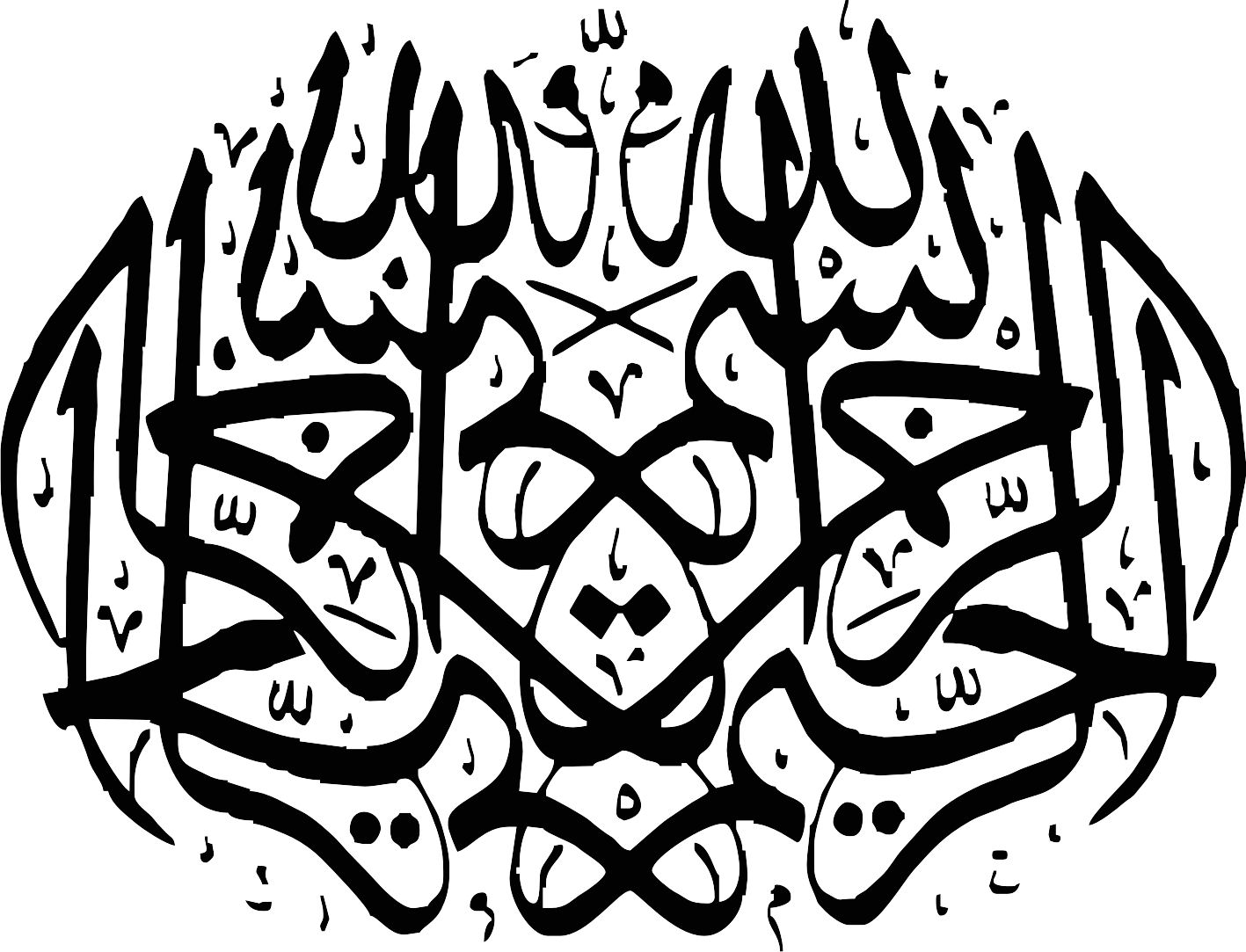


**گزارش تکلیف اول درس یادگیری ماشین کاربردی**

**استاد درس: دکتر ناظرفرد**

تهیه کننده: سید نیما محمودیان

شماره دانشجویی: 402125005



فهرست مطالب

[1-پاسخ سوال اول 1](#_Toc163208856)

[1-1-لود کردن دیتاست، نمایش ده ردیف به صورت تصادفی، و بررسی کلی داده‌ها 1](#_Toc163208857)

[2-1-مدیریت مقادیر گمشده 5](#_Toc163208858)

[3-1-مصور سازی داده‌ها 8](#_Toc163208859)

