هوش مصنوعی

پروژه شماره‌ی ۴

مدرس: **دکتر فدایی و دکتر یعقوب‌زاده**

طراح: کیانوش عرشی، نگار مرادی، آتیه آرمین

مهلت تحویل: شنبه ۱۰دی، ساعت ۲۳:۵۵

مقدمه **1**

معرفی مجموعه داده **2**

بررسی مجموعه داده **2**

پیش پردازش مجموعه داده **2**

آموزش، ارزیابی و تنظیم **2**

روش‌های یادگیری جمعی **2**

نکات پایانی **2**

مقدمه

در این پروژه هدف آشنایی با روش های یادگیری ماشین و تخمین داشتن دیابت بر اساس مجموعه داده جمع‌آوری شده توسط NIDDK با کمک کتابخانه Scikit-Learn است. پروژه شامل ۴ فاز می‌باشد:

۱. بررسی مجموعه داده: در این فاز به تجزیه و تحلیل داده های اکتشافی می‌پردازید و یک تحلیل ساده روی مجموعه داده انجام می‌دهید.

۲. پیش پردازش مجموعه داده: این فاز که مهمترین فاز یک پروژه یادگیری ماشین هست به پیش پردازش مجموعه داده می‌پردازد تا برای مراحل بعدی مناسب‌تر باشد و کارایی مدل با وجود داده‌های نامناسب به مشکل نخورد.

۳. آموزش، ارزیابی و تنظیم: در این فاز چند مدل آماده کتابخانه scikit-learn را برای پیش‌بینی ویژگی مطرح شده آموزش می‌دهید و پس از بررسی کارایی هر مدل و تنظیم hyper parameterها کارایی مدل‌ها را بهبود می‌دهید.

۴. روش‌های یادگیری جمعی: در فاز آخر هم با برخی روش‌های یادگیری جمعی آشنا می‌شوید.

دقت کنید که در این پروژه باید تغییرات زیادی در پارامترهای مدل‌ها، روش‌های پیش پردازش و... بدهید بنابراین سعی کنید کد مرتب و خوانایی بزنید تا اعمال این تغییرات را راحت‌تر کنید.

معرفی مجموعه داده

مجموعه داده‌ای که در اختیار شما قرار دارد از موسسه ملی دیابت و بیماری‌های گوارشی و کلیوی دریافت شده. با کمک این مجموعه داده براساس ویژگی‌های متفاوتی که در ادامه توضیح داده خواهند شد، دیابت داشتن یک شخص را بررسی می‌کنید.

Column	Description
Pregnancies	تعداد دفعات حاملگی
Glucose	سطح گلوکز در خون
BloodPressure	فشار خون
SkinThickness	زخامت پوست
Insulin	سطح انسولین در خون
BMI	Body Mass Index
DiabetesPedigreeFunction	ریسک دیابت نوع ۲
Age	سن
Outcome	اینکه شخص دیابت دارد یا خیر

بررسی مجموعه داده

در این فاز داده‌های خام را بررسی خواهید کرد. این تجزیه و تحلیل داده‌ها با نام EDA شناخته می‌شود و برای دریافت یک دید کلی نسبت مجموعه داده به کار می‌رود. مراحل زیر را انجام دهید و در هر مرحله نتیجه را تحلیل کرده و در گزارش بیاورید.

۱. ساختار کلی داده‌ها را با متدهای info و describe بدست بیاورید.

۲. برای هر ویژگی[1] تعداد و نسبت داده‌های از دست رفته را بدست بیاورید.

۳. نمودار وابستگی ویژگی‌ها به یکدیگر را رسم کنید. کدام ویژگی‌ها وابستگی بیشتری به نتیجه دارند؟

۴. برای ویژگی‌های بدست آمده در مرحله قبل نمودار تعداد مشاهدات هر مقدار منحصر به فرد را رسم کنید.

۵. ارتباط ویژگی‌ها با {متغیر} را دقیق‌تر بررسی کنید، از نمودارهای scatter و hexbin می‌توانید استفاده کنید.

۶. شما می‌توانید هر بررسی دیگری که به شناخت مجموعه کمک می‌کند را پیاده و تحلیل کنید.

[1] feature

پیش پردازش مجموعه داده

در دنیای واقعی، اطلاعات جمع‌آوری شده به راحتی کنترل نمی‌شوند و در نتیجه مقادیر خارج از محدوده، ناممکن، از دست رفته و به طور کلی گمراه‌کننده برای آموزش مدل در مجموعه داده‌ها وجود دارند. در نتیجه قبل از ادامه پروژه باید این موارد را شناسایی و اصلاح کنیم. همچنین گاهی برای بهبود کارایی مدل و سرعت یادگیری میتوان فرمت این داده‌ها را تغییر داد و خلاصه‌تر کرد. در نهایت این فاز مهمترین فاز یک پروژه یادگیری ماشین است در غیر این صورت خروجی هم خروجی بسیار نادقیقی خواهد بود (به عبارتی "garbage in, garbage out").

در موارد زیر، علت انتخاب روش خود برای حل مسئله را نیز توضیح دهید.

