علی صالح ۹۷۲۲۲۰۵۳

پروژه سوم ماشین لرنینگ

---

۱. در خصوص کرنل های پرکاربرد روش SVM تحقیق کنید. به صورت کلی چرا ما از ایده کرنل در بحث SVM بهره میبریم . آیا میتوان در خصوص کرنل ها و استفاده ی آنها حکم کلی داد . به طور مثال بگوییم از کرنل RBF در این مواقع خاص استفاده میکنیم.

به طور کلی ایده کرنل ها برای این مطرح شد زیرا در خیلی از اوقات ما نمیتوانیم با یک هایپر پلین دیتا ها کلاس ها را جدا کنیم و هر جایی که هایپر پلین را می‌گذاریم با خطای بسیار زیادی مواجه می‌شویم.

در این موارد به کمک کرنل های مختلف ما دیتا هایمان را به بعد های بالاتر می‌بریم و در بعد های بالا تر سعی می‌کنیم کلاس ها را با یک هایپر پلین جدا کنیم.

کمکی بزرگی که کرنل ها به ما می‌کنند این است که بدون اینکه همه ی دیتا ها را به n بعد بالاتر ببریم میتوانیم به کمک فرمول کرنل ارتباط دیتا ها در بعد n ام را محاسبه کنیم و مقدار زیادی از حجم محاسبات ما کم می‌شود.

ما نمی‌توانیم یک حکم کلی در مورد استفاده از کرنل های مختلف بدهیم. ولی به طور مثال روش rbf کرنل پرکاربردی است زیرا دیتای ما را به بعد بی نهایت میبرد و جدا کردن دیتا ها با یک هایپر پلین در بعد بینهایت کار دشواری نیست.

۲ . قبلا با دیتاست کلاس بندی قیمت موبایل در ککگل کار کرده ایم . بر روی دیتاست ، روش SVM را اجرا کنید. (استفاده از پکیج ها همانند sklearn مجاز است.)

همانطور که در نوت بوک ضمیمه قابل مشاهده است ما روش svm را از طریق پکیج sklearn روی دیتا هایمان استفاده میکنیم و به دقت ۸۹ درصد میرسیم

۳ . برای سوال ۲ حداقل ۵ حالت مختلف از قبیل کرنل ها و پارامترها را بررسی کنید و نتایج آن را گزارش دهید

---

۵ آزمایش مختلف انجام میدهیم

- کرنل چند جمله ای درجه ۲

  دقت: ۴۷ درصد

- کرنل چند جمله ای درجه ۱۰

  دقت: ۲۵ درصد

- کرنل چند جمله ای دیفالت (بهینه ترین حالت)

  دقت: ۸۰ درصد

  درجه کرنل به دست آمده : ۳

- کرنل rbf

  دقت: ۸۹ درصد

- کرنل سیگموید

  دقت: ۹۱ درصد

۴ . برای سوال ۲ سعی کنید مبحث margin soft و margin hard را بررسی کنید و نتایج آن را گزارش دهید

ما وقتی از هارد مارجین استفاده میکنیم اجازه میس کلسیفیکیشن را نمی‌دهیم برای همین دقت ترین در این روش بالاتر است اما باعث میشد به نوعی دیتا ها اورفیت شود و از جنرال بودن مدل کم میکند برای همین دقت تست به نسبت سافت مارجین پایین تر است

۵ . مهندسی ویژگی یکی از بخش های مهم در فرایندهای علم داده میباشد. بر روی دیتاست موارد زیر را اجرا کنید .

