چه در اسپاتیفای فارسی میگذرد؟ به زبان داده

فهرست مطالب

مقدمه

در این پروژه میخواهیم تحلیل داده و نیز چندین تسک پیش‌بینی را بر روی مجموعه‌داده‌ای از موسیقی‌های فارسی در اسپاتیفای[1] انجام بدهیم.
شما در این پروژه ابتدا با مجموعه‌داده و ویژگی‌های آن آشنا می‌شوید. سپس از شما خواسته می‌شود با آنچه از درس داده‌کاوی یاد گرفته‌اید، به پیش‌بینی بپردازید!
از آنجایی که تا کنون آشنایی لازم برای درگیر شدن با مجموعه‌داده‌های ناشناخته و تحلیل و بررسی آنها را دارید، سه تسک زیر برایتان تعریف می‌شود و شما آزادی عمل کامل در انجام آن دارید. تنها خروجی نهایی شما اهمیت دارد و معیار ارزیابی شما مدنظر قرار می‌گیرد. پس روی خلاقیت خودتون حساب ویژه باز کنید!

این تسک ها عبارت‌اند از:

- تحلیل و بررسی داده‌ها
- رگرسیون برای پیش‌بینی محبوبیت موسیقی
- دسته‌بندی موسیقی‌ها به سنتی و غیرسنتی

در ادامه، هر بخش به تفصیل توضیح داده شده است.

امیدواریم از انجام این پروژه لذت ببرید.

[1] Spotify

آشنایی با مجموعه داده‌ها

این مجموعه داده شامل ۱۰۶۳۲ موسیقی از ۶۹ هنرمند ایرانی است. ۳۲ ویژگی برای توصیف موسیقی ها وجود دارد. که توضیح تعدادی از آنها در ادامه آمده و بررسی سایر ویژگی‌ها بر عهده خودتان است.

- ★ Track_id
  - The ID that Spotify specified for the song track
- ★ Duration_ms
  - Duration of the track in milliseconds
- ★ Explicit
  - Having strong language; If the track is considered offensive or unsuitable for children
- ★ Track_name
  - Official name of the track
- ★ Artist_name
  - Artist name
- ★ Popularity
  - A value will be between 0 and 100, with 100 being the most popular, based on total number of plays the track has had and how recent those plays are.
- ★ Danceability
  - A value will be between 0 and 1, with 1 being the most danceable, based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity.
- ★ Energy
  - A value will be between 0 and 1, with 1 being the most energetic, representing a perceptual measure of intensity and activity.
- ★ Key
  - The key the track is in. Values between 0 and 11.
- ★ Loudness
  - Overall loudness of a track in decibels (dB) averaged across the entire track. Values between -60 dB and 0 dB, with 0 being the loudest.
- ★ Mode
  - Indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0.
- ★ Speechiness
  - A value between 0 and 1, with 1 being the most exclusively speech-like the recording.
- ★ Acousticness
  - A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.
- ★ Instrumentalness

- Predicts whether a track contains no vocals. The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content.

★ Liveness
- Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live.

★ Valence
- A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).

★ Tempo
- The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.

★ Time_signature
- An estimated overall time signature of a track. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure).

★ Key_name
- The key name the track is in. Integers in 'key' map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C♯/D♭, 2 = D, and so on.

★ Mode_name
- Indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived.

★ Key_mode
- The key name AND the mode name the track is in.

در اصطلاح تحلیل داده، EDA یا Exploratory data analysis به عنوان گام اول برای آشنایی با داده‌های ناشناخته است. با استفاده از این فرآیند درک عمیقتر داده ها و یادگیری ویژگی های مختلف داده ها میسر می‌شود که اغلب با ابزارهای بصری هستند. با طی این فرآیند شما احساس بهتری نسبت به داده های خود پیدا خواهید کرد و الگوهای مفیدی را در آنها پیدا می‌کنید.

