به نام خدا

پروژه پایانی درس داده کاوی

نیم‌سال اول سال تحصیلی ۱۴۰۳-۱۴۰۴ @ دانشگاه تربیت دبیر شهید رجائی

فهرست

مقدمه

این پروژه نهایی با هدف آشنایی شما با فرآیندهای مختلف تحلیل داده‌ها و به‌کارگیری الگوریتم‌های مختلف یادگیری ماشین طراحی شده است. در دنیای امروز، مهندسی داده و تحلیل داده‌ها بخش اساسی از بسیاری از حوزه‌های کاری و صنعتی است. داده‌ها می‌توانند به‌عنوان منبعی گرانبها برای تصمیم‌گیری‌های استراتژیک و پیش‌بینی‌های دقیق در کسب‌وکارها، علوم پزشکی، تحلیل‌های مالی، و بسیاری دیگر از حوزه‌ها استفاده شوند. با این حال، قبل از این‌که از داده‌ها در مدل‌های یادگیری ماشین استفاده کنیم، لازم است که داده‌ها به‌طور کامل تحلیل و پاکسازی شوند تا اطمینان حاصل شود که نتایج مدل‌ها به درستی منعکس‌کننده واقعیت‌ها هستند.

در این پروژه، شما درگیر انجام تحلیل داده‌ها از مراحل ابتدایی جمع‌آوری داده‌ها تا مراحل پیشرفته مانند مدل‌سازی و ارزیابی مدل خواهید بود. هدف اصلی این پروژه، آشنایی شما با مراحل مختلف پردازش داده‌ها، از جمله **پاکسازی داده‌ها(Data Cleaning)** ، **تحلیل خوشه‌بندی (Clustering)** و **طبقه‌بندی (Classification)** است. همچنین، در این پروژه، از شما خواسته خواهد شد که نه‌تنها الگوریتم‌های مختلف را پیاده‌سازی کنید، بلکه به تحلیل نتایج و نتایج به‌دست‌آمده نیز توجه کنید. این فرآیند به شما این امکان را می‌دهد که به‌عنوان یک مهندس داده، تحلیل‌های ژرف و دقیقی را ارائه دهید که فراتر از صرف اجرای الگوریتم‌ها باشد.

روش تهیه گزارش نهایی

گزارش نهایی شما باید شامل تمام مراحل انجام شده در پروژه باشد و تحلیل‌های خود را در هر بخش به طور کامل توضیح دهید. در این گزارش، بخش‌های مختلف باید به شرح زیر باشند:

۱. **مقدمه و هدف پروژه** :در این بخش، به طور مختصر توضیح دهید که پروژه شامل چه مراحلی است و هدف از انجام آن چیست. همچنین، توضیح دهید که چرا هر یک از مراحل (مثل خوشه‌بندی یا طبقه‌بندی) اهمیت دارند و چه مفهومی را در تحلیل داده‌ها به همراه دارند.
۲. **داده‌های ورودی و پیش‌پردازش داده‌ها** :در این قسمت، ابتدا توضیح دهید که داده‌های ورودی کدام‌ها هستند و نحوه بارگذاری و پیش‌پردازش آن‌ها را شرح دهید. توضیح دهید که برای پاکسازی داده‌ها چه کارهایی انجام دادید و چرا این مراحل ضروری بودند (مثل حذف داده‌های گمشده، اصلاح داده‌های غلط و ...).
۳. **الگوریتم‌های انتخابی و پیاده‌سازی**:
   - **خوشه‌بندی** :در این بخش، الگوریتم‌های **K-means** و **خوشه‌بندی سلسله‌مراتبی** را شرح دهید و توضیح دهید که چرا این الگوریتم‌ها را برای تحلیل داده‌های خود انتخاب کرده‌اید. نتایج حاصل از این خوشه‌بندی‌ها را به همراه تحلیل‌های خود شرح دهید.

