Amirkabir University of Technology
(Tehran Polytechnic)

Applied Machine Learning Course

By Dr. Nazerfard
CE5501 | Spring 2024

Assignment (1)

Name: Esmaeil Khosravi

S_ID: 402131046

Email: es.khosravi@aut.ac.ir

سوال اول

a.

.ابتدا کتابخانه‌های مورد نیاز را فراخوانی می‌کنیم

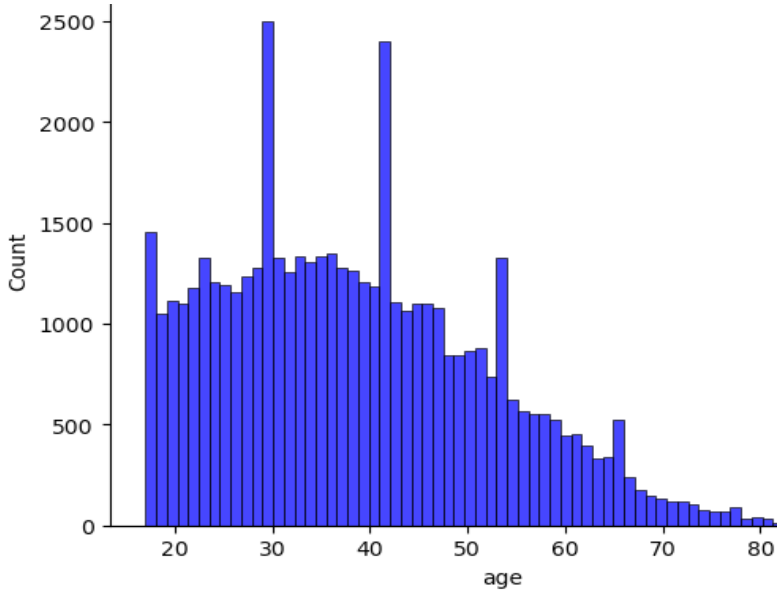
با متد read_csv از کتابخانه پانداس فایل csv را load میکنیم

با متد sample ۱۰ دیتای نمونه از دیتاست را چاپ می‌کنیم.

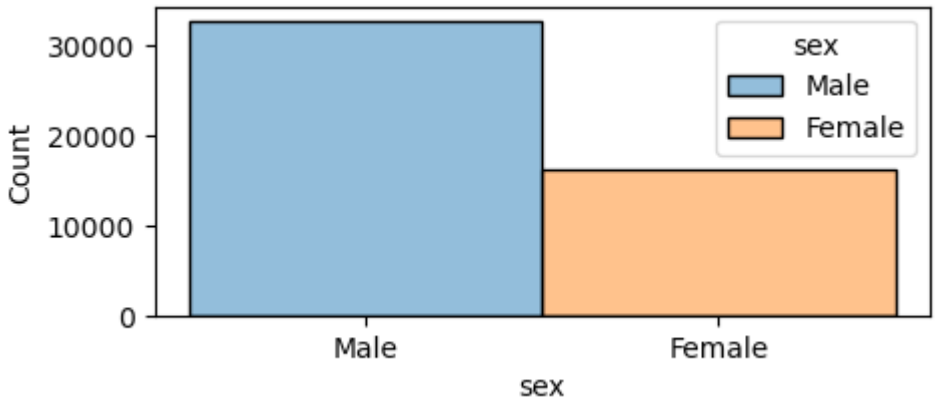
.b

با استفاده از کتابخانه seaborn نمودار هیستوگرام را برای صفت‌های دیتاست رسم می‌کنیم.

نمودارهیستوگرام سن نشان می‌دهد بیشتر سن افراد در محدوده ۲۰ تا ۴۰ سال است



.نمودار هیستوگرام جنسیت نشان می‌دهد بیشتر افراد مرد هستند



نمودار نژاد نشان می‌دهد بیشتر افراد سفید پوست و بعد از آن سیاه پوست هستند (به دلیل بزرگ بودن نمودار، نشان داده نشد)

.و نمودار هیستوگرام حقوق نشان می‌دهد بیشتر افراد کمتر از ۵۰ حقوق می‌گیرند

نمودار وضعیت تأهل نشان می‌دهد بیشتر افراد ازدواج کرده‌اند و بعد از آن بیشتر افراد مجردها می باشند

از نمودار ضرر و زیان می‌شود فهمید تعداد ضررکنندگان کم است اما میزان ضرر آنها زیاد است

---

c.

