

**گزارش تکلیف اول درس یادگیری ماشین کاربردی**

**استاد درس: احمدی**

تهیه کنندگان:

سید نیما محمودیان

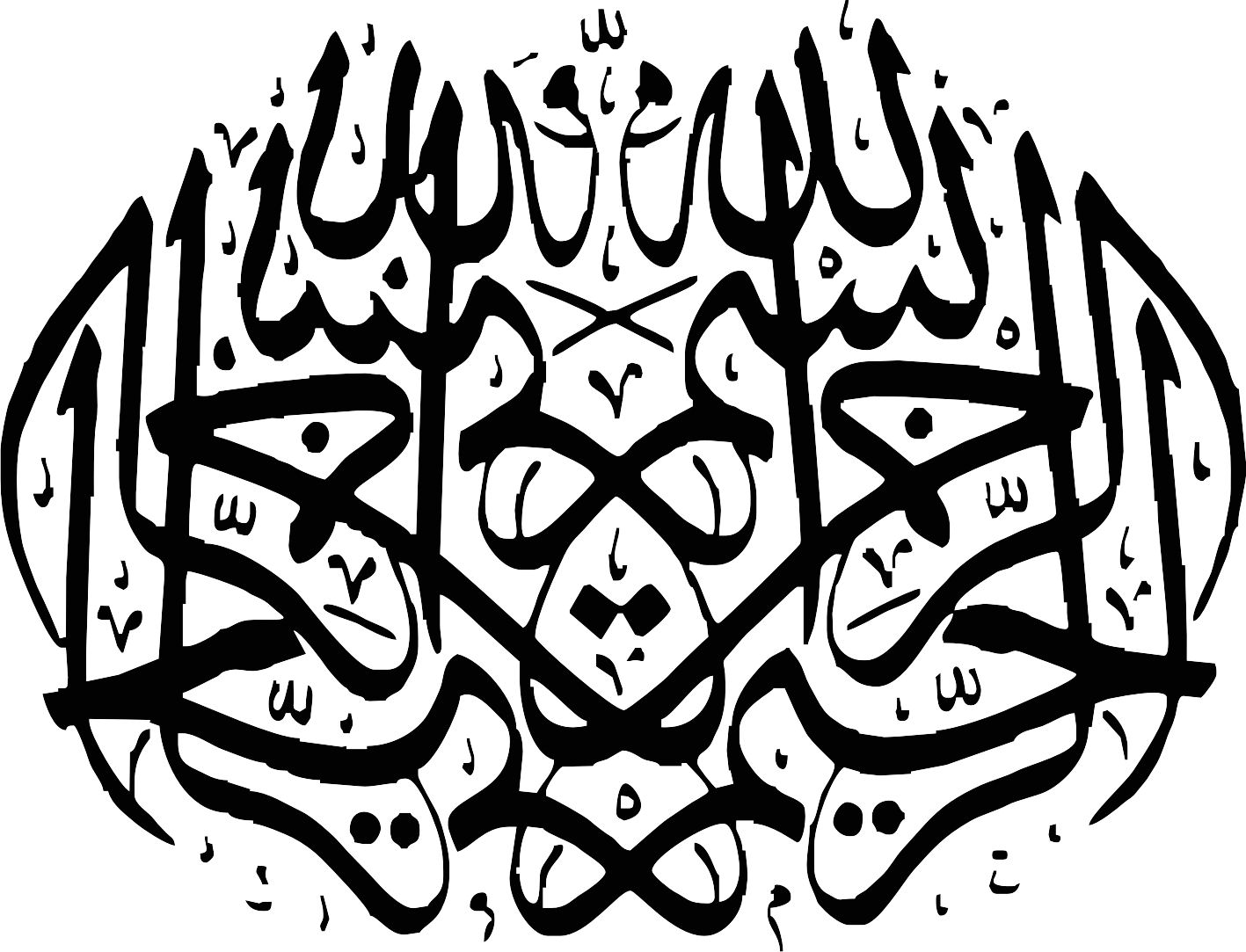
402125005

معصومه بهبهانی‌فر

402191003

محسن فراغه

401225002



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)