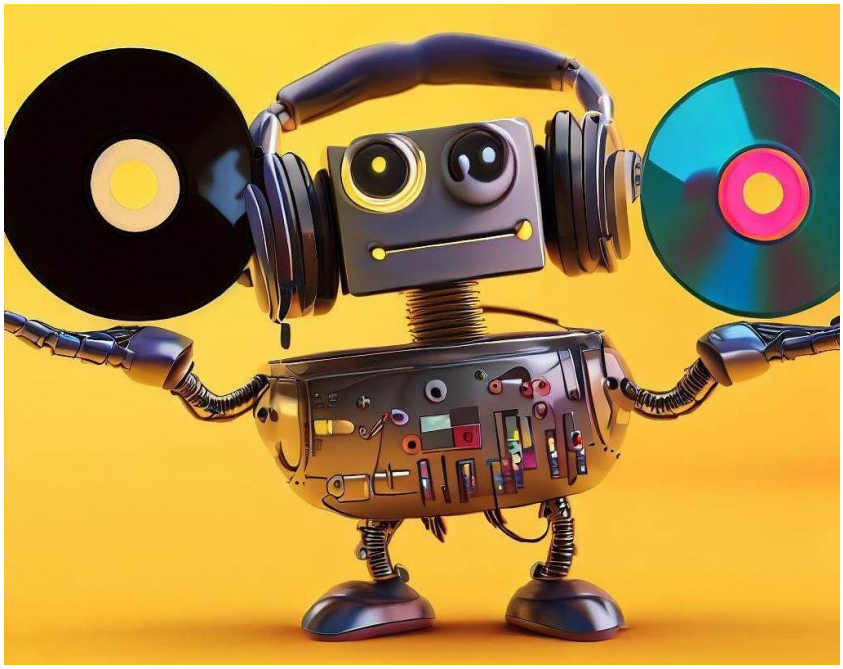
نوت‌بوک همراه پروژه را باز کنید و موارد خواسته شده را تکمیل کنید.
برای اینکه درک خوبی از داده‌ها پیدا کنید احتمالا نیاز به بررسی موارد بیشتری از آنچه خواسته شده است خواهید داشت.

توجه داشته باشید آنچه برای شما آورده شده است حداقل کار ممکن است. بسته به خلاقیت شما در تحلیل و بررسی داده‌ها نمره امتیازی دریافت میکنید.

رگرسیون

از آنجایی که هر پروژه یادگیری ماشین به دنبال هدفی است، ما نیز هدف انجام این گام را به صورت زیر تعریف میکنیم:

- میخواهیم الگو سلیقه موسیقی کاربران را از درون این دیتاست پیدا کنیم!



شکل ۱- تصویر تولید شده مدل DALL-E 2 به واسطه دستور "ربات پیش‌بینی میزان محبوبیت یک موسیقی"

در گام دوم این پروژه میخواهیم با استفاده از ویژگی‌های هر موسیقی و مدل‌های یادگیری ماشین رگرسیونی، پیش‌بینی کنیم هر موسیقی در مجموعه داده تست، چه میزان محبوبیت خواهد داشت.

همانطور که در قسمت قبل دیدید، میزان محبوبیت هر موسیقی به صورت زیر تعریف می‌شود.

A value will be between 0 and 100, with 100 being the most popular, based on total number of plays the track has had and how recent those plays are

الگوی سلیقه کاربران در مسئله یادگیری ماشین ما به رابطه بین ویژگی‌های هر موسیقی و نیز میزان محبوبیت آن تفسیر می‌شود. حال ما میخواهیم بر اساس ویژگی‌های یک موسیقی پیش‌بینی کنیم از چه میزان محبوبیتی برخوردار خواهد بود.

به طور دقیقتر، شما باید در ابتدا ویژگی‌های مدنظر خود را انتخاب کنید و بعد از انجام پیش‌پردازش بر روی آنها، ستون popularity در مجموعه داده را پیش‌بینی کنید.

همانطور که گفته شد در بررسی، انتخاب و پیش‌پردازش فیچرها و نیز تیون کردن مدل آزادی عمل دارید.
مدل رگرسیون شما باید حداقل معیارهای زیر را در پیش‌بینی میزان محبوبیت داشته باشد تا نمره کامل را دریافت کنید.
اگر مدل شما از این معیارهای پایه عملکرد بهتری داشت(خطای کمتر)، نمره امتیازی دریافت خواهید کرد.