خوشه‌ها را بر اساس ویژگی‌های شاخص تحلیل کنید و توضیح دهید که هر خوشه چه ویژگی‌های مشترکی داشته و چرا این ویژگی‌ها در کنار هم قرار گرفته‌اند.

- **طبقه‌بندی** :در این بخش، الگوریتم‌های **طبقه‌بندی**)مانند درخت تصمیم، SVM، KNN)را پیاده‌سازی کرده و نتایج هر یک را توضیح دهید. علاوه بر مقایسه دقت، تحلیل خود را از نتایج ارزیابی مدل‌ها بیان کنید.

٤. **تحلیل نتایج** :پس از پیاده‌سازی هر الگوریتم و ارزیابی آن‌ها با معیارهایی مانند دقت (Accuracy)، حساسیت(Recall) ، دقت مثبت(Precision) ، F1-Scoreو ماتریس گیج‌زنی (Confusion Matrix)، تحلیل خود را ارائه دهید. تنها ارائه اعداد و خروجی‌ها کافی نیست، بلکه باید تفکر تحلیلی شما در خصوص نحوه عملکرد مدل‌ها و علت موفقیت یا شکست آن‌ها مشخص باشد.

٥. **نتیجه‌گیری** :در نهایت، پس از تحلیل‌های مختلف، یک نتیجه‌گیری کلی از تمامی نتایج و تحلیل‌های خود بیان کنید. چه مدل یا الگوریتمی برای داده‌های شما بهتر عمل کرد؟ چرا؟ و چه پیشنهاداتی برای بهبود عملکرد مدل‌ها دارید؟

اهمیت ژرف‌نگری در تحلیل داده‌ها

در این پروژه، شما به‌عنوان یک تحلیلگر داده باید از یک دیدگاه تحلیلی عمیق وارد فرآیند شوید. تنها پیاده‌سازی الگوریتم‌ها کافی نیست. هدف این است که شما بتوانید دلایل انتخاب هر الگوریتم و نحوه به‌کارگیری آن‌ها را توضیح دهید، نکات قوت و ضعف هر مدل را شناسایی کنید و تحلیل‌های خود را به‌طور دقیق و علمی ارائه دهید. به‌عنوان مثال، در انتخاب تعداد خوشه‌ها برای K-means باید از روش‌هایی مانند Elbow Method و Silhouette Score استفاده کنید و توضیح دهید چرا تعداد مشخصی از خوشه‌ها را انتخاب کرده‌اید.

در دنیای واقعی، داده‌ها به‌مراتب پیچیده‌تر از آن چیزی هستند که در این پروژه خواهید دید. با این حال، این پروژه به شما فرصت می‌دهد تا تجربه‌ای اولیه از تحلیل داده‌ها را به‌دست آورید و با فرایندهای اصلی آشنا شوید. تحلیل عمیق و ژرف شما در این پروژه، شما را برای مواجهه با چالش‌های پیچیده‌تر در دنیای حرفه‌ای آماده می‌کند.

شرایط انجام پروژه

- این پروژه به‌صورت **انفرادی** است و انجام آن به صورت گروهی قابل قبول نیست(شما به صورت انفرادی در یک شرکت استخدام خواهید شد).
- گزارش نهایی باید به‌صورت دقیق و منظم نوشته شود و شامل تحلیل‌های عمیق و علمی باشد.
- دانشجویان باید تمامی مراحل انجام پروژه را به‌طور کامل توضیح دهند و فقط به الگوریتم‌ها و خروجی‌ها بسنده نکنند(ژرف اندیشی شما به عنوان مهندس داده، از مهم ترین شاخص‌های ارزیابی است).