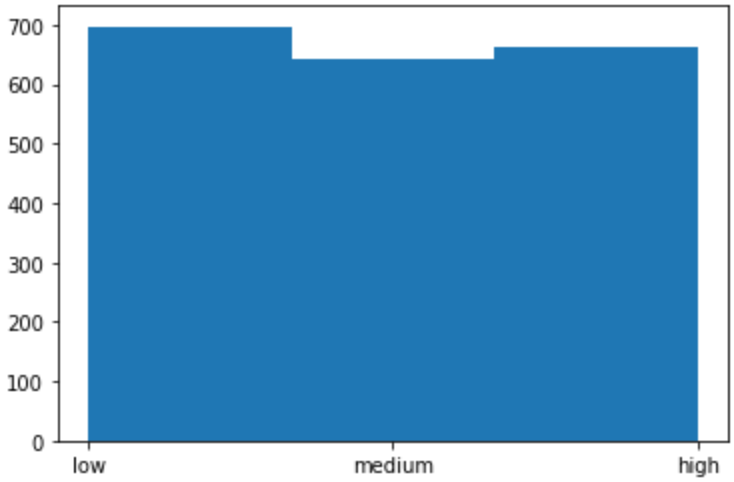
آ) بر روی فیچر power battery از روش binning استفاده کنید. (حداقل سه اندازه مختلف برای بین ها در نظر بگیرید و حتی سایز بین ها را نامساوی در نظر بگیرید.)

سه اندازه مختلف bin های مختلف را در نظر میگیریم

- سه قسمت مساوی در دامنه اعداد:

(501-1000), (1000-1499), (1499-1998)

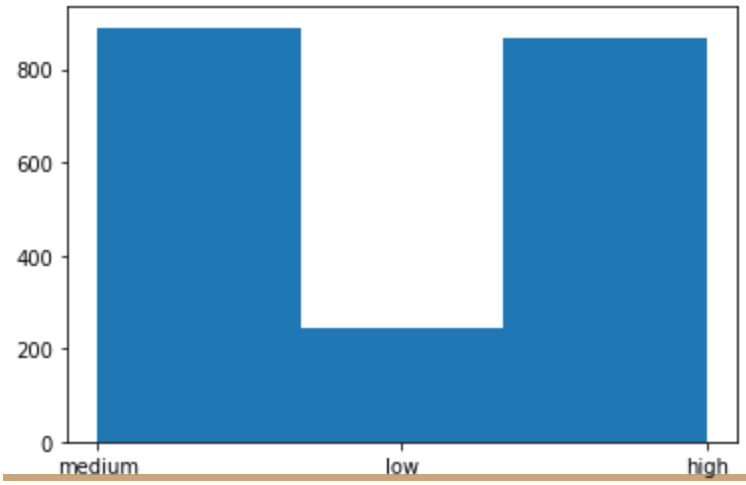
توزیع:



- دومین بین ها:

(0-666), (666-1333), (1333-2000)

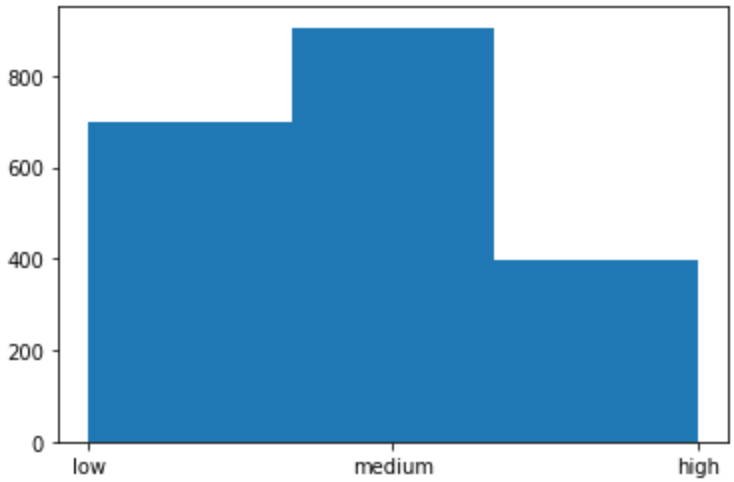
توزیع:



- سومین آزمایش:

(0-1000), (1000-1700), (1700-2000)

توزیع:



ب) بر فیچرهای کتگوریکال در دیتاست encoding hot one را اعمال کنید . چرا ما باید به صورت کلی از این کدگذاری بهره ببریم

این دیتاست فیچر کتگوریکال ندارد.