[4-1-مدیریت داده‌های پرت 14](#_Toc163208860)

[5-1-مهندسی ویژگی‌ها 16](#_Toc163208861)

[2-پیش‌پردازش تصویر 18](#_Toc163208862)

[1-2- gray scale کردن عکس‌ها 20](#_Toc163208863)

[2-2-تنظیم روشنایی و کنتراست تصاویر 21](#_Toc163208864)

[3-2-نرمال سازی تصاویر 23](#_Toc163208865)

[3-پیش‌پردازش متن 25](#_Toc163208866)

[1-3-خواندن داده‌ها از پیکره همشهری 25](#_Toc163208867)

[2-3-پیش‌پردازش متن‌ها 29](#_Toc163208868)

[3-3-TF-IDF 32](#_Toc163208869)

[4-3-مصورسازی داده‌ها 33](#_Toc163208870)

[4-پیوست 40](#_Toc163208871)

فهرست شکل‌ها

[شکل 1: وارد کردن کتابخانه‌های مورد نیاز 1](#_Toc163166210)

[شکل 2: حذف ستون‌های اضافه 1](#_Toc163166211)

[شکل 3: خواندن دیتاست و نمایش ده سطر تصادفی 2](#_Toc163166212)

[شکل 4: بررسی مجموع مقادیر ناموجود در هر ستون 2](#_Toc163166213)

[شکل 5: اطلاعات نمایش داده شده توسط متد info() 3](#_Toc163166214)

[شکل 6: تعداد مقادیر موجود در هر یک از متغیرهای دسته‌ای 3](#_Toc163166215)

[شکل 7: دسته‌بندی مجدد ستون nationality 4](#_Toc163166216)

[شکل 8: دسته‌بندی مجدد ستون edu 5](#_Toc163166217)

[شکل 9: جایگذاری ؟ با np.nan 6](#_Toc163166218)

[شکل 10: train test split 6](#_Toc163166219)

[شکل 11: پایپ‌لاین پیش پردازش داده‌های دسته‌ای 6](#_Toc163166220)

[شکل 12: نحوه تعریف کلاس ColumnTransformer() 7](#_Toc163166221)

[شکل 13: تعریف مدل و پایپ‌لاین اصلی 7](#_Toc163166222)

[شکل 14: شمای کلی پایپ‌لاین 8](#_Toc163166223)

[شکل 15:معیارسنجی مدل 8](#_Toc163166224)

[شکل 16: حذف سطرهای دارای مقادیر خالی 8](#_Toc163166225)

[شکل 17: کد رسم نمودار ستونی 9](#_Toc163166226)

[شکل 18: کد مربوط به نمودار جعبه‌ای 9](#_Toc163166227)

[شکل 19: نمودارهای ستونی رسم شده از داده‌های دسته‌ای 10](#_Toc163166228)

[شکل 20: نمودار جعبه‌ای از داده‌های عددی 11](#_Toc163166229)

[شکل 21: کد مربوط به هیستوگرام 11](#_Toc163166230)

[شکل 22: هیستوگرام‌های رسم شده 11](#_Toc163166231)

[شکل 23: نحوه انتخاب ستون‌های عددی و محاسبه ماتریس همبستگی 12](#_Toc163166232)

[شکل 24: هیت‌مپ رسم شده 12](#_Toc163166233)

[شکل 25: نحوه محاسبه همبستگی با ستون هدف 13](#_Toc163166234)

[شکل 26: نمودار میله‌ای همبستگی با ستون هدف 13](#_Toc163166235)

[شکل 27: نحوه محاسبه هیت‌مپ برای داده‌های دسته‌ای 14](#_Toc163166236)

[شکل 28: هیت‌مپ ایجاد شده از ستون‌های دسته‌ای 15](#_Toc163166237)

[شکل 29: نحوه حذف سطرهایی که داده پرت دارند 15](#_Toc163166238)

[شکل 30: one hot encoding 16](#_Toc163166239)

[شکل 31: انتقال ستون هدف به انتهای دیتافریم 17](#_Toc163166240)

[شکل 32: پایپ‌لاین ارزیابی روش PCA 17](#_Toc163166241)

[شکل 33: ایجاد کلاس و پایپ‌لاین برای mutual information 18](#_Toc163166242)

[شکل 34: شمای کلی پایپ‌لاین ایجاد شده برای mutual information 18](#_Toc163166243)

[شکل 35: وارد کردن کتابخانه‌ها و تولید سه عدد تصادفی 19](#_Toc163166244)