۱. دو روش برای حل مشکل Missing Values، حذف کل ستون و پر کردن مقادیر خالی با آماره‌ها(برای مثال مد) می‌باشد. باقی روش‌ها را توضیح دهید و مقایسه کنید.

۲. بر اساس نتایج فاز قبل، کدام داده‌ها بیشترین میزان داده گم شده را دارند؟ برای تمامی ویژگی‌ها مشکل داده های گم شده را با کمک روش‌های مطرح شده حل کنید.

۳. در ویژگی‌های عددی[2]، normalizing یا standardizing به چه منظور انجام می‌شود؟ در این پروژه نیاز به انجام این کار هست؟

۴. برای استفاده ویژگی‌های دسته‌ای[3]، که معمولا بصورت یک string یا object در مجموعه داده ذخیره شده‌اند، در آموزش مدل چه پیش پردازش‌هایی کارگر هستند؟ آیا همه داده‌های دسته این نیازمند این روش‌ها هستند؟

۵. آیا امکان حذف برخی ستون‌ها وجود دارد؟ چرا؟

۶. برای آموزش و در نهایت ارزیابی مدل یادگیری ماشین نیاز است که داده ها را به دو دسته test و train تقسیم کنیم. نسبت این تقسیم به چه صورت است؟ چه روش‌های برای تقسیم و ساخت این دو دسته وجود دارد؟

۷. گاهی علاوه بر دو دسته بالا یک دسته سومی هم وجود دارد. در مورد این دسته (validation) توضیح دهید.

[2] numerical

[3] categorical

آموزش، ارزیابی و تنظیم

در این فاز از پروژه، سه مدل بر پایه Decision Trees، K-Nearest-Neighbours و Logistic Regression با استفاده از کتابخانه scikit learn پیاده سازی می‌کنید. سپس هایپرپارامترها را تغییر دهید و مدل را بهینه کنید. بهینه‌سازی مدلها به این منظور است که خطا کمینه شود اما overfitting رخ ندهد.

برای KNN تغییر تعداد همسایه‌ها کافی‌ست.

۱. دقت هر مدل را بر اساس confusion matrix رسم شده بدست آورید و نتایج را توضیح دهید.

۲. برای مدل‌هایی که پارامترهای زیادی دارند با کمک تابع GridSearchCV، مقادیر بهینه برای پارامترها را بدست آورید.

۳. در مورد underfitting و overfitting تحقیق کنید. آیا در مدل‌های شما این پدیده‌ها رخ دادند؟

۴. سعی کنید برخی از پیش پردازش‌هایی که انجام دادید را تغییر دهید. تاثیر آنها بر دقت مدل‌هایتان را بررسی کنید.

روش‌های یادگیری جمعی

یادگیری گروهی به این معناست که پیشبینی نهایی را با تجمیع نتایج حاصل از چند مدل انجام دهیم. در این فاز به پیاده‌سازی و تحلیل نتایج مدل‌های Random Forest میپردازیم.

در این مدل، تعدادی Decision Tree ساخته میشود که هرکدام جداگانه و با ویژگی‌های متفاوت آموزش میبینند. سپس برای تخمین نهایی بین نتایج درخت‌ها نوعی رای‌گیری انجام میشود.

۱. در مورد حداقل دو عدد از هایپرپارامترهای این مدل مطالعه کنید و تاثیر تغییر این هایپرپارامترها را روی نتایجتان را با رسم نمودار و ذکر دقیق نتایج بسنجید.

۲. نتایج این مدل را با مدل Decision Tree مقایسه کنید. در مورد bias و variance و ارتباط بین آن ها مطالعه کنید. به نظر شما از نظر هر کدام از دو مورد bias و variance یک مدل تنها Decision Tree بهتر عمل میکند یا یک مدل تجمیعی Random Forest؟ آیا نتایجی که به دست‌آوردید با نظرتان مطابقت دارد؟

نکات پایانی

۱ .دقت کنید که هدف پروژه تحلیل نتایج است بنابراین از ابزارهای تحلیل داده مانند نمودارها استفاده کنید و توضیحات مربوط به هر بخش از پروژه را به طور خلاصه و در عین حال مفید در گزارش خود ذکر کنید. اگر در جایی ذکر شده مقایسه‌ای انجام دهید، حتما نتایج را دقیق ذکر کنید و سپس آنها را تحلیل و مقایسه کنید.

۲ .در همه‌ی بخش‌ها مجازید از متدهای کتابخانه‌ی Matplotlib، Seaborn، Scikit-Learn و Pandas استفاده کنید ولی باید اطلاعات لازم در مورد هر کاری که انجام میدهید را داشته باشید، در هنگام تحویل ممکن است در مورد هرکدام از شما سوال پرسیده شود.

۳ .نتایج و گزارش خود را در یک فایل فشرده با عنوان AI_CA4_<SID>.zip تحویل دهید. محتویات پوشه باید شامل فایل notebook، خروجی html و فایل‌های مورد نیاز برای اجرای آن باشد. توضیح و نمایش خروجی‌های خواسته شده بخشی از نمره این تمرین را تشکیل میدهد. از نمایش درست خروجی های مورد نیاز در فایل html مطمئن شوید.