پروژه اول درس داده کاوی

عنوان :

پیش‌بینی بیماری با مدل رگرسیون

استاد:

دکتر حسین رحمانی

پاییز ۱۴۰۲

راهنمای پروژه

- مهلت ارسال پروژه تا **ساعت ۲۳:۵۹ تاریخ ۲۶ / ۰۸ / ۱۴۰۲** است و قابل تمدید نخواهد بود.
- به ازای هر روز تاخیر **۲۵ درصد** از نمره پروژه کسر خواهد شد.
- پاسخ به سوالات این پروژه باید در قالب یک گزارش با **فرمت PDF** یا به همراه **توضیحات فایل نوتبوک (Markdown)** ارائه شود.
- در صورت ارائه گزارش در قالب **توضیحات فایل نوتبوک**، توضیحات باید کامل، جامع و شفاف باشد.
- در صورت ارائه گزارش با **فرمت PDF** ، فایل **کدهای اجراشده** نیز پیوست شود.
- تمامی فایل‌های این پروژه (گزارش و کدها) در قالب یک **فایل فشرده rar یا zip** با نام‌گذاری زیر ارسال شود.

StudentNumber_FirstName_LastName_Prj01.zip

- فایل تمرین را **حتما** در سامانه LMS آپلود نمایید. بدیهی است که تحویل از طریق ایمیل و یا سایر راه‌های ارتباطی قابل پذیرش نخواهد بود.
- **رعایت نکات نگارشی** در نوشتن گزارش نمره مثبت خواهد داشت.
- برای پاسخ به سوالات این پروژه **حتما** باید از زبان برنامه‌نویسی **پایتون** استفاده شود.

مباحث تحت پوشش: Dimension Reduction, Feature Selection, Correlation, Regression

هدف پروژه: هدف از پروژه ساخت مدل رگرسیونی برای پیش‌بینی ابتلای افراد به Cancer و Liver Disease می‌باشد.

۱- فایل ورودی:

بررسی ملی سلامت و تغذیه (NHANES) یک برنامه مطالعاتی است که برای ارزیابی وضعیت سلامت و تغذیه بزرگسالان و کودکان در ایالات متحده طراحی شده است.

این دیتاست به صورت مجموعه ای از نظرسنجی‌ها با تمرکز بر گروه‌های مختلف جمعیتی یا موضوعات بهداشتی انجام شده است و شامل سوالات جمعیت شناختی، اجتماعی، اقتصادی، رژیم غذایی و سلامتی است. به جز معاینه شامل اندازه گیری‌های پزشکی، دندان‌پزشکی و فیزیولوژیکی و همچنین تست‌های آزمایشگاهی است. از آن‌جا که ویژگی‌های این دیتاست رمزگذاری شده است، در فایل ColumnDefinitions توضیحات کاملی از معنی هرکدام از ویژگی‌ها قرار داده شده است.

ابتدا مجموعه داده را بشناسید و با متغیرهای آن آشنا شوید. فرآیند جمع‌آوری داده‌ها، ویژگی‌های موجود و ارتباط بین دیتاست‌ها و ویژگی‌هایشان را درک کنید.

برای درک بیشتر از معانی داده‌های عددی موجود برای هر ستون در دیتاست را می‌توانید در لینک‌های زیر مشاهده کنید.

Demographic, Diet, Examination, Labs

۲- پیش‌پردازش داده:

یکی از مهم‌ترین مراحل قبل از شروع هر نوع تحلیلی پیش‌پردازش داده‌ها می‌باشد. در داده‌های خود کاوش کنید و درک کاملی از ساختار و ویژگی‌های آماری آن بدست آورید، همچنین به منظور نتیجه‌گیری بهتر از بسیاری از الگوریتم‌های داده‌کاوی، لازم است تغییرات و یا اصلاحاتی بر روی داده‌های خام انجام شود. به دنبال مقادیر گم شده، outliers و یا ناسازگاری‌ها بگردید و استراتژی خود را برای رفع آن‌ها اعمال کنید. برای مثال ستون‌های با بیش از X درصد مقادیر خالی را حذف کند.