[شکل 36: لود کردن تصاویر و ذخیره آنها در متغیرها 19](#_Toc163166245)

[شکل 37: نمایش ابعاد عکس‌ها 19](#_Toc163166246)

[شکل 38: تبدیل عکس‌ها به rgb و نمایش آنها 20](#_Toc163166247)

[شکل 39: gary scale کردن عکس‌ها 22](#_Toc163166248)

[شکل 40: تابع تنظیم کنتراست و روشنایی 23](#_Toc163166249)

[شکل 41: تنظیم روشنایی و کنتراست شکل اول 24](#_Toc163166250)

[شکل 42: تابع نرمال کننده عکس 25](#_Toc163166251)

[شکل 43: نرمال کردن عکس با استفاده از تابع 25](#_Toc163166252)

[شکل 44: کتابخانه‌های استفاده شده در سوال سوم 26](#_Toc163166253)

[شکل 45: جداسازی رکوردها و قرار دادن آنها در چهار فیلد 26](#_Toc163166254)

[شکل 46: تابع تقسیم کننده فایل به یازده بخش 27](#_Toc163166255)

[شکل 47: تمیز کردن متن‌ها و ذخیره آنها در اکسل 28](#_Toc163166256)

[شکل 48: تبدیل فایل‌های اکسل به یک دیتافریم 28](#_Toc163166257)

[شکل 49: پیدا کردن نحوه انکودینگ متن 29](#_Toc163166258)

[شکل 50: حذف و جایگذینی کاراکترها و type casting 29](#_Toc163166259)

[شکل 51: حذف علائم نگارشی 30](#_Toc163166260)

[شکل 52: نحوه حذف اعداد 30](#_Toc163166261)

[شکل 53: توکنایز کردن متن 31](#_Toc163166262)

[شکل 54: حذف stop wordsها 31](#_Toc163166263)

[شکل 55: نمایش 5 توکنی که بیشتری استفاده را داشته‌اند 31](#_Toc163166264)

[شکل 56: تبدیل توکن‌ها به متن و نرمال سازی متن 32](#_Toc163166265)

[شکل 57: اجرای TF-IDF 32](#_Toc163166266)

[شکل 58: نمایش 5 کلمه مهم هر متن 33](#_Toc163166267)

[شکل 59: نمایش 5 کلمه مهم در همه متون 33](#_Toc163166268)

[شکل 60: فرایند ایجاد wordcloud 34](#_Toc163166269)

[شکل 61: ابر کلمه تولید شده 35](#_Toc163166270)

[شکل 62: انجام تحلیل عواطف با استفاده از کتابخانه polyglot 36](#_Toc163166271)

[شکل 63: نمودار سری زمانی 37](#_Toc163166272)

[شکل 64: نمودار میله‌ای تعداد متن‌های مثبت، خنثی و منفی 37](#_Toc163166273)

[شکل 65: نمودار سهم هر دسته از قطبیت‌ها از کل متون 38](#_Toc163166274)

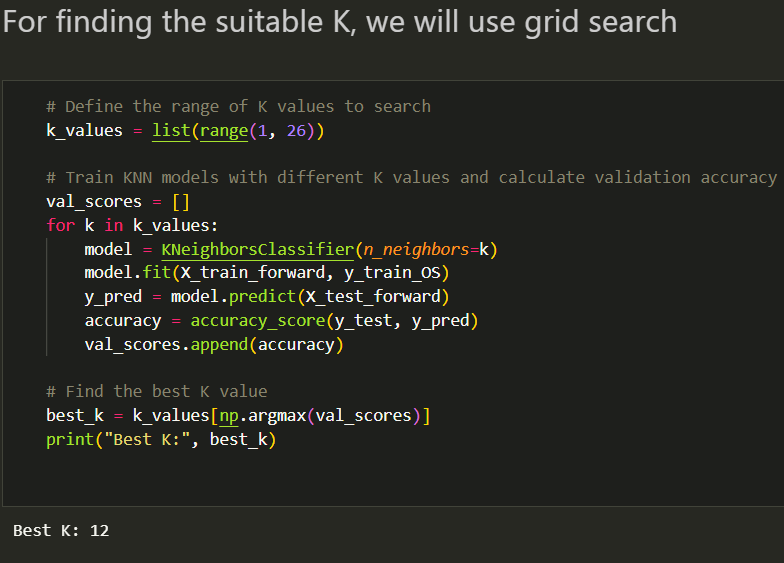
[شکل 66: هسیتوگرام میزان قطبیت 38](#_Toc163166275)