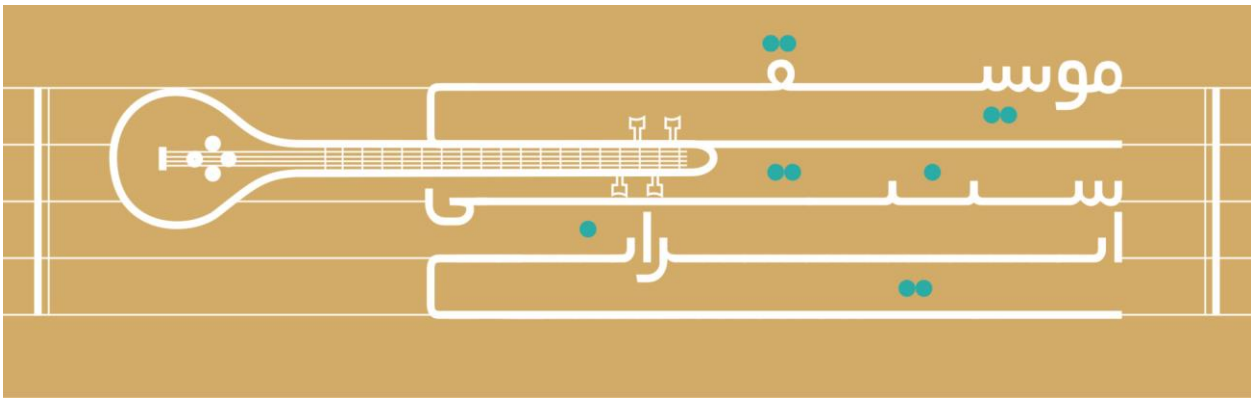
دقت کنید که حتما RANDOM_SEED = 42 باشه و ۳۰ درصد از دیتاست رو برای ارزیابی (دیتاست تست) استفاده کنین تا بتونین دقت مدلتون رو با مدل پایه به درستی مقایسه کنید.

معیار پایه:

Mean Squared Error = 8.8

دسته‌بندی

حال که به درک مناسبی از داده‌ها دست پیدا کرده‌اید و توانسته‌اید ارتباط میزان محبوبیت موسیقی‌ها را به واسطه ویژگی‌های آن به واسطه الگوریتم رگرسیون به دست آورید، زمان آن فرا رسیده است که به سراغ انجام تسک دسته‌بندی بروید.



موسیقی سنتی ایرانی یکی از کهن‌ترین سبک‌های موسیقایی است که از دوران باستان تا به امروز وجود داشته است. موسیقی سنتی ایرانی دارای ویژگی‌های متفاوت و متنوعی است که براساس شیوه اجرا دارای ویژگی‌های منحصربه‌فردی می‌باشد. این موسیقی دارای دستگاه‌های مختلفی مانند سه‌تار، تنبک، نی، کمانچه و... است که هر کدام دارای صدایی منحصر به فرد هستند. علاوه بر این، موسیقی سنتی ایرانی دارای ویژگی‌هایی مانند تکرار، تندی، کندی، تغییر فاز، تغییر مقام و... است.

شکل ۲- تصویر تولید شده مدل DALL-E 2 به واسطه دستور "ربات تشخیص موسیقی سنتی ایرانی از سایر موسیقی ها"

در این قسمت ما به دنبال یادگیری ویژگی‌های خاص و متمایز کننده موسیقی سنتی و توانایی جدا کردن این نوع از موسیقی از دیگر انواع موسیقی نظیر موسیقی‌های پاپ، فولکلور، هیپ هاپ و ... می‌باشیم. در این قسمت، شما می‌توانید بسته به نیازمندی خود، ویژگی‌هایی از موسیقی که به نظر شما تاثیر بیشتری در یادگیری این مدل خواهند داشت را انتخاب کنید و پس از پاک‌سازی این ویژگی‌ها، مدل پیشنهادی خود را بر روی آنها آموزش داده و در نهایت نتیجه نهایی را بر روی دادگان تست گزارش کنید.

همانطور که در قسمت‌های قبلی نیز مشاهده کردید، دیتاست فعلی ستونی برای مشخص کردن نوع سنتی بودن یا غیرسنتی بودن موسیقی ندارد. برای اضافه کردن این ستون به دیتاست فعلی، می‌بایست با استفاده از لیست خوانندگان سنتی که فایل نوتبوک برای شما مشخص شده است، این ستون را به دیتاست اضافه کنید. بنابراین تمامی موسیقی‌هایی که توسط یک خواننده سنتی خوانده شده است دارای مقدار ۱ و تمامی موسیقی‌های خوانندگان دیگر دارای مقدار ۰ به ازای این ستون خواهند بود.