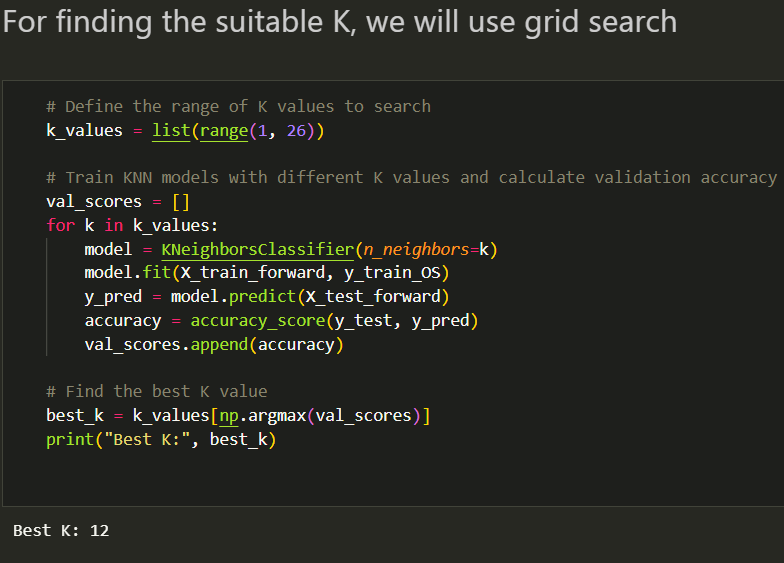
[4-پیوست 40](#_Toc163208871)

فهرست شکل‌ها

**No table of figures entries found.**

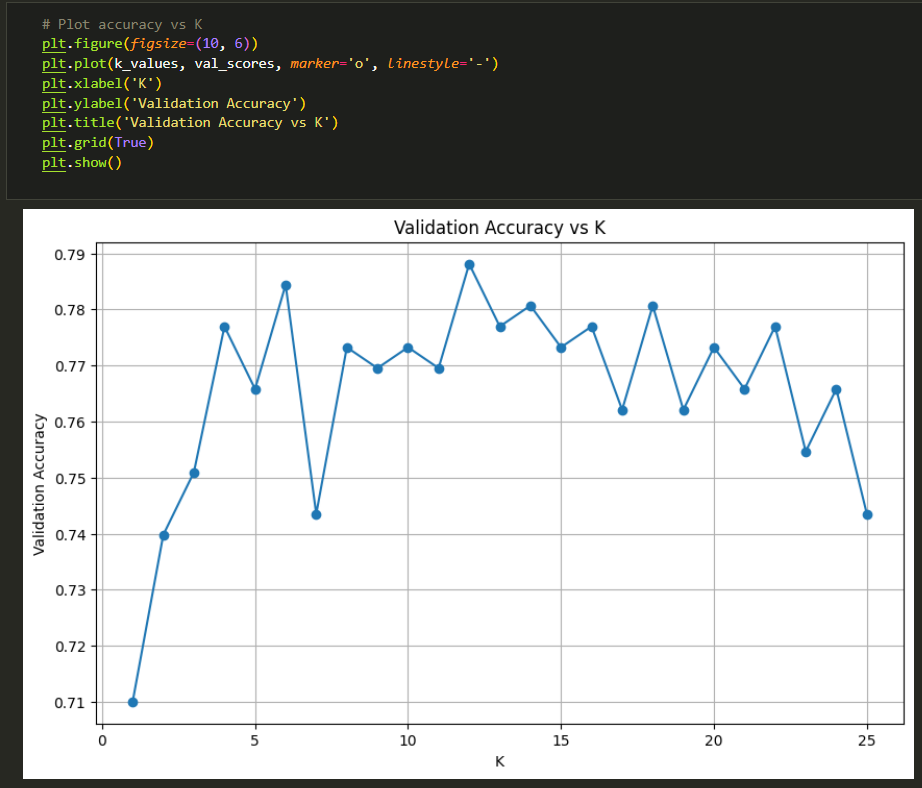
در این بخش به اجرای الگوریتم KNN بر روی دیتاست جدید می‌پردازیم. دیتاست حاضر استاندارد شده، 65 درصد داده‌ها برای آموزش جدا شده‌اند و تعداد لیبل‌های صفر و یک در آن برابر است.

