گزارش پروژه داده کاوی – پیاده سازی KNN

امیرحسین ابراهیمی 9813666013

1-1 یادگیری ماشین

یادگیری ماشین (Machine learning) به عنوان یکی از شاخه‌های وسیع و پرکاربرد هوش مصنوعی، به تنظیم و اکتشاف شیوه‌ها و الگوریتم‌هایی می‌پردازد که بر اساس آن‌ها رایانه‌ها و سامانه‌ها توانایی تعلّم و یادگیری پیدا می‌کنند. یادگیری ماشین کمک فراوانی به صرفه جویی در هزینه‌های عملیاتی و بهبود سرعت عمل تجزیه و تحلیل داده‌ها می‌کند. در حالت کلی یادگیری ماشین به دو حالت کلی "یادگیری نظارت شده" (Supervised Learning) و "یادگیری نظارت نشده" (Unsupervised Learning) تقسیم بندی می‌شود.

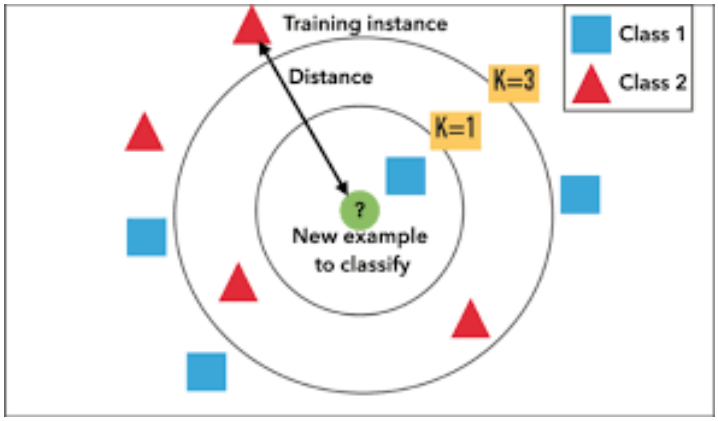
روش‌های یادگیری ماشین که به صورت نظارت شده عمل مینمایند، به این صورت کار می‌کنند که مجموعه‌ای از بردارهای ورودی مانند X و بردارهای خروجی متناظر با آنها مانند T داده می‌شود. هدف این است که ماشین قادر باشد با استفاده از این داده‌های آموزشی برای ورودی x جدید، t را پیش‌بینی نماید. از جمله روش‌های یادگیری نظارت شده می‌توان به روش‌های طبقه بندی (Classification) مانند شبکه‌های عصبی (Artificial Neural Network)، درخت تصمیم (Decision Tree)، بیزین ساده (Naïve Bayesian)، K-نزدیک ترین همسایگی (K Nearest Neighbor) و ماشین‌های بردار پشتیبان (Support Vector Machine) و روش‌های رگرسیون (Regression) مانند رگرسیون خطی (Linear Regression)، رگرسیون غیرخطی (Non Linear Regression)، رگرسیون بردار پشتیبان (Support Vector Regression) اشاره کرد.

اما در روش‌های یادگیری نظارت نشده، یادگیری ماشین تنها از طریق داده‌های ورودی انجام می‌شود و به این معنی است که مجموعه داده‌ها تنها شامل متغیرهای ورودی است و هیچ خروجی متناسبی با ورودی‌ها وجود ندارد. بنابراین در یادگیری نظارت نشده، الگوریتم یادگیری خودش به دنبال الگو و ساختار میان داده می‌گردد. در واقع یادگیری نظارت نشده روشی است که برای یافتن الگوهای (Pattern) میان داده‌ها استفاده می‌شود. به عبارت دیگر از طریق یادگیری نظارت نشده می‌توانیم ساختار و الگوهای پنهان میان داده‌ها را پیدا کنیم. از جمله روش‌های یادگیری نظارت نشده می‌توان به روش‌های خوشه‌بندی (Clustering) مانند K-Means، K-

Mediods، وDBSCAN روش‌های کاهش ابعاد (Dimensionality Reduction) مانند PCA و LDA اشـــاره کرد.