- استفاده از مدل‌های زبانی، به عنوان دستیار، بسیار خوب و پسندیده است. اما حذف نقش انسانی شما و بسنده کردن به خروجی‌های این مدل‌ها، بدون هیچ توضیحی، منجر به از دست دادن نمره این بخش خواهد شد.
- تمام منابع مورد استفاده شما در تحلیل (شامل کدها، کتابخانه ها و هر آنچه که برای ضمیمه کردن مدنظر دارید) باید در یک مخزن گیت‌هاب ارسال شده و در گزارش شما، لینک این مخزن در دسترس باشد.
- کیفیت و حرفه‌ای بودن گزارش، مورد ارزیابی است. طبیعتاً مدت زمانی که شما برای زیبایی گزارش نویسی گذاشتید، بر زمانی که گزارش ما خوانده می‌شود و مورد ارزیابی مثبت قرار خواهد گرفت، موثر است (در صورت امکان از LaTex استفاده کنید).

این پروژه به شما کمک خواهد کرد تا تجربه عملی از تحلیل داده‌ها و استفاده از الگوریتم‌های یادگیری ماشین کسب کنید و به‌عنوان یک مهندس داده، توانایی تحلیل و استخراج نتایج دقیق از داده‌ها را به‌دست آورید.

پاک‌سازی داده

در پروژه نهایی دوره، یکی از مراحل مهم و ضروری، **پاکسازی داده‌ها** (Data Cleaning)است. این مرحله به معنای شناسایی و اصلاح مشکلات مختلفی است که ممکن است در داده‌ها وجود داشته باشد و مانع از تحلیل صحیح و مدل‌سازی درست داده‌ها شوند. در دنیای واقعی، داده‌ها به ندرت به‌صورت تمیز و آماده برای تحلیل هستند و به همین دلیل پاکسازی داده‌ها به‌عنوان یکی از مهارت‌های کلیدی در علم داده و یادگیری ماشین مطرح است.

انواع مشکلات داده‌ها که باید پاکسازی شوند

۱. **مقدار گمشده (Missing Data):** یکی از رایج‌ترین مشکلات در داده‌ها، وجود مقادیر گمشده است. این مقادیر می‌توانند به دلایل مختلفی نظیر خطای انسانی، مشکلات سیستمی یا نقص در فرایند جمع‌آوری داده‌ها ایجاد شوند. برای رفع این مشکل، ممکن است از روش‌هایی چون حذف سطرهای دارای داده ناقص، پر کردن مقادیر گمشده با میانگین یا میانه، یا استفاده از مدل‌های پیش‌بینی برای جایگزینی مقادیر گمشده استفاده کنیم.
۲. **داده‌های تکراری (Duplicate Data):** در برخی مواقع ممکن است داده‌ها به‌طور تصادفی یا در فرایند جمع‌آوری داده‌ها تکرار شوند. شناسایی و حذف این داده‌های تکراری از اهمیت بالایی برخوردار است چرا که می‌تواند منجر به تحلیل‌های نادرست و مدل‌های ناقص شود.
۳. **مقادیر خارج از محدوده (Outliers):** داده‌هایی که به‌طور غیرعادی بالا یا پایین هستند و ممکن است تاثیر زیادی در نتایج تحلیل و مدل‌سازی داشته باشند. شناسایی و تصمیم‌گیری در مورد اینکه آیا این مقادیر باید حذف شوند یا اصلاح گردند، یکی از چالش‌های مهم در مرحله پاکسازی است.
٤. **عدم سازگاری در داده‌ها (Inconsistent Data):** گاهی اوقات داده‌ها به‌صورت غیرسازگار وارد می‌شوند. به عنوان مثال، مقادیر در یک ستون ممکن است با فرمت‌های مختلف وارد شوند)مثل "Yes" و "yes" یا "Male" و "M"). اصلاح این ناسازگاری‌ها برای جلوگیری از خطا در تحلیل و مدل‌سازی ضروری است.
۵. **تعارضات و خطاهای منطقی (Logical Errors):** در برخی از داده‌ها ممکن است تناقضات منطقی وجود داشته باشد، مانند وارد کردن سن یک فرد به عنوان منفی یا وجود سطح قند خون غیرقابل تصور. این نوع خطاها باید شناسایی و اصلاح شوند تا تحلیل‌ها به درستی انجام شوند.