توجه: در دیتاست هنگامی که داده‌ها قابل استفاده نیستند یا زمانی که بیمار از پاسخ دادن امتناع می‌ورزد از ترکیب‌های ۷ و ترکیب‌های ۹ استفاده شده است.

۳- انتخاب ویژگی‌ها[1]:

الف. تجزیه و تحلیل همبستگی[2] را برای شناسایی روابط بین متغیرها در مجموعه داده انجام دهید. ضریب همبستگی بین هر جفت صفات[3] را برای درک وابستگی بین آن‌ها محاسبه کنید.

ب. سپس برای نمایش این همبستگی بین جفت صفت‌ها از روش نمایش Heat Map استفاده نمایید. به منظور سادگی نمایش و افزایش قابلیت درک تنها ۱۰۰ صفتی که به صورت قدر مطلق، بالاترین همبستگی را با یکدیگر دارند را جدا کرده و فقط آن‌ها را نمایش دهید.

ج. از بین تمامی صفات، آن‌هایی که با یک‌دیگر همبستگی بالایی دارند می‌توانند نمایان‌گر یک ویژگی باشند؛ بنابراین در این مرحله با تعیین یک حد آستانه مناسب از همبستگی، از بین آن‌هایی که با یکدیگر هم بستگی بالایی دارند، یکی را انتخاب کنید.

۴- کاهش ابعاد[4]:

الف. حال بر روی صفات بدست آمده در مرحله‌ی سوم، با استفاده از روش PCA ابعاد (صفات) را کاهش دهید.

ب. از یک تکنیک دیگر به جز PCA برای کاهش ابعاد استفاده کنید.

۵- مدل رگرسیون:

بهترین مدل رگرسیون از بین مدل‌های موجود را بر اساس ماهیت داده‌هایتان انتخاب کنید. دو مجموعه داده‌ی به دست آمده در مرحله قبل خود را جداگانه به مجموعه‌های آموزشی (Train) و آزمایشی (Test) تقسیم کنید. از مجموعه‌های آموزشی برای آموزش مدل رگرسیون بر روی ویژگی‌های انتخاب شده و ویژگی‌های هدف یعنی Cancer (MCQ220) و Liver (MCQ160L) موجود در فایل Questionnaire برای پیش بینی ابتلا به این دو بیماری‌ها جداگانه استفاده کنید. در این فایل عدد ۱ نشان‌دهنده ابتلا و عدد ۲ نشان دهنده عدم ابتلا به بیماری‌ها می‌باشد.

۶- ارزیابی مدل:

[1] Feature Selection
[2] Correlation
[3] Feature
[4] Dimension Reduction

مدل‌های رگرسیون آموزش دیده را جداگانه با استفاده از مجموعه‌ی آزمایشی ارزیابی کنید (یکی برای PCA و دیگری برای روش انتخابی خود). این مرحله به شما کمک می‌کند تا بفهمید مدلتان چقدر توانسته متغیرهای هدف را بر اساس ویژگی‌های انتخاب شده پیش‌بینی کند. برای مثال با محاسبه‌ی confusion matrix دقت مدل خود را ارزیابی کنید.

۷- نتیجه‌گیری:

نتیجه‌ی ارزیابی‌ها بر روی روش‌های PCA و روش انتخابی خود را مقایسه کنید.

- به نظر شما انتخاب صفات و کاهش ابعاد در این پروژه چه کمکی به ما کرده است؟ آیا تاثیر منفی هم داشته‌است؟

- ویژگی‌های مختلف مانند سرعت، دقت در دو مدل کاهش ابعاد را باهم مقایسه کنید.

نکته: هرگونه ابتکار عمل، خلاقیت، تحلیل‌های دیگر و استفاده از روش‌های بیشتر نمره‌ی اضافه خواهد داشت.

موفق باشید.