1-2 K نزدیکترین همسایگی

روش K نزدیک‌ترین همســایگی[1] یک روش یادگیری موردی اســت و از جمله ســاده‌ترین الگوریتم‌های یادگیری ماشـین می‌باشـد که به روش K همسـایه نزدیک نیز معروف اسـت. در این الگوریتم یک نمونه با رای اکثریت از همسایه‌هایش دسـته‌بندی می‌شـود و این نمونه در عمومی‌ترین کلاس مابین k همسـایه نزدیک تعیین می‌شـود. K یک مقدار مثبت صحیح و عموماً کوچک است. اگر k=1 باشد نمونه به سادگی در کلاس همسایگان نزدیکش تعیین می‌گردد. فرد بودن مقدار k مفید می‌باشـــد چون با این کار جلوی آراء برابر گرفته می‌شـــود. روش k همســـایه نزدیک، برای بســـیاری از روش‌ها کاربرد دارد، زیرا اثربخش، غیرپارامتریک و دارای پیاده‌ســـازی راحت می‌باشد. با این حال زمان دسته‌بندی‌اش طولانی است و یافتن مقدار k بهینه مشکل است. بهترین انتخاب از k، وابسته به داده‌ها می‌باشد به طور کلی مقدار بزرگ از k اثر نویز روی دسته‌بندی را کاهش می‌دهد، اما مرز مابین کلاس‌ها کمتر متمایز می‌شود.



شکل .Error! No text of specified style in document-**1: الگوریتم K نزدیکترین همسایگی (KNN)**

[1] K Nearest Neighbors (KNN)

شکل بالا مثالی از الگوریتم دسته‌بندی K نزدیک ترین همسایه، با استفاده از بردار ویژگی چندبعدی می‌باشد که مثلث‌ها کلاس اول و مربع‌ها کلاس دوم رانشـان می‌دهند. دایره کوچک در داخل دوایر، نمونه تسـتی را نشـان می‌دهد. حال اگر مقدار k=3 باشـد (یعنی 3 همسـایه‌ی نزدیک به نمونه)، نمونه تسـتی متعلق به کلاس مثلث و اگر k=5 باشد نمونه متعلق به کلاس مربع می‌باشد.

1-2-1 مراحل آموزش K همسایه نزدیک

مراحل آموزش K نزدیک‌ترین همسـایه به صـورت زیر می‌باشـد، این الگوریتم یک نمونه تسـتی را بر اسـاس k همسـایه نزدیک دسـته‌بندی می‌کند. نمونه‌های آموزشـی به عنوان بردارهایی در فضـای ویژگی چند بعدی مطرح می‌شـوند. فضـا به ناحیه‌هایی با نمونه‌های آموزشـی پارتیشـن‌بندی می‌شـود. یک نقطه در فضـا به کلاسـی تعلق می‌یابد که بیشـترین نقاط آموزشی متعلق به آن کلاس در داخل نزدیک‌ترین نمونه‌ی آموزشی به k در آن باشد. معمولا فاصـله‌ی اقلیدسـی یا تشـابه کسـینوسـی در این روش اسـتفاده می‌شـود. در فاز دسـته‌بندی KNN، نمونه تسـتی به عنوان یک بردار در فضـای ویژگی نمایش داده می‌شـود و فاصـله‌ی اقلیدسـی یا تشـابه کسـینوسـی بردار تستی با کل بردارهای آموزشی محاسبه می‌شود و نزدیکترین نمونه‌ی آموزشی به k انتخاب می‌شود. البته راه‌های زیادی برای دسته‌بندی بردار تستی وجود دارد و بنابراین الگوریتم k همسایه نزدیک کلاسیک، یک نمونه تستی را بر اساس بیشترین آراء از k همسایه‌ی نزدیکش تعیین می‌کند. سه فاکتور مهم در الگوریتم روش k همسایه‌ی نزدیک، به شرح زیر می‌باشد:

- معیار فاصله یا شباهت، برای پیداکردن k همسایه نزدیک استفاده می‌شود.
- K تعداد همسایه‌های نزدیک است.
- قانون تصمیم‌گیری برای تعیین (شناسایی) یک کلاس برای سند تستی از k همسایه نزدیک می‌باشد.