چرا پاکسازی داده‌ها مهم است؟

پاکسازی داده‌ها اهمیت زیادی در فرایند تحلیل و مدل‌سازی دارد زیرا داده‌های کثیف می‌توانند به نتایج گمراه‌کننده و تصمیمات نادرست منجر شوند. این فرایند تضمین می‌کند که داده‌ها از دقت و کیفیت بالایی برخوردار هستند و می‌توان از آن‌ها برای ایجاد مدل‌های پیش‌بینی دقیق و تحلیل‌های

معتبر استفاده کرد. به‌ویژه در پروژه‌های داده‌کاوی و یادگیری ماشین، کیفیت داده‌ها مستقیماً بر دقت مدل‌های ساخته‌شده تأثیر می‌گذارد.

روش‌های معمول برای پاکسازی داده‌ها

- **حذف یا پر کردن مقادیر گمشده** :استفاده از میانگین، میانه یا پیش‌بینی مدل‌ها برای پر کردن مقادیر گمشده.
- **حذف داده‌های تکراری** :شناسایی و حذف سطرهای تکراری از دیتاست.
- **شناسایی و حذف مقادیر خارج از محدوده** :شناسایی داده‌های غیرمعمول و خارج از محدوده مجاز.
- **یکسان‌سازی داده‌های متنی** :اصلاح تفاوت‌های جزئی در فرمت‌ها یا مقادیر متنی.
- **اصلاح خطاهای منطقی** :شناسایی و اصلاح داده‌هایی که از نظر منطقی نادرست هستند.

در این بخش، هدف این است که با شبیه‌سازی مشکلات رایج در داده‌ها، دانش‌آموزان با استفاده از تکنیک‌های مختلف پاکسازی داده‌ها، توانمندی‌های خود را در پردازش داده‌ها به‌طور کامل تقویت کنند. این فرایند، علاوه بر بهبود کیفیت داده‌ها، دانش‌آموزان را برای مقابله با مشکلات واقعی در داده‌ها آماده می‌کند و آن‌ها را به یک تحلیل‌گر داده ماهر تبدیل می‌کند.

پیاده‌سازی و مقایسه الگوریتم‌های طبقه‌بندی

در این پروژه، شما با استفاده از دیتاست **پیش‌بینی دیابت** و بر اساس برچسب diabetes(که ۱ برای "دیابت" و ۰ برای "بدون دیابت" است)، باید دو الگوریتم مختلف طبقه‌بندی را پیاده‌سازی کرده و نتایج آن‌ها را مقایسه کنید. هدف این است که دانش‌آموزان بتوانند الگوریتم‌های مختلف یادگیری ماشین را پیاده‌سازی کرده، نتایج آن‌ها را تحلیل کنند و با استفاده از معیارهای ارزیابی مدل، تصمیم بگیرند که کدام الگوریتم عملکرد بهتری دارد.

مراحل انجام کار

بارگذاری و آماده‌سازی داده‌ها

- ابتدا دیتاست **پیش‌بینی دیابت** را بارگذاری کرده و آن را برای انجام عملیات مختلف آماده کنید.
- داده‌های گمشده را شناسایی و در صورت لزوم آن‌ها را حذف یا تکمیل کنید.
- اطمینان حاصل کنید که داده‌ها به درستی نرمال‌سازی یا مقیاس‌بندی شده‌اند، به‌خصوص اگر از الگوریتم‌هایی مانند KNN یا SVM استفاده می‌کنید که حساس به مقیاس داده‌ها هستند.

