# بـــه نـــام خـــدا


دانشکده مهندسی کامپیوتر

**هوش مصنوعی و سیستم های خبره**

**پروژه اول – درخت تصمیم**

**دکتر آرش عبدی**

**پاییز 1403**

**طراحان :**

**محمدصادق نعمت‌پور**

**نیایش خانی**

- **در صورت وجود هر گونه ابهام در سوالات تنها به طراح آن سوال پیام دهید.**
- **با توجه به تنظیم شدن ددلاین تمارین توسط خود شما امکان تمدید وجود ندارد.**

- **زبان برنامه‌نویسی و قالب تمپلیت پایتون است ولی برای تمرین‌های اول می‌توانید از C# نیز استفاده کنید.**
- **کل محتوای ارسالی را زیپ کرده و نام آن را شماره دانشجویی خود قرار دهید.**
- **نیازی به نوشتن داک نیست.**
- **به سوالات انتهایی با دقت و به صورت کامل پاسخ دهید.**
- **انجام تمرین تک نفره است. لطفا به تنهایی انجام شود، در غیر اینصورت نمره منفی در نظر گرفته خواهد شد.**
- **بدیهی است که نه قرار بر پر کردن هر صفحه سؤال است نه تک کلمه‌ای و انتزاعی مشوش از افکارتان، پس حتی اگر جواب شما بله یا خیر است دلیل یا توضیحی برای آن ارائه کنید (نه سخنان قصار گاندی نه قصه هزار و یک شب). معیار را دوستتان قرار بدهید که بعد خواندن از شما سوال نپرسد.**

**آیدی تلگرام طراحان :**

@msnp1381
@mainlynia

# سوالات تمرین

1. در روند پروژه با چه چالش‌هایی مواجه شدید؟

تایتانیک:

ستون هایی مثل سن و قیمت موارد مختلف زیادی داشت برای همین آنها را دسته بندی کردم تا به حالت کتگوری در بیاید. موارد دیگری هم که به صورت کتگوری ولی استرینگ بودند را هم با اعداد 0 به بالا جایگزین کردم مثل جنسیت.

کرونا:

ستون های اضافی زیادی داشت که شامل مقدار زیادی nan بودند برای همین آنها بهتر بود که حذف شوند زیرا پر کردن اشتباه آنها باعث ایجاد جهت گیری خلاف واقع میشد. آنها را حذف کردم و اعمال دیگر را دقیقا عین مورد بالا انجام دادم.

2. درخت به دست آمده برای هر کدام از دیتاست ها به چه صورت بوده است؟

تایتانیک:

کرونا:

3. دو معیار آنتروپی و Gini index را مقایسه کنید.

تایتانیک و کرونا:

ناخالصی جینی با استفاده از فرمول زیر محاسبه می شود:

$$GiniIndex = 1 - \sum_j p_j^2$$

جایی که pj احتمال کلاس j است.

ناخالصی جینی فرکانس برچسب گذاری اشتباه هر عنصر مجموعه داده را هنگامی که به طور تصادفی برچسب گذاری می شود اندازه گیری می کند.

حداقل مقدار شاخص جینی 0 است. این زمانی اتفاق می افتد که گره خالص باشد، به این معنی که تمام عناصر موجود در گره از یک کلاس منحصر به فرد هستند. بنابراین، این گره دوباره تقسیم نخواهد شد. بنابراین، تقسیم بهینه توسط ویژگی هایی با شاخص جینی کمتر انتخاب می شود. علاوه بر این، زمانی که احتمال دو کلاس یکسان باشد، حداکثر مقدار را دریافت می کند.

---

آنتروپی با استفاده از فرمول زیر محاسبه می شود:

$$Entropy = -\sum_j p_j \cdot log_2 \cdot p_j$$

جایی که مثل قبل pj احتمال کلاس j است.

آنتروپی معیاری از اطلاعات است که نشان دهنده بی نظمی ویژگی ها با هدفمان است. مشابه شاخص جینی، تقسیم بهینه توسط ویژگی با آنتروپی کمتر انتخاب می شود. زمانی حداکثر مقدار خود را به دست می آورد که احتمال دو کلاس یکسان باشد و یک گره زمانی خالص باشد که آنتروپی حداقل مقدار خود را داشته باشد که 0 است.

4. برای افزایش دقت چه ایده‌ای دارید؟

تایتانیک:

تغییرات در عمق درخت میتواند کمک کند. همچنین عوض کردن تعداد داده هاس تست و ترین نیز میتواند موثر باشد

کرونا:

تغییرات در عمق درخت میتواند کمک کند. همچنین عوض کردن تعداد داده هاس تست و ترین نیز میتواند موثر باشد

5. آیا بیش‌برازش داشته‌اید؟ توضیح دهید.

تایتانیک:

خیر

کرونا:

خیر

6. چه نکات و کارهایی پروژه شما را متمایز می‌کند؟

تایتانیک:

مصور سازی و حذف ستون ها و داده های اضافی بدون اینکه در نتیجه نهایی تاثیر بدی بگذارد

کرونا:

مصور سازی و حذف ستون ها و داده های اضافی بدون اینکه در نتیجه نهایی تاثیر بدی بگذارد

7. مفهوم cross-validation چیست و در چه مواقعی استفاده می‌شود؟

هدف از cross-validation آزمایش توانایی مدل برای پیش‌بینی داده‌های جدیدی است که در تخمین آن استفاده نشده‌اند، به منظور پیدا کردن مشکلاتی مانند بیش برازش یا سوگیری انتخاب و ارائه روشی در مورد چگونگی تعمیم مدل به یک مجموعه داده مستقل. به عنوان مثال، (یک مجموعه داده ناشناخته از یک مشکل واقعی).