[شکل 67: نمودار تعداد اخبار مثبت، خنثی و منفی در طول زمان 39](#_Toc163166276)

[شکل 68: نمودار میله‌ای تعداد متون در بیست دسته اول موضوعات 40](#_Toc163166277)

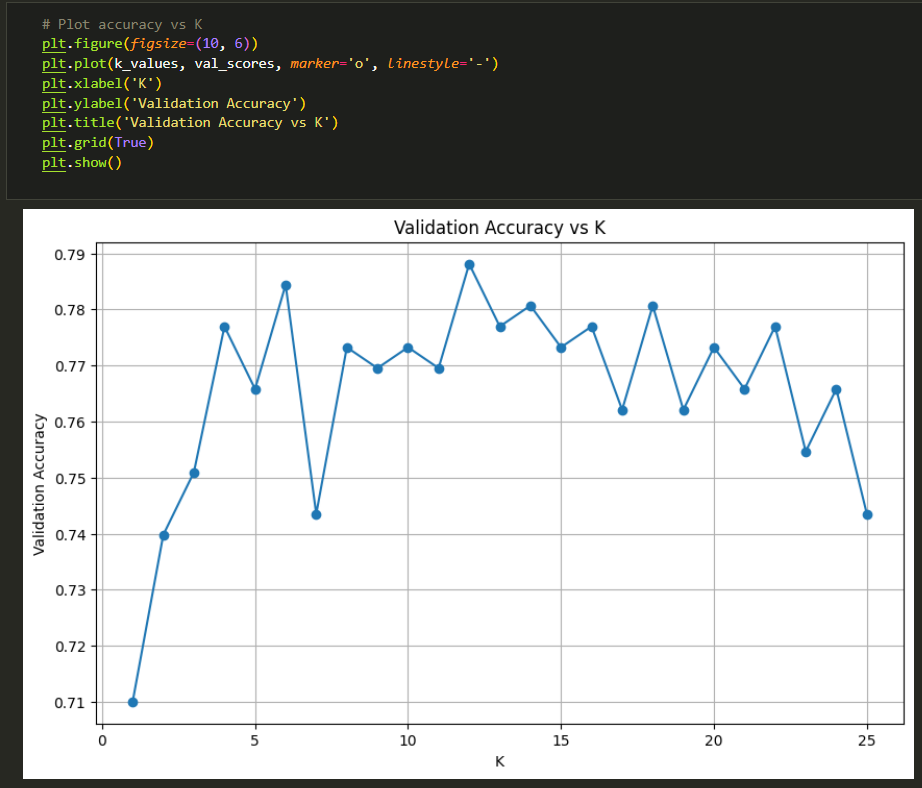
در این بخش به اجرای الگوریتم KNN بر روی دیتاست جدید می‌پردازیم. دیتاست حاضر استاندارد شده، 65 درصد داده‌ها برای آموزش جدا شده‌اند و تعداد لیبل‌های صفر و یک در آن برابر است.

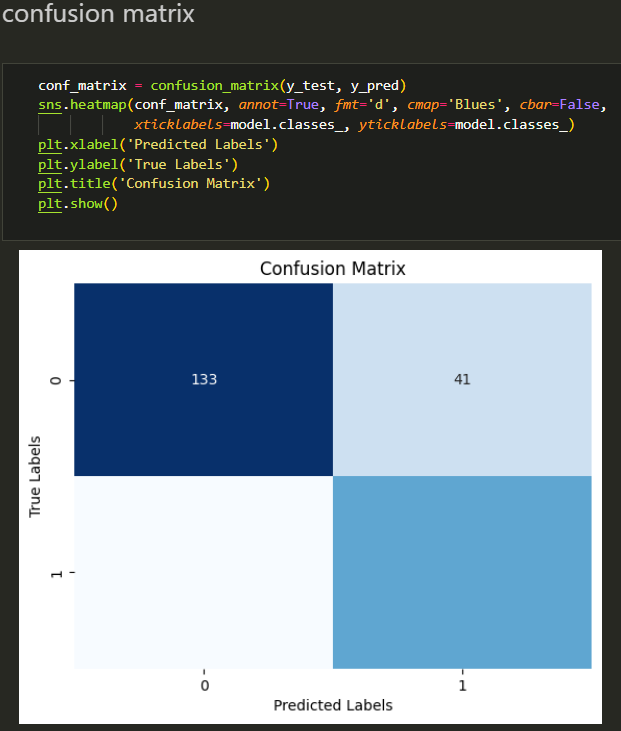
برای پیدا کردن K مناسب ابتدا محدوده‌ی مناسب برای K را تعریف می‌کنیم. حد بالای مقدار K برابر با است. 652 داده آموزشی داریم، ریشه دوم این عدد برابر با 25.53 است. بنابراین یک حد بالای مناسب برای جست و جو را می‌توان در نظر گرفت. برای پیدا کردن K مناسب از یک حلقه استفاده می‌کنیم. به طوری که ابتدا K را برابر 1 قرار می‌دهیم، مدلی را با این پارامتر آموزش می‌دهیم و تست می‌کنیم، مقدار دقت را محاسبه می‌کنیم و مقدار به دست آمده را در یک لیست ذخیره می‌کنیم. سپس با استفاده از تابع argmax() از کتابخانه نامپای، K متناظر با بیشترین مقدار دقت را نمایش می‌دهیم. شکل () کدی که وظیفه پیدا کردن K را به عهده دارد را نمایش می‌دهد.