انتخاب دو الگوریتم طبقه‌بندی

- دو الگوریتم مختلف طبقه‌بندی را انتخاب کرده و پیاده‌سازی کنید. به عنوان مثال:
  - **درخت تصمیم (Decision Tree)**: الگوریتمی ساده و قابل تفسیر که از داده‌ها برای ساخت درختی استفاده می‌کند که به طبقه‌بندی کمک می‌کند.
  - **ماشین بردار پشتیبانی (Support Vector Machine - SVM)**: الگوریتمی قوی برای مسائل طبقه‌بندی که از یک مرز تصمیم برای تفکیک داده‌ها استفاده می‌کند.
  - سایر الگوریتم‌ها مانند **k-نزدیک‌ترین همسایه (KNN)** یا **رگرسیون لجستیک** نیز می‌توانند به عنوان گزینه‌های جایگزین انتخاب شوند.

آموزش و تست مدل‌ها

- داده‌ها را به دو بخش **آموزش** و **تست** تقسیم کنید (حدود ۷۰٪ آموزش و ۳۰٪ تست).
- هر دو الگوریتم طبقه‌بندی انتخاب شده را بر روی داده‌های آموزشی آموزش دهید و سپس مدل‌ها را بر روی داده‌های تست ارزیابی کنید.

ارزیابی مدل‌ها

- پس از آموزش مدل‌ها، نتایج آن‌ها را با استفاده از معیارهای ارزیابی مختلف مقایسه کنید. معیارهایی که باید محاسبه شوند عبارتند از:
  - **دقت (Accuracy):** نسبت پیش‌بینی‌های صحیح به کل پیش‌بینی‌ها.
  - **دقت مثبت (Precision):** نسبت پیش‌بینی‌های مثبت صحیح به تمام پیش‌بینی‌های مثبت.
  - **حساسیت (Recall):** نسبت پیش‌بینی‌های مثبت صحیح به کل موارد مثبت واقعی.
  - **امتیاز F1 (F1 Score):** میانگین هندسی دقت و حساسیت.
  - **ماتریس گیج‌زنی (Confusion Matrix):** برای ارزیابی دقیق‌تر عملکرد مدل‌ها و مشاهده تعداد پیش‌بینی‌های صحیح و غلط.

مقایسه نتایج

- نتایج مدل‌های مختلف را با استفاده از معیارهای ارزیابی فوق مقایسه کنید.
- بررسی کنید که کدام الگوریتم بهترین دقت را دارد و چرا ممکن است یکی از مدل‌ها نسبت به دیگری عملکرد بهتری داشته باشد.
- تحلیل کنید که کدام الگوریتم در تشخیص موارد مثبت (افرادی که دیابت دارند) عملکرد بهتری دارد.

نتیجه‌گیری

- پس از مقایسه نتایج، در یک گزارش کوتاه، دلایل انتخاب هر الگوریتم، نتایج ارزیابی و تحلیل خود را ارائه دهید.
- الگوریتمی که به نظر شما بهترین عملکرد را دارد، انتخاب کنید و دلایل خود را برای این انتخاب بیان کنید.

انجام تحلیل خوشه‌بندی(Clustering)

در این مرحله از پروژه، شما باید داده‌ها را با استفاده از دو الگوریتم **خوشه‌بندی** مختلف تحلیل کنید : **K-means**و **خوشه‌بندی سلسله‌مراتبی (Hierarchical Clustering)** .هدف این بخش، شناسایی خوشه‌ها یا گروه‌های مشابه در داده‌ها است. پس از انجام خوشه‌بندی، شما باید تحلیل دقیقی از هر خوشه انجام داده و رابطه میان اعضای هر خوشه را توصیف کنید.

مراحل انجام کار

انتخاب ویژگی‌ها برای خوشه‌بندی

- ابتدا ویژگی‌هایی را که می‌خواهید از آن‌ها برای انجام خوشه‌بندی استفاده کنید، انتخاب کنید. اصولاً استفاده از تمام ویژگی‌ها مطلوب است مگر دلیل منطقی برای نبود آن ذکر شود.