توجه داشته باشید در این قسمت، دست شما برای انتخاب مدل آموزشی، معماری مدل و ویژگی‌های انتخابی جهت آموزش مدل کاملا باز است و شما می‌بایست با بررسی حالت‌های گوناگون و استفاده از شهود استخراج شده از داده‌ها، بهترین مدل را آموزش دهید تا بهترین عملکرد را بر روی دادگان تست از خود نشان دهد.

برای سادگی کار شما در این مرحله، یک تابع با نام `fit_and_eval` آماده شده است. شما می‌توانید تنها با پاس دادن مدل آموزشی (از نوع scikit-learn) و همچنین دیتاست (به صورت زوج x و y) به این تابع، مدل را بر روی ۷۰ درصد دیتاست آموزش داده و سپس نتیجه معیارهای ارزیابی مورد نیاز را بر روی ۳۰ درصد دیتا تست به دست آورید.

در این قسمت برای کسب نمره پایه این قسمت، می‌بایست بهترین مدل شما دارای حداقل امتیاز ۸۰ درصد در معیار $F_1\ Score$ به دست آورد. هر چه امتیاز شما در این قسمت از مقدار پایه بالاتر باشد، نمره امتیازی به شما تعلق خواهد گرفت.

ارزیابی

برای ارزیابی کیفیت مدل آموزش داده شما بر روی دیتاست تست، ۵ معیار زیر گزارش می‌شوند:

۱. معیار Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

۲. معیار Precision

$$Precision = \frac{TP}{TP + FP}$$

۳. معیار Recall

$$Recall = \frac{TP}{TP + FN}$$

۴. معیار $F_1\ Score$

$$F_1\ Score = \frac{2\ Recall \times Precision}{Recall + Precision}$$

۵. ماتریس درهم‌ریختگی

	Positive	Negative
True	TP	FN
False	FP	TN

نکات

- به همراه صورت پروژه، یک فایل نوت‌بوک پایتونی قرار داده شده است. برای تکمیل این پروژه، باید از همین نوت‌بوک استفاده کنید و پیاده‌سازی خود برای هر قسمت را به آن اضافه کنید. فرایند ارزیابی نوتبوک‌ها به صورت خودکار توسط autograder انجام خواهد شد. بنابراین پس از پایان تکمیل نوتبوک، حتما یک‌بار تمام سلول‌های نوتبوک را به ترتیب اجرا کرده تا از صحت عملکرد اجرای کد و تعریف متغیرها اطمینان حاصل کنید. سلول‌های مختص به autograder با کامنت `#autograde` مشخص شده‌اند. بنابراین به هیچ‌وجه این کامنت را تغییر ندهید.

- این پروژه به صورت انفرادی است. هم‌فکری و مشورت در رابطه با سوالات پروژه مشکلی ندارد با این‌حال در صورت تشخیص تقلب در پاسخ‌های افراد، نمره صفر برای آن‌ها در نظر گرفته خواهد شد.

- آخرین مهلت تحویل پروژه تا **ساعت ۲۳:۵۹ روز چهارشنبه ۷ تیر** می‌باشد. پس از این زمان، امکان تحویل با تاخیر پروژه وجود نخواهد داشت.

- در این پروژه استفاده از کتابخانه‌های Numpy، Pandas و Scikit-Learn محدودیتی ندارد و می‌توانید بدون مشکل از آن‌ها استفاده کنید. توجه داشته باشید امکان پیاده‌سازی الگوریتم‌ها از صفر نیز وجود دارد با این‌حال نمره امتیازی به این بخش تخصیص داده نخواهد شد.

- در صورت وجود ابهام در رابطه با پروژه می‌توانید سوالات خود را از طریق ایمیل درس و یا گروه متصل به کانال مطرح کنید. علاوه بر این، می توانید با آیدی‌های زیر در تلگرام در ارتباط باشید.

  @Amir_fal_01
  @AliAsad059

همیشه شاد و موفق باشید.

تیم تدریسیاری درس داده‌کاوی

بهار ۱۴۰۲