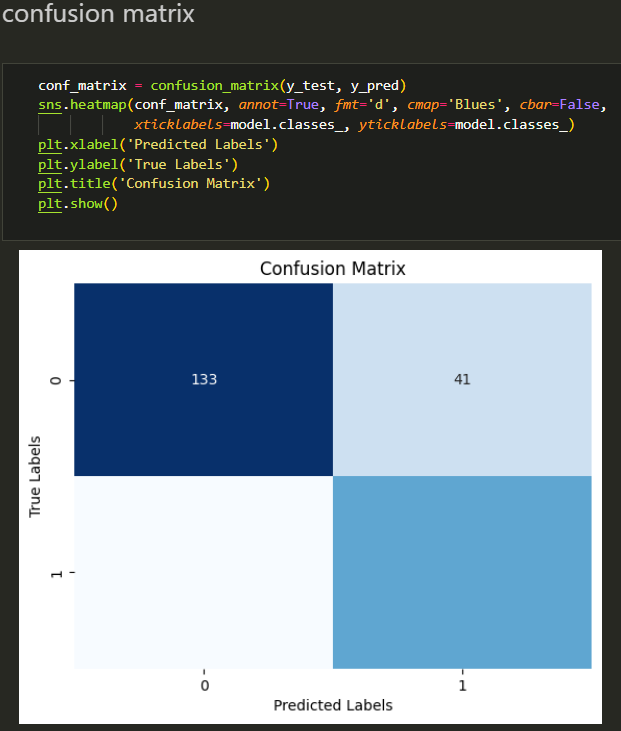
برای پیدا کردن K مناسب ابتدا محدوده‌ی مناسب برای K را تعریف می‌کنیم. حد بالای مقدار K برابر با است. 652 داده آموزشی داریم، ریشه دوم این عدد برابر با 25.53 است. بنابراین یک حد بالای مناسب برای جست و جو را می‌توان در نظر گرفت. برای پیدا کردن K مناسب از یک حلقه استفاده می‌کنیم. به طوری که ابتدا K را برابر 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.

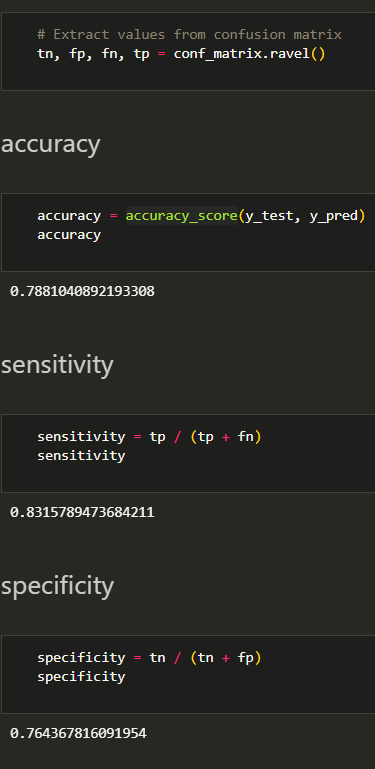


با توجه به بررسی صورت گرفته، بهترین دقت را دارد. همچنین نمودار دقت بر اساس K را نیز رسم می‌کنیم. شکل () این نمودار را نمایش می‌دهد.

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.

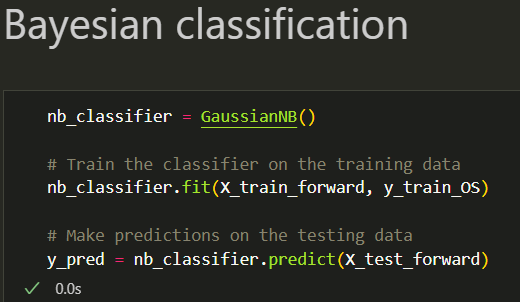






## دسته‌بندی بیزی

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.