اما به طور کلی ما باید دیتا هایی که کتگوریکال هستند را به شکلی انکود کنیم تا بتوانیم به مدل های ماشین لرنینگی ورودی بدهیم.

یکی از این روش ها روش وان هات است.

در روش وان هات هر کدام از کتگوری ها یک ستون در نظر گرفته می شوند و اینکه هر دیتا مقدار این فیچرش کدام کتگوری است به این شکل نشان داده میشود که ستون بقیه ی کتگوری ها ۰ و ستون کتگوری آن دیتا مقدار ۱ دارد.

مزیت اصلی این انکود این است که باعث نمیشود دیتا های ما به سمت یک کتگوری خاص بایاس شوند.

از آنجایی که ما دیتای کتگوریکال نداریم با روش باینینگ با بین های شماره ۱ فیچر های کتگوریکال درست میکنیم و آن را با روش one hot انکود میکنیم.

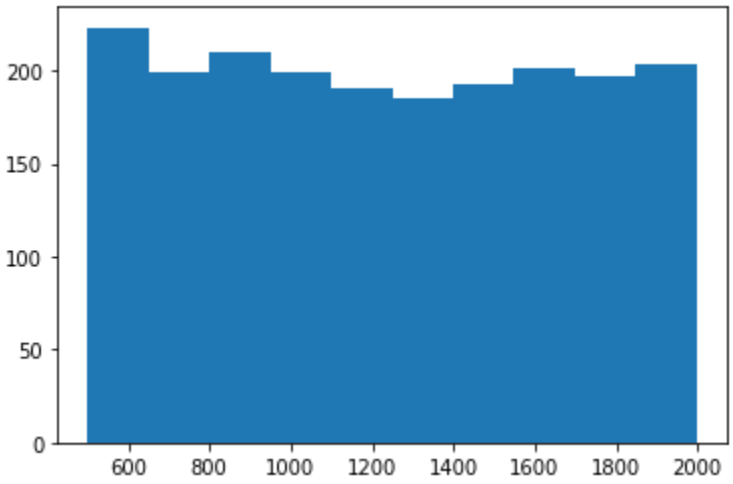
روی آن با روش svm فرایند کلاس بندی را انجام می دهیم و به دقت ۸۳ درصد میرسیم.

ج) بررسی کنید آیا استفاده از تبدیل هایی از قبیل transform log و یا تبدیل نمایی در اینجا کاربرد دارد . به صورت کلی چرا از این دست تبدیلات بهره میبریم . (در این بخش شما مجاز هستید اگر تبدیل دیگری را مناسب میدانید اعمال کنید این بخش نمره امتیازی برای شما خواهد داشت . حتما دلیل استفاده از تبدیل استفاده شده را بیان کنید)

معمولا استفاده از تبدیل ها برای این است که دیتا ها را با توزیع بهتری داشته باشیم

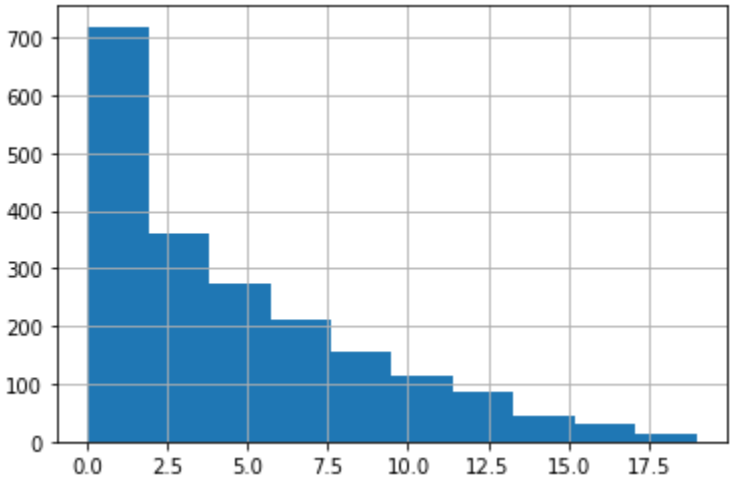
و تبدیل log transform برای دیتا هایی به کار میرود که توزیع نمایی دارند و با نگه داشتن log آن به جای مقدار اصلی میتوانیم ورودی بهتری به مدل ها بدهیم.

