باشد؟
13. کدام فاکتورها بیشترین تاثیر را دارند؟ تست کنید که فاکتورها hdl, bp 1s, bp 1d آیا با سطح معناداری ۰,۰۵ مقدار پیش بینی دارند؟
14. با استفاده از نتیجه تفاکتور بین دو نسبت، از اطلاعات مربوطه، جهت بهبود رگرسیون استفاده کنید؟
15. مقدار residual ها را به عنوان تابعی از stab.glu و نسبت های BMI, WHR رسم کنید. برای پایداری مقدار واریانس residual error راهکاری ارایه دهید؟
16. Logistic Regression را نیز با استفاده از فاکتورها نرمالیزه شده انجام دهید و نتیجه را با رگرسیون خطی بررسی کنید.
17. در نهایت با استفاده از دیتاست diabetes_test.csv این دو نوع رگرسیون را تست کنید. آیا به نتیجه مشابه در این قسمت رسیدید؟ توجیه کنید.

:Tables

Underweight	Healthy	Overweight	Level 1 Obese	Level 2 Obese	Level 3 Obese
<18.5	18.5-24.99	25-29.99	30-34.99	35-39.99	>40

Table 1: BMI Standards

Gender	Age	Low	Moderate	High	Very High
Men	20 - 29	<0.83	0.83 – 0.88	0.89 – 0.94	> 0.94
	30 – 39	<0.84	0.84 – 0.91	0.92 – 0.96	>0.96
	40 – 49	<0.88	0.88 - .95	0.96 – 1.00	> 1.00

	50 – 59 >60	<0.90 <0.91	0.90 – 0.96 0.91 – 0.98	0.97 – 1.02 0.99 – 1.03	> 1.02 > 1.03
Women	20 - 29 30 – 39 40 – 49 50 – 59 >60	< 0.71 < 0.72 < 0.73 < 0.74 < 0.76	0.71 – 0.77 0.72 - 0.78 0.73 – 0.79 0.74 – 0.81 0.76 – 0.83	0.78 – 0.82 0.79 – 0.84 0.80 – 0.87 0.82 – 0.88 0.84 - 0.90	> 0.82 > 0.84 > 0.87 > 0.88 > 0.90

Table 2: WHR Standards for Men and Women