خوشه‌بندی با الگوریتم:K-means

- ابتدا داده‌ها را برای الگوریتم K-means آماده کنید. داده‌ها باید نرمال‌سازی یا مقیاس‌بندی شوند، زیرا K-means حساس به مقیاس داده‌ها است.
- سپس الگوریتم **K-means**را اجرا کنید و تعداد خوشه‌ها را به‌صورت بهینه انتخاب کنید. این تعداد خوشه‌ها را می‌توانید با استفاده از روش‌هایی مانند **Elbow Method** یا **Silhouette Score**تعیین کنید.
- پس از اجرای خوشه‌بندی، نتایج خوشه‌ها را تحلیل کنید و مشخص کنید که کدام ویژگی‌ها برای هر خوشه به‌طور برجسته‌تر هستند.

خوشه‌بندی با روش سلسله‌مراتبی:(Hierarchical Clustering)

- در این مرحله، از الگوریتم **خوشه‌بندی سلسله‌مراتبی** برای شبیه‌سازی خوشه‌ها استفاده کنید. این الگوریتم ساختار درختی (Dendrogram) ایجاد می‌کند که نشان می‌دهد چطور خوشه‌ها به‌تدریج با هم ترکیب می‌شوند.
- الگوریتم سلسله‌مراتبی را روی داده‌ها اعمال کرده و یک درخت خوشه‌بندی ترسیم کنید. سپس با استفاده از یک آستانه مناسب، تعداد خوشه‌ها را انتخاب کنید.
- نتایج خوشه‌بندی را تحلیل کرده و تفاوت‌های اصلی بین خوشه‌ها را بررسی کنید.

تحلیل خوشه‌ها

- پس از اجرای هر دو روش خوشه‌بندی (K-means وHierarchical Clustering) ، برای هر خوشه یک تحلیل دقیق انجام دهید. نشان دهید هر خوشه نماینده چه

جمعیتی است؟ در این بخش، نیاز است که ابتداً برای هر خوشه یک برچسب و نام تعیین نمایید.

- در تحلیل هر خوشه، ویژگی‌های مهم و تمایزات آن خوشه با سایر خوشه‌ها را توضیح دهید. به عنوان مثال:
  - **ویژگی‌های برجسته** :مثلا ممکن است یک خوشه افرادی با BMI بالا و سطح قند خون بالا باشد.
  - **رابطه بین اعضای خوشه‌ها** :مشخص کنید که چه ویژگی‌هایی موجب تشابه اعضای یک خوشه شده است.
  - **تحلیل منطقی** :اگر خوشه‌ای شامل افرادی با شرایط خاص (مثل سن بالا و فشار خون بالا) است، توضیح دهید که چرا این افراد در یک خوشه قرار گرفته‌اند.

مقایسه نتایج

- نتایج هر دو الگوریتم را مقایسه کنید. چه شباهت‌ها و تفاوت‌هایی در خوشه‌ها وجود دارد؟
- تحلیل کنید که چرا ممکن است نتایج یکی از الگوریتم‌ها بهتر از دیگری باشد.
- روش‌های خوشه‌بندی سلسله‌مراتبی و K-means هرکدام مزایا و معایب خاص خود را دارند. مزایای هرکدام را تحلیل کرده و دلایل انتخاب بهترین الگوریتم را ارائه دهید.

نکات مهم

- استفاده از **نرمال‌سازی یا مقیاس‌بندی داده‌ها** برای هر دو الگوریتم K-means و Hierarchical Clusteringبسیار مهم است، زیرا این الگوریتم‌ها تحت تأثیر مقیاس داده‌ها قرار می‌گیرند.
- برای انتخاب تعداد خوشه‌ها در الگوریتمK-means ، از روش‌هایی مانند **Elbow Method**یا **Silhouette Score**استفاده کنید تا بهترین نتیجه را بدست آورید.
- برای ترسیم درخت خوشه‌بندی (Dendrogram) در روش سلسله‌مراتبی، از کتابخانه‌هایی مانند SciPy یا seaborn استفاده کنید.

درنهایت باز یادآوری می‌شود که ژرف‌نگری شما، اهمیت بالایی دارد!

شاد و پیروز باشید.