در اینجا توزیع فیچر battery power به صورت زیر است.

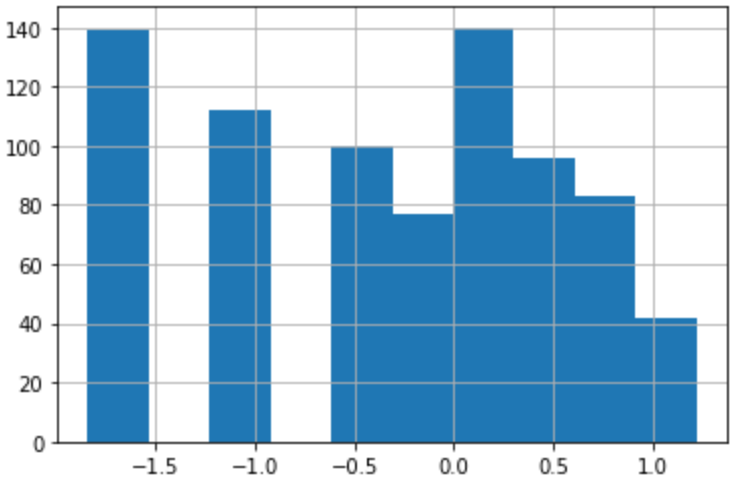


همانطور که مشخص است از یک توزیع نسبتا خطی برخوردار است و نیازی به تبدیل log transform نیست.

اما به طور مثال فیچر fc توزیعی به شکل زیر دارد.



که بعد از اسکیل کردن و انجام log transform به توزیع زیر میرسیم



که توزیع نرمال تری است.

د) یک فیچر جدید به نام مساحت یا حجم گوشی بسازید

فیچر مساحت را می‌سازیم و روی آن SVM میزنیم.

۶. برای هریک از حالت های سوال ۵ یک مدل SVM بسازید و بررسی کنید یک بار هم هر ۵ حالت را باهم اعمال کنید و مدل SVM روی آنها اجرا کنید . حاصل این مدل ها را گزارش کنید

در هر قسمت در صورت امکان مدل svm روی فیچرهای خروجی اجرا شده

یک بار برای همه ی پروسس ها با هم svm را اجرا میکنیم.

۷ . در خصوص الگوریتم های مختلف ساخت درخت تصمیم (همانند CART ، ID3 و...) تحقیق کنید . به صورت کلی تفاوت الگوریتم های مختلف ساخت درخت تصمیم در چیست ؟

ما در درخت تصمیم در هر مرحله یک فیچر را انتخاب کرده و بنا بر میانگین و واریانس آن فیچر یک مرزی را مشخص کرده و طبق آن تصمیم گیری میکنیم

تفاوت های الگوریتم های مختلف در روش انتخاب هر فیچر برای هر مرز و انتخاب مرز و همچنین الگوریتم ها در مواجهه با فیچر ها کتگوریکال و میسینگ ولیو ها متفاوت اند.

