به نام خدا

1- روش انتخاب ویژگی Forward Selection را پیاده سازی کنید . برای معیار انتخاب فیچر جدید در هر مرحله از AUCاستفاده کنید . **استفاده از پکیج مجاز نمی باشد** و باید این بخش را خودتان پیاده سازی کنید. برای سادگی این پیاده سازی موبایل ها را به دو گروه قیمت بالا (دو کلاس گران را ادغام کنید و یک کلاس در نظر بگیرید) و گروه با قیمت پایین تقسیم بندی کنید.
در روش انتخاب پیشرو ما از یک مجموعه تهی شروع کرده و در هرگام سعی داریم فیچر را به مجموعه فیچرهای انتخابی اضافه کنیم که AUCرا افرایش دهد.

2- با استفاده از کد پیاده سازی شده در بخش قبل به انتخاب ویژگی ها از فیچر ها بپردازید و سپس مدل لجستیک (با استفاده از پکیج) را بر روی فیچرهای انتخاب شده اجرا کنید و معیارهای f1-score، recall، precision را گزارش کنید.

3- با استفاده از الگوریتم PCAدر حالتی که تعداد Componentها با تعداد فیچرهای انتخابی حاصل روش انتخاب ویژگی پیشرو برابر باشد (یعنی اگر در سوال ۱شما با استفاده از انتخاب ویژگی پیشرو به طور مثال ۵فیچر را انتخاب کردید در الگوریتم PCA هم به عنوان آرگومان ورودی تعداد Componentرا ۵درج کنید) دیتاست را تغییر دهید.

4- با استفاده از دیتاست تغییر یافته در سوال قبلی و به کمک پکیج ها یک رگرسیون لجستیک را پیاده سازی کنید و معیار های f1-score، recall، precision را گزارش کنید.

5- مهندسی ویژگی یکی از بخش های مهم در فرایندهای یادگیری ماشین میباشد . بر روی دیتاست موارد زیر را اجرا کنید.

الف) بر روی فیچر battery power 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).

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

ج) بررسی کنید آیا استفاده از تبدیل هایی از قبیل log transform 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).

و) یک فیچر جدید به نام مساحت یا حجم گوشی بسازید.

6- برای هریک از حالت های سوال قبلی یک مدل رگرسیون لجستیک بسازید و بررسی کنید یکبار هم هر ۵ حالت را باهم اعمال کنید و مدل رگرسیون لجستیک روی آنها اجرا کنید. حاصل این مدل ها را گزارش کنید.

7- سعی کنید به دلخواه با استفاده از پکیج ها بر روی دیتاست مطرح شده یک درخت تصمیم بسازید و هرس کردن درخت ها که در سوالات تشریحی نیز مطرح شده را در مدل خود اجرا کنید و بررسی کنید آیا این هرس کردن در نتایج شما تاثیر داشته است.

8- روش انتخاب ویژگی Backward Selection را پیاده سازی کنید و با استفاده از فیچرهای انتخاب شده و کمک پکیج یک رگرسیون لجستیک را پیاده سازی کنید . معیار های f1-score، recall، precision را گزارش کنید و نتایج را با الگوریتم انتخاب ویژگی پیشرو در تسک 1 مقایسه کنید.

سوالات تحقیقی تشریحی

1- Bootstrapping چیست و چه تفاوتی با Cross Validation دارد؟ در کجا ها از Bootstrapping استفاده میشود؟

2- 5x2 cross validation را در یک پاراگراف توضیح دهید سپس بیان کنید در چه جاهایی استفاده از این روش میتواند مفید باشد؟

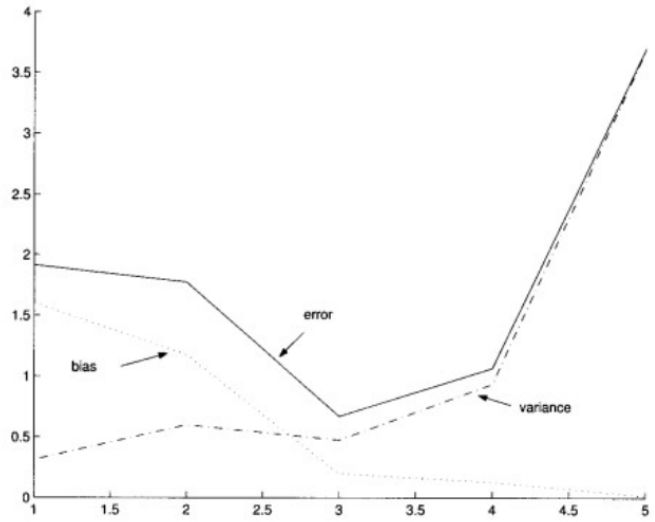
3- در خصوص الگوریتم های مختلف ساخت درخت تصمیم (همانند CART، ID3و ...) تحقیق کنید . به صورت کلی تفاوت الگوریتم های مختلف ساخت درخت تصمیم در چیست؟

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

5- 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؟

6- در خصوص هرس کردن Pruning درخت تصمیم تحقیق کنید . چرا ما به بحث هرس کردن درخت تصمیم نیاز داریم و چه کمکی به ما میکند؟

7- آیا میتوان با استفاده از روش Elbow با استفاده نموداری مشابه نمودار زیر که نمایان گر بایاس ، واریانس و مرتبه مدل است . بهترین مرتبه مدل برای پیچیدگی مدل را یافت ؟
به طور مثال با استفاده از روش Elbow میتوان در نظر گرفت که بر روی دیتاست ، مدلی از مرتبه ۳ جواب خوبی به ما میدهد . آیا همواره در تمامی مسائل و نه صرفا بحث تحلیلی میتوان اینگونه قضاوت کرد و مرتبه مناسب را به دست آورد ؟(راهنمایی برای پاسخ به این سوال توجه به مفهوم بایاس میتواند کمک کننده باشد).
(در شکل زیر خطا از حاصل جمع توان بایاس و واریانس به دست می آید).



شکل ۱: نمودار بایاس و واریانس بنا بر مرتبه های مختلف مدل

8- چگونه میتوان با استفاده از statistical significance tests به مقایسه مدل ها پرداخت ؟ (توضیح کامل) راهنمایی: Statistical Significance Tests for Comparing Models را جستجو کنید.

9- معیار Matthews Correlation Coefficient(MCC) چیست و در چه جاهایی استفاده میشود؟