اگر تارگت پیش بینی میزان درآمد باشد رابطه این فیچر با بقیه به صورت زیر است

نمودار هیستوگرام ازدواج و حقوق نشان می‌دهد بیشتر افراد شاغل مجرد بوده و حقوق کمتر از ۵۰ دریافت می‌کنند

نمودار نوع شغل و حقوق نشان می‌دهد بیشتر شغل‌ها حقوق کمتر از ۵۰ دارند به جز self empl ها

مودار kdeplot نشان می‌دهد کسانی که حقوق بیشتر از ۵۰ دارند بازه سواد آنها بیشتر از 7.5 و کمتر از ۱۵ بوده است

تعداد کسانی که بازه سواد آنها بین 7.5 تا ۱۰ بوده از همه بیشتر است. رابطه بین ساعات کاری و تحصیلات نشان می‌دهد افراد دارای تحصیلات ۱۴ تا ۱۶ بیشترین حقوق را می‌گیرند.

d.

داده‌هایی که ناموجود هستند علامت سوال برای آنها ثبت شده، با متد replace این مقدار را با با مقدار مناسب جایگزین می‌کنیم

تعداد داده‌های ناموجود را با متد is null و sum حساب می‌کنیم

.سه ستون، شغل، دسته شغلی و ملیت دارای داده‌های ناموجود هستند

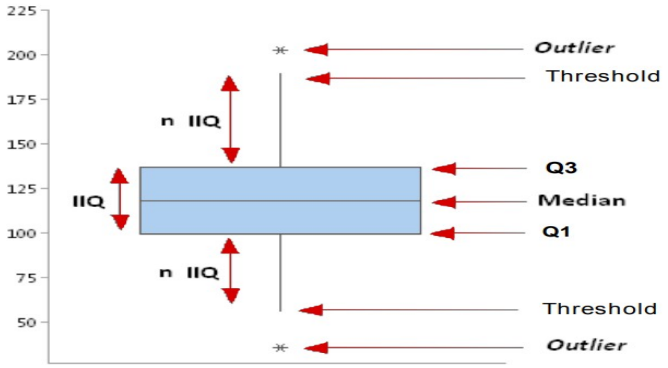
برای هندل کردن داده‌های ناموجود سه روش وجود دارد که هر کدام استفاده شده است

- حذف کردن سطر های با داده ناموجود
- حذف کردن ستون دارای داده ناموجود
- پیش بینی مقدار برای داده‌های ناموجود

با استفاده از متد duplicated فهمیدیم دیتا ست داده‌های تکراری ندارد.

یک تابع برای تشخیص بازه outlier بر اساس روش IQR[1] نوشته شده است این روش طبق تصویر زیر عمل میکند.

و بازه outlier برای برخی ستون‌ها حساب شـده و نمـودارboxplot بـرای نشـان دادن outlier استفاده شده است. در هر مورد تعداد داده‌های outlier برای ستون‌های عددی حساب شده و از دیتا ست حذف شده است.



---

e.

در مهندسی ویژگی ستون های‌edu، unnamed و rel به دلیل اینکه غیر کاربردی بوده‌اند حذف شده‌اند ستون جدیدی به نام سود profit که از اختلاف ضرر و و درآمد به دست آمده اضافه شده‌اند

سه ستون فوق با متد jlfit_transformاز کتابخانه skeat learn نرمال شده اند .

یک تابع به نام plot_normalized نوشته شده که وضعیت ستون‌های عددی را با نمودار، بعد از نرمال کردن نشان می‌دهد. در این تابع از متد stat.boxcox که از یک روش آماری برای تبدیل نمودار یک توزیع به توزیع نرمال است، استفاده میکند.

سوال دوم

[1] interquartile range

.a

با استفاده از کتابخانه open cv و متد imread عکس ها را خوانده و در یک لیست ذخیره میکنیم.

