هدف از این پروژه مروری بر مفاهیم کلیدی درس یادگیری ماشین آماری و تمرکز روی مدل‌های رگرسیون آماری می‌باشد. این پروژه از سه مجموعه داده که یک مورد از آن‌ها مرتبط با دادگان بومی نمرات دانشجویان دو درس یکی از دانشکده‌های دانشگاه برای سال‌های مشخص و چند مجموعه داده مصنوعی تشکیل می‌شود. دادگان را می‌توانید از طریق لینک دریافت نمایید. آزمایشهای خواسته شده را به دقت انجام داده و نتایج هر آزمایش را با نتیجه‌گیری و ارائه تحلیل علمی عملی ارائه کنید.

۱- آزمایش مجموعه داده مصنوعی

به منظور انجام آزمایش روی مجموعه داده مصنوعی موجود در فایل dataset01.csv به سوالات ذیل پاسخ دهید. این مجموعه داده متشکل از ۶۰۰ نمونه می‌باشد که هر یک از آنها با ۹ ویژگی مختلف مشخص شده‌اند. از میان این ویژگی‌ها هشت ویژگی اول متغیرهای مستقل و ویژگی نهم متغیر وابسته (هدف) است. از ۵۰۰ داده ابتدایی برای آموزش و از ۱۰۰داده بعدی برای اعتبارسنجی مدل استفاده کنید :

الف) نمودار نقطه‌ای مربوط به هریک از ویژگی‌های موجود در مجموعه داده به همراه متغیر هدف را ترسیم نمایید. باتوجه به نمودار رسم شده، ارتباط هر کدام از ویژگی‌ها با متغیر هدف را مورد بررسی قرار دهید.

ب) به ازای هر کدام از ویژگی‌های موجود، مدل رگرسیون خطی ساده‌ای برای پیش‌بینی متغیر هدف ارائه دهید. و سپس به سوالات ذیل با تفکیک مشخص و مرتب پاسخ دهید:

- **ب-۱)** پارامترهای β_0 و β_0 مدل با استفاده از کمترین مربعات تخمین زده و مقادیر حاصل را ذکر کنید.
- **ب-۲)** به ازای هر یک از مدل‌های بدست آمده، ابتدا خط پیش‌بینی شده را به همراه داده‌های موجود ترسیم کنید.
- **ب-۳)** به ازای مجموعه داده‌های آموزشی و آزمایشی معیارهای RSS و ضریب تشخیص[۱] را محاسبه کنید.
- **ب-٤)** برای هرمدام از پارامترهای تخمین زده شده، انحراف معیار متناظر را تخمین زده و ثبت نمایید.
- **ب-۵)** مقدار تخمین زده شده برای σ^2 را نیز در هر کدام از مدل‌های بدست آمده محاسبه نمایید.

ج) در بخش قبل به ازای هر کدام از متغیرهای مستقل و ویژگی هدف، مدل رگرسیون خطی جهت پیش‌بینی متغیر هدف بدست آمد. مجدد به سوالات ذیل با تفکیک مشخص و مرتب پاسخ دهید:

- **ج-۱)** با توجه به آزمایش‌های انجام شده، کدام ویژگی بهترین گزینه برای پیش‌بینی متغیر هدف است؟ چرا؟
- **ج-۲)** پس از انتخاب یکی از ویژگی‌ها به عنوان بهترین ویژگی، در یک فرآیند رو به جلو، ویژگی دوم را به ویژگی انتخابی اول اضافه کنید. در تمامی ۷ حالت بدست آمده، معیار AIC را محاسبه کنید.
- **ج-۳)** با بررسی تغییر حاصل در معیار AIC، ویژگی دوم انتخابی را مشخص کرده و به مدل اضافه کنید.
- **ج-٤)** پس از افزودن ویژگی دوم، معیارهای RSS و R^2 را محاسبه کنید.

[۱] R^2

د) همانند موارد ذکر شده در بخش ج، سایر ویژگی‌های موجود را با توجه به بهبود معیار AIC به مدل اضافه کنید. در هرکدام از مراحل ویژگی افزوده شده و معیارهای RSS و R^2 را محاسبه کنید. نمودار معیار RSS را در حین افزودن ویژگی‌ها ترسیم کنید. چه تغییری در این معیار رخ می‌دهد؟

ه) در این مرحله قصد داریم تا با استفاده از تمامی ویژگی‌ها به تخمین هدف بپردازیم. با استفاده از Least square برای تخمین پارامترهای مدل رگرسیون خطی، مدل را تشکیل دهید. و به سوالات ذیل پاسخ دهید:

- **ه-۱)** ماتریس واریانس-کواریانس پارامترهای β را بدست آورید.
- **ه-۲)** تخمین غیربایاس شده σ^2 را بدست آورده و ثبت کنید.
- **ه-۳)** برای مدل رگرسیون خطی ارائه شده، معیار خطا را با استفاده از Leave-one-out-cross-validation بدست آورید. برای محاسبه این معیار از دو روش ذکر شده در کتاب (n بار آموزش مدل و یک بار آموزش مدل) استفاده کنید.