Features	*ID3*	*C4.5*	*CART*
Type of data	Categorical	Continuous and Categorical	continuous and nominal attributes data
Speed	Low	Faster than ID3	Average
Boosting	Not supported	Not supported	Supported
Pruning	No	Pre-pruning	Post pruning
Missing Values	Can't deal with	Can't deal with	Can deal with
Formula	Use information entropy and information Gain	Use split info and gain ratio	Use Gini diversity index

۸ . به دلخواه با استفاده از پکیج ها بر روی دیتاست مطرح شده یک درخت تصمیم بسازید

با پکیج sklearn و مدل tree.DecisionTreeClassifier دیتاست را کلاسبندی میکنیم و به دقت ۸۳ درصد میرسیم.

۹ . برای درخت تصمیم پارامتر های مختلف مورد ارزیابی قرار دهید . آیا عمق درخت و تعداد نمونه های موجود در هر هر گره تاثیری در عملکرد درخت تصمیم دارد ؟

سه پارامتر را عوض میکنیم

- عمق درخت: ماکسیمم عمق درخت را ۳ می‌گذاریم و دقت ۷۵ درصد میگیریم.
- تعداد فیچر ها: ماکسیمم تعداد فیچر ها را ۱۰ میگذاریم و دقت ۸۱ درصد میگیریم.
- تعداد نمونه های هر گره: این پارامتر را ۵ میگذاریم و دقت ۷۳ درصد میگیریم.

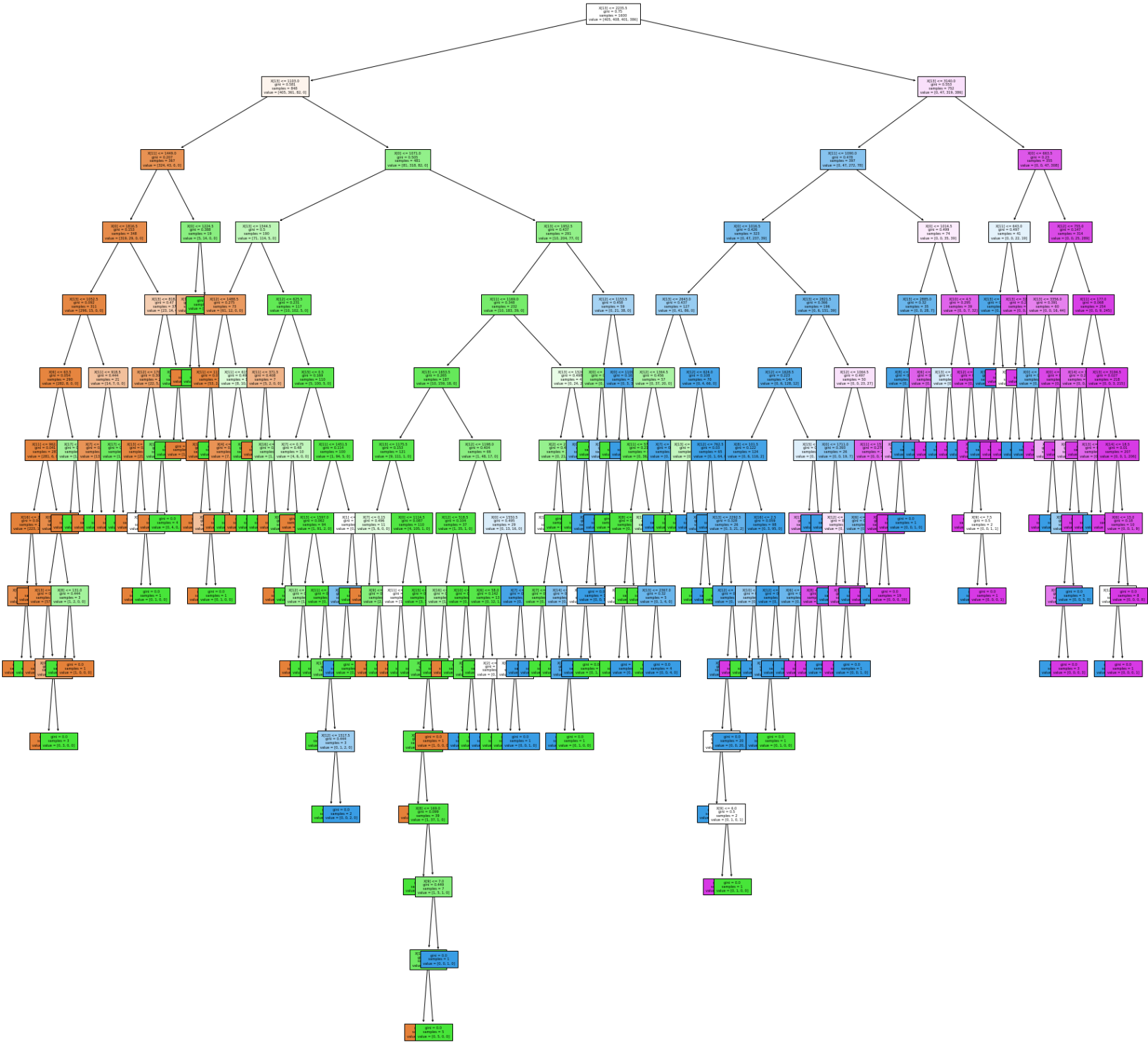
۰۱۰ در خصوص هرس کردن Pruning درخت تصمیم تحقیق کنید . چرا ما به بحث هرس کردن درخت تصمیم نیاز دارد و چه کمکی به ما میکند