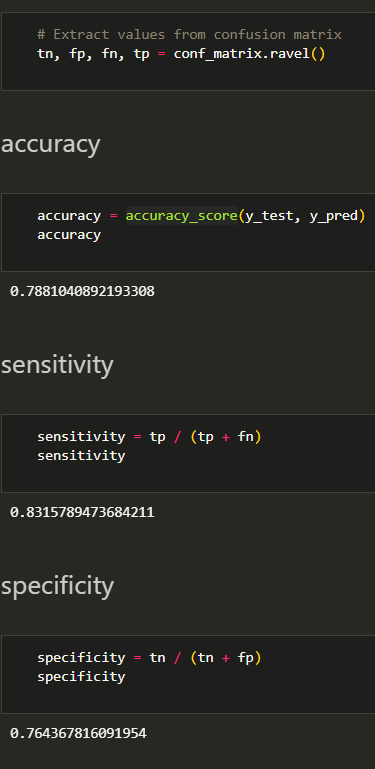


با توجه به بررسی صورت گرفته، بهترین دقت را دارد. همچنین نمودار دقت بر اساس K را نیز رسم می‌کنیم. شکل () این نمودار را نمایش می‌دهد.

حال بار دیگر با پارامتر و همین دیتاست، مدل را آموزش و ارزیابی می‌کنیم. سپس ماتریس درهم‌ریختگی، و معیارهای دقت، حساسیت و اختصاصیت را با استفاده از فرمول‌های مربوطه محاسبه می‌کنیم. برای محاسبه دقت از تابع پیش‌ساخته accuracy\_score() در سایکیت لرن استفاده می‌کنیم. شکل () ماتریس درهم ریختگی و شکل () محاسبه سه معیار را نمایش می‌دهد.

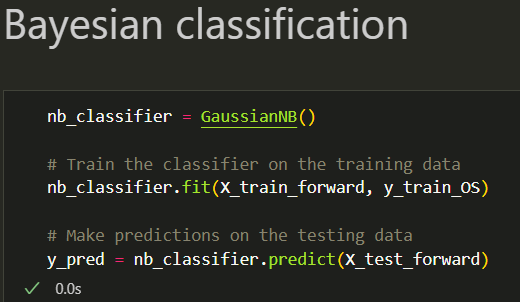






## دسته‌بندی بیزی

برای انجام دسته‌بندی بیزی از همان دیتاست بخش قبلی استفاده می‌کنیم. با استفاده از کلاس GaussianNB() از ماژول scikit-learn مدل بیزی را آموزش می‌دهیم. برای این منظور ابتدا یک شی به نام nb\_classifier از این کلاس می‌سازیم و با پاس دادن دیتاست ویژگی‌ها و ستون لیبل متناظر آنها به عنوان آرگومان به متد fit()، مدل دسته‌بندی بیزی راآموزش می‌دهیم. در ادامه با استفاده از متد predict() و پاس دادن دیتاست تست، پیش‌بینی‌های مدل از این دیتاست را به دست می‌آوریم و در متغیر y\_pred ذخیره می‌کنیم. شکل () نحوه انجام این عمل را نمایش می‌دهد.



سپس همانند دو بخش قبلی به محاسبه ماتریس درهم‌ریختگی و معیارهای ارزیابی می‌پردازیم. شکل () ماتریس درهم‌ریختگی روش دسته‌بندی بیز را نشان می‌دهد. همچنین شکل () مقدار سه معیار دقت حساسیت و اختصاصیت را برای روش دسته‌بندی بیز نشان می‌دهد.

