بسمه تعالی

تمرین پیاده سازی 2

پیاده سازی درخت تصمیم برای تشخیص دیابت

- مهلت تحویل دوشنبه، ۱۳۹۹/۰۲/۲۲ ساعت ۲۳:۵۵
- مهلت ارسال قابل تغییر نیست.
- مواردی که بعد از تاریخ فوق و حداکثر تا تاریخ ۱۳۹۹/۰۲/۲۶ ساعت ۲۳:۵۵ ارسال شوند، با سقف نمره 80% ای بررسی خواهند شد و ارسال های بعد از تاریخ ۱۳۹۹/۰۲/۲۶ قابل قبول نبوده و نمره ای نخواهد داشت.
- انجام تمرین تک نفره است. لطفا به تنهایی انجام شود، در غیر اینصورت نمره منفی در نظر گرفته خواهد شد.
- کل محتوای ارسالی زیپ شود و نام فایل زیپ ارسالی IMP2_studentNumber باشد.
- محتوای ارسالی دارای راهنما (read me) جهت تسهیل اجرا باشد.
- زبان برنامه نویسی دلخواه است. (پیشنهاد: پایتون)
- موارد ارسال شده در تاریخی که بعدا مشخص می‌شود به صورت حضوری نیز تحویل گرفته خواهند شد (صرفا آنچه در CW طبق تاریخ های فوق تحویل داده شده است بعدا به صورت حضوری تست شده و توضیح داده می‌شود)
- تنها تکالیفی که به CW و قبل از مهلت ارسال، فرستاده می‌شوند بررسی خواهند شد.
- حداقل یک ساعت قبل از مهلت ارسال را احتیاطا هدف قرار دهید، تا مشکلات غیرقابل پیش بینی باعث عدم آپلود پاسخ ها در CW نشوند.

شرح:

در این تمرین به پیاده سازی درخت تصمیم می‌پردازیم.

در مرحله اول، درخت تصمیم برای داده های گسسته طبق شبه کد ارائه شده در اسلایدهای کلاس پیاده سازی می‌شود. برای تست پیاده سازی صورت گرفته، داده های ۱۲ گانه مثال رستوران مورد آزمایش قرار میگیرد (می‌توانید تمام ۱۲ داده را به عنوان مجموعه آموزشی در نظر بگیرید، بدون مجموعه آزمایشی مجزا). درختی که ایجاد میکنید را با درختی که در اسلایدها ارائه شده مقایسه کنید.

در مرحله دوم، داده های مربوط به تشخیص دیابت مورد استفاده قرار می‌گیرند. این داده ها در فایلی به نام diabetes.csv تحویل شده است. در این پایگاه داده نمونه هایی با ۸ ویژگی و یک خروجی باینری وجود دارد. هدف آن است که با کمک این ورودی های ۸ گانه وجود یا عدم وجود دیابت تشخیص داده شود.

ابتدا داده ها را به دو مجموعه آموزشی و آزمایشی تقسیم کنید (چند درصد برای آموزش و چند درصد برای آزمایش؟ مثلا ۸۰٪ و ۲۰٪ میتواند خوب باشد، ۵۰٪ و ۵۰٪ هم قابل تست است).

برای گسسته سازی ورودی های از نوع پیوسته (یا ورودی های دارای مقادیر خیلی زیاد) بازه های عددی در نظر بگیرید. ساده ترین ایده (که در این تمرین قابل قبول است) آن است که برای چنین ویژگی هایی، بازه مینیمم تا ماکزیمم اعداد **در مجموعه آموزشی** را به تعدادی بازه مساوی تقسیم کنید (چه تعداد؟ تعدادهای مختلف را آزمایش کنید) و دو بازه اضافی هم برای مقادیر کمتر از مینیمم و بیشتر از ماکزیمم در نظر بگیرید (زیرا ممکن است در داده های آزمایشی مقادیر کمتر از مینیمم و بیشتر از ماکزیمم هم وجود داشته باشد). ایده های بهتر برای گسسته سازی مانند مرتب سازی و انتخاب نقاط برش در هر گره از درخت بر اساس نمونه هایی که در آن گره حاضرند را نیز میتوانید امتحان کنید. همچنین میتوانید ایده های جدید و خلاقانه خود را آزمایش کنید و نتایج آن را با حالت های قبل (بازه های مساوی یا انتخاب نقاط برش بر حسب مرتب سازی) مقایسه کنید.
پیشنهاد میکنم ابتدا همان بازه های مساوی را پیاده کنید (چیزی از نمره را از دست نخواهید داد) و در صورتی که فرصت کردید سراغ ایده های بعدی بروید (نمره اضافی).

در نظر داشته باشید برای پیاده سازی درخت تصمیم نباید از توابع آماده استفاده کنید. لذا فرمول آنتروپی و ... را باید خودتان پیاده کنید. استفاده از توابع آماده برای بخش های بعدی بلامانع است (و حتی توصیه میشود) .
مثلا برای خواندن اکسل، احیانا نمایش گرافیکی خروجی درخت (که الزامی نیست)، نمایش دقت خروجی و ... .

آنچه تحویل داده میشود:

۱- کداجرایی برنامه با توضیحات لازم برای اجرا
۲- درختی که برای مرحله اول و دوم پیدا کرده اید (میتوانید گرافیکی نمایش دهید (به هر نحوی که میتوانید) یا به صورت Text با پروتکلی که توضیح میدهید و قابل فهم باشد (بشود فهمید در هر گره کدام ویژگی با چه مقادیری خروجی تست شده اند و زیر شاخه هایش کدامند و ...))
۳- در هر گره، کدام ویژگی تست میشود، مقدار دست‌آورد اطلاعات، آنتروپی، باقی مانده آنتروپی در زیرشاخه ها چقدر است.

۴- گزارشی کامل از مسیر انجام کار، چالشهایی که احتمالا مواجه شدید، اجراهایی که گرفتید و نتایجی که حاصل شده است. دقت در داده های آموزشی و آزمایشی چقدر بوده و چقدر تفاوت داشته؟ آیا بیش برازش داشته اید؟ ایده ای برای افزایش دقت دارید (حتی اگر پیاده نکرده باشید)؟ + کشف و شهود خاصی اگر داشتید!