و) با توجه به بخش قبل، مدلی متشکل از ۸ ویژگی برای پیش‌بینی متغیر هدف داریم. حال در یک فرآیند رو به عقب، در هر مرحله یک ویژگی را با استفاده از leave-one-out-cross-validation حذف کنید تا جایی که تنها یک متغیر باقی بماند. در هنگام حذف اولین ویژگی، معیار leave-one-out-cross-validation را در ازای حالت‌های ممکن ذکر کرده و دلیل انتخاب ویژگی نهایی را ذکر کنید. در روند حذف متغیرها تا رسیدن به تک متغیر معیار RSS را محاسبه کرده و نمودار آن را رسم کنید. این معیار در حین حذف چه تغییری دارد؟

ز) بهترین مدل بدست آمده از بخش قبل را در نظر بگیرید. با تغییر درصد داده‌های آموزش و آزمایش، پارامترهای مدل را مجدد آموزش دهید. خطای RSS حاصل از مدل بر روی داده‌های آموزش و آزمایش را بدست آورده و نمودار مربوطه را رسم کنید. تغییرات RSS را تحلیل کنید.

۲- مجموعه داده نمرات

این مجموعه داده در فایل با نام dataset02.csv متشکل از ۶ ویژگی و متغیر هدف است. ۲۳۰ داده ابتدایی را به عنوان داده آموزشی و ۴۲ داده انتهایی را به عنوان داده آزمایشی در نظر بگیرید:

الف) نمودار نقطه‌ای مربوط به هریک از ویژگی‌های موجود در مجموعه داده به همراه متغیر هدف را ترسیم نمایید. باتوجه به نمودار رسم شده، ارتباط هر کدام از ویژگی‌ها با متغیر هدف را مورد بررسی قرار دهید.

ب) این مجموعه داده دارای مقادیر نامشخص است. مقادیر نامشخص را با عدد صفر مقداردهی کنید. روشی برای پرکردن مقادیر نامشخص جستجو و سپس ارائه کنید. پس از پرکردن مقادیر نامشخص، مجددا نمودار نقطه‌ای را همانند بخش الف رسم کرده و تغییرات حاصل را تحلیل کنید.

ج) روش Lasso را بر روی مجموعه داده اجرا نمایید.

د) با تغییر پارامتر λ ، مدل‌های مختلف را آموزش دهید. نمودار معیار Lasso را برحسب پارامتر λ ترسیم کنید. بهترین مدل را مشخص کرده و برای آن مدل معیارهای RSS و R^2 را به ازای مجموعه داده آموزشی و آزمایشی تحویل دهید.

ه) با استفاده از بهترین مدل بدست آمده، مقدار متغیر هدف را برای مجموعه داده بدون برچسب مجموعه داده سوم (dataset02_unlabeled) بدست آورید. فایل خروجی را در کنار فایل نهایی با نام dataset02_mylabel قرار دهید.

و) مجموعه داده dataset02_extended با اندکی تغییر در داده قبلی و افزوده شدن ستونی جدید به ابتدای داده‌ها بدست آمده است. روش Lasso را با پارامتر 0.001 بر روی این مجوعه داده اجرا کنید. مقادیر β حاصل را مورد بررسی قرار دهید.

۳- آشنایی با مدل‌های احتمالاتی گرافی

این بخش از پروژه قرار است با استفاده از مجموعه داده تشخیص بیماری‌های سلولی (همچون نئوپلاسم) موجود در فایل dataset03.csv پیش رود. برای این قسمت از پروژه از مجموعه داده مذکور استفاده نمایید. جزئیات مرتبط با مجموعه داده در فایل متنی dataset03-info.txt نوشته شده است:

الف) پیش‌پردازش‌های مورد نیاز مانند پرکردن مقادیر نامشخص و گسسته‌سازی ویژگی‌های پیوسته را انجام داده و در گزارش خود توضیح دهید (توضیحات مرتبط با روش های پیشنهادی را ارائه کنید).

ب) نمودار نقطه‌ای مربوط به هر یک از ویژگی‌ها (با توجه به کلاس مورد نظر) را رسم کرده و در مورد قدرت جداکنندگی هرکدام توضیح دهید.

ج) مدل بیز ساده را با در نظر گرفتن تمام متغیرهای اصلی آموزش داده و دقت مدل را با استفاده از 10 Fold Cross Validation گزارش کنید

د) حداقل سه زیرمجموعه از ویژگیها را انتخاب کرده (با ذکر دلیل) و مدل بیز ساده را آموزش داده و همانند قسمت ج دقت حاصل را گزارش کرده و با آن مقایسه نمایید.

ه) ویژگیهای مجموعه داده را بررسی کرده و حداقل دو مدل گرافی را با کمک گرفتن از دانش خبره یا تحلیل خودتان ساخته و دلایل انتخاب هر مدل را نیز توضیح دهید. دقت مدلها را همانند بخشهای قبل گزارش کرده و نتایج را با یکدیگر و با بخشهای قبل مقایسه نمایید. مدلهای ساخته شده را در گزارش خود رسم کرده و احتمالهای شرطی لازم برای هرکدام را بیان کنید. در هر مدل حداقل از 7ویژگی استفاده کنید