هرس کردن یا pruning به این معنا است که ما قسمتی از درخت تصمیم گیریمان را حذف کنیم.

مثلا فرض کنید ما در لایه n ام درخت تصمیم گیری در مورد فیچر a تصمیم میگیریم. و در صورتی که کمتر از x باشد کلاس c1 را برمیگردانیم و در غیر این صورت کلاس c2.

اما اگر درخت را هرس کنیم باید بگوییم بدون توجه به فیچر a یکی از کلاس c1 یا c2 را برگردانیم.

این کار به وضوح از پیچیدگی مدل ما کم میکند و باعث جلوگیری از اورفیت می‌شود.



۱۱. (بخش امتیازی) سعی کنید این هرس کردن درخت در مدل خود اجرا کنید و بررسی کنید آیا این هرس کردن در نتایج شما تاثیر داشته است

۱۲ . بر روی دیتاست یک forest random اجرا کنید و نتایج با یک درخت تصمیم مقایسه کنید . بررسی کنید آیا نتایج درخت تصمیم بهتر است و یا forest random و چرا ؟

روی دیتا ها random forest را اجرا میکنیم و دقت ۸۷ درصد می‌گیریم که به نسبت درخت تصمیم ۴ درصد بهتر است.

رندوم فارست به دلیل اینکه یک متود انسبل است و از ترکیب زیر مجموعه ای از فیچر ها به صورت رندوم استفاده میکند با احتمال خیلی کمتری به سمت فیچر خاصی بایاس میشود و از احتمال اور فیت شدن آن هم بسیار کمتر است.

۱۳. تحقیق کنید چرا با وجود روش های جدید از قبیل یادگیری عمیق و شبکه عصبی ، هم چنان روشی مانند درخت تصمیم محبوب است ؟ (راهنمایی : میتوانید در خصوص بحث تفسیری پذیری این مدل ها صحبت کنید.)

زیرا اولا بسیار فرایندی که در این روش انجام میشود قابل درک تر از بقیه روش ها است و همه یک شهود بسیار بالایی نسبت به این روش دارند

مثلا خیلی قابل درک نیست در لایه های وسطی شبیه عصبی چه اتفاقی میافتد و فقط میدانیم لرن میشوند اما کاملا مشخص است که در هر نود درخت تصمیم چگونه تصمیم گیری میکنیم.

همچنین این روش به نسبت روشی مثل شبکه عصبی بسیار بار محاسباتی کمتری دارد این موارد باعث میشود که این روش همچنان روشی بسیار محبوب باشد.