.b

سه عکس تصادفی با کتابخانه random و متد sample روی یک محور نشان داده شده است. عکس های رنگی به طور معمول 3 بعد دارند. بعد اول تعداد پیکسل ها در بعد عرض و بعد دوم تعداد پیکسل ها در بعد طول را نشان میدهد و بعد سوم تعداد کانال های رنگی را نشان میدهد تصاویر رنگی دارای 3 کانال رنگی قرمز، سبز و آبی (RGB) هستند و تصاویر خاکستری دارای یک کانال میباشند.

.c

استفاده از تصاویر رنگی در پردازش تصویر

مزایا

- تصاویر رنگی اطلاعات بیشتری نسبت به تصاویر خاکستری ارائه می‌دهند
- تصاویر رنگی برای تجزیه و تحلیل دقیق‌تر تصاویر مناسب اند و الگو ها در این تصاویر بهتر قابل تشخیص اند

معایب

- پردازش تصاویر رنگی پیچیده‌تر از پردازش تصاویر خاکستری است و هزینه بیشتری دارد
- به منابع پردازشی و حافظه بیشتری نیاز است.

.d

روشنایی

- به طور کلی، به میزان نور موجود در یک تصویر اشاره دارد
- تصویری با روشنایی مناسب، جزئیات را به طور واضح نمایش می‌دهد و نه خیلی تاریک و نه خیلی روشن است
- تنظیم روشنایی می‌تواند به بهبود وضوح تصویر، مخصوصاً در تصاویر کم نور، کمک کند
- افزایش بیش از حد روشنایی ممکن است جزئیات تصویر را از بین ببرد.

کنتراست

- به تفاوت بین روشن‌ترین و تیره‌ترین نقاط در یک تصویر اشاره دارد
- تصویری با کنتراست بالا، دارای جزئیات واضح و رنگ‌های زنده است
- تصویری با کنتراست پایین، مات و بی‌روح به نظر می‌رسد

- تنظیم کنتراست می‌تواند به برجسته کردن جزئیات تصویر و افزایش وضوح آن کمک کند

روشنایی و کنتراست را می توان با پارامتر های alpha و beta در هنگام تبدیل یک عکس رنگی به خاکستری در متد convertScaleAbs تنظیم نمود

الگوریتم‌های پردازش تصویر برای عملکرد صحیح به تصاویر با کیفیت مناسب نیاز دارند و تنظیم روشنایی و کنتراست،میتواند نویز تصاویر را کاهش دهد و دقت الگوریتم‌ها در پردازش تصاویر ارتقا می‌یابد. تصاویر با روشنایی و کنتراست مناسب، پردازش سریع‌تری دارند. تنظیم روشنایی و کنتراست، تصاویر را به یک سطح استاندارد از کیفیت می‌رساند. در مجموع، تنظیم کنتراست و روشنایی در مرحله پیش پردازش، گامی ضروری برای بهبود کیفیت تصاویر و افزایش دقت و سرعت الگوریتم‌های پردازش تصویر است.

---

e.

نرمالسازی تصویر

نرمالسازی تصویر، فرآیندی برای تبدیل مقادیر پیکسل‌های یک تصویر به محدوده مشخصی است. این کار به منظور استانداردسازی تصاویر و افزایش کارایی الگوریتم‌های پردازش تصویر انجام می‌شود.

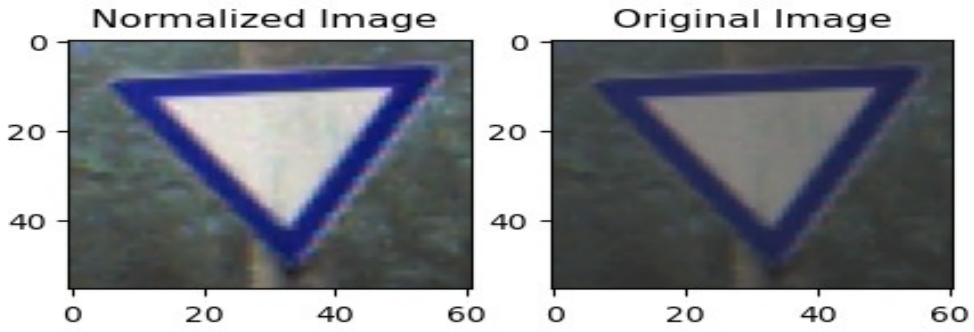
چالش های عدم نرمالسازی

- الگوریتم‌های پردازش تصویر برای عملکرد صحیح به تصاویر با کیفیت مناسب نیاز دارند.
- تصاویر بدون نرمالسازی، کیفیت پایینی دارند و می‌توانند دقت الگوریتم‌ها را به طور قابل توجهی کاهش دهند.
- پردازش تصاویر بدون نرمالسازی، به زمان و منابع محاسباتی بیشتری نیاز دارد
- تصاویر بدون نرمالسازی، در شرایط نوری مختلف عملکرد متفاوتی دارند و این باعث میشود دقت الگوریتم‌ها کاهش یابد.

تصاویر دیتاست با سه روش نرمال شده اند:

- روش min- max
- روش z-score
- استفاده از متد normalize از کتابخانه cv2

یک نمونه از تصویر نرمال شده در زیر پس از اعمال روش سوم:



سوال سوم

.a

ابتدا کتابخانه های موردنیاز را فراخوانی میکنیم

با دستور !gdown فایل را از طریق لینک داده شده دانلود کرده و بادستورات unrar و unzip فایل موردنظر را استخراج میکنیم و با دستورات open و متد read محتوای فایل را ذخیره میکنیم

.b

فایل خام حاوی دیتاست را میتوانیم با did جدا کنیم از طریق دستور متد split

حال محتوای ما به data object هایی تبدیل شده که هر data object نشان دهنده ی یک متن خبری از روزنامه همشهری است. نتیجه split یک لیست است که میتوان از آن برای تبدییل دیتاست به فرمت جدولی استفاده کرد. با یک حلقه for روی فایل split شده، روی هر data object ، با متد split محتوای متنی، تاریخ و دسته هر data object را جدا میکنیم. پس از آن با ذخیره کردن محتوای حلقه for در یک دیکشنری و دادن آن به متد DataFrame از کتابخانه pandas ، دیتاست به یک جدول تبدیل می شود

.c

با استفاده از متد pie از کتابخانه plot فراوانی دسته های خبری نشان داده شده است و طبق این نمودار دسته های سیاسی و اقتصادی و خارجی و akhar (!) و ورزشی جز 5 دسته اول از نظر فراوانی خبرها بوده است.

نمودار برای سال های خبر ها هم نشان می دهد بیشترین خبر مربوط به سال های 81 و بعد ازآن 76 بوده است نمودار ماه و روز توزیع معناداری را نشان نمیدهد و تقریبا فراوانی خبر ها در این زمان ها یکنواخت بوده است.

.d

پیش پردازش داده ها

برای پیش پردازش داده های متنی مراحل مختلفی باید طی شود برای این کار از کتاخانه hazm که مخصوص متن فارسی است استفاده شده است.

- حذف کلمات stop

- یک تابع به نام stop words نوشته شده است که متن هر نمونه دیتا را از دیتافرم گرفته و کلمات توقف را از آن فیلتر میکند و در ستون جدیدی به نام witout stop در دیتا فریم ذخیره میکند

- نرمال کردن متن
  - از متد normalize این کتابخانه برای نرمال کردن متن داخل یک تابع به نام normal استفاده شده است که ستون جدیدی به نام normal به دیتا فریم اضافه میکند که حاوی متن نرمال شده است.
- استخراج کلمات ((tokenization
  - یک تابع به نام tokenنوشته شده است که متن هر dataobject را از دیتافرم گرفته و کلمات را از هم جدا میکند و در ستون جدیدی به نام tokens در دیتا فریم ذخیره میکند

در نهایت، کلمات ستون without stop با متد join به هم متصل شده و ستونی جدیدی به نام o_content که بدون کلمات توقف می باشد اضافه شده است.

.e

TF-IDF به طور کلی به معنای تشخیص فراوانی کلمات کلیدی و مقایسه آن با دیگر متن ها می باشد.

.f

با کمک ماژول Tfidfvectorizer و متد fit_transform روش tf-idf روی دیتا اجرا شده است و مهم ترین کلمات متن استخراج شده اند .روش tf-idf بر روی ستون normaliezed دیتافریم اجرا شده و 5 کلمه مهم عبارتنداز: ایران، امیرملاحی، آئورینو، ازماندلا، تاکمبود.