در الگوریتم K نزدیک‌ترین همسایگی، یک نمونه مطابق با رای اکثریت از همسایگان خود دسته بندی می شود، که نزدیکترین همسـایگان با اسـتفاده از تابع فاصـله[2] اندازه گیری شـده اسـت. اگر K = 1، سـپس مورد به کلاس نزدیکترین همسـایه اختصـاص داده می‌شـود. تابع فاصـله اقلیدسـی، منهتن و مینکوسـکی به ترتیب در روابط (**Error! No text of specified style in document.**1)، (2) و (**Error! No text of specified style in**

[2] Distance Function

document.) ذکر شده است.

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

Error! No) text of specified style in (1document.

$$\sum_{i=1}^{k}|x_i - y_i|$$

(2)

$$\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$$

Error! No) text of specified style (in document.

همچنین لازم به ذکر است که تمام سه اندازه گیری فاصله ذکر شده، تنها برای متغیرهای پیوسته معتبر است. در مورد متغیرهای گسسته باید از فاصله همینگ[3] استفاده شود که مسئله استانداردسازی متغیرهای عددی بین 0 و 1 را به وجود می آورد در حالی که مخلوطی از متغیرهای عددی و گسسته در مجموعه داده وجود دارد. فاصله نوع همینگ در رابطه (4**Error! No text of specified style in document.**) نمایش داده شده است.

$$D_H = \sum_{i=1}^{k}|x_i - y_i|$$

Error! No) text of specified style in (4document.

در روش K نزدیک‌ترین همسایه، انتخاب بهترین مقدار برای K بهتر است با اولین بررسی داده ها انجام شود. به طور کلی، یک مقدار بزرگ K دقیق تر است زیرا نویز کلی را کاهش می دهد اما هیچ تضمینی برای اعتبار آن وجود ندارد. اعتبار سنجی متقاطع[4] یکی دیگر از راه های به دست آوردن یک K خوب با استفاده از یک مجموعه داده مستقل برای اعتبار سنجی K می باشد. به لحاظ تجربه، K مطلوب برای بیشتر مجموعه داده ها بین 3 تا 10 است.

[3] Hamming distance
[4] Cross-Validation

آماده‌سازی داده‌ها

جهت مدل سازی با مدل KNN، تقسیم بندی داده ها برای مراحل آموزش (Train)، آزمایش (Test) به این صورت انجام گرفت که از 80 درصد داده مربوط به صورت تصادفی برای آموزش مدل KNN و 20 درصد باقیملنده برای آزمایش مدل KNN استفاده گردید. مجموعه داده‌های مورد استفاده در این تحقیق از بخش دیتاست‌های یادگیری ماشین دانشگاه کالیفرنیا آمریکا تهیه شده است و در پایگاه داده UCI[5] قابل دسترسی است. این مجموعه داده شامل 116 ردیف است که هر کدام 9 ویژگی دارند. 64 بیمار مبتلا به سرطان سینه و 52 بیمار سالم وجود دارد. اطلاعات مربوط به ویژگی‌های این مجموعه داده عبارتند از: سن[6]، شاخص توده بدنی[7]، گلوکز[8]، انسولین[9]، مدل ارزیابی هموستاتیک[10]، لپتین[11]، ادیپونکتین[12]، رزیستین[13]، پروتئین کموتاکسی مونوسیت یک[14].

متغیرهای ورودی به مدل KNN در جدول زیر ارائه شده است.

متغیرهای ورودی و خروجی مدل KNN

نام متغیر	
Age	**متغیرهای ورودی مدل**
BMI	
Glucose	
Insulin	
HOMA	
Leptin	
Adiponectin	
Resistin	
MCP.1	
Labels	**متغیرهای خروجی مدل**

[5] https://archive.ics.uci.edu/ml/datasets.php
[6] Age
[7] BMI
[8] Glucose
[9] Insulin
[10] homeostatic model assessment (HOMA)
[11] Leptin
[12] Adiponectin
[13] Resistin
[14] MCP.1

تقسیم‌بندی داده‌ها

در آموزش ماشین (Machine Learning) معمولاً داده‌ها را به دو قسمت تفکیک می‌کنند. مجموعه داده‌های آموزش و آزمایش. در این تحقیق از 80 درصد از مجموعه داده‌ها به عنوان داده‌های آموزش و 20 درصد باقی‌مانده به عنوان داده‌های آزمایش استفاده شده است.

داده‌های آموزشی (Training set): از این بخش از داده‌ها به منظور ایجاد و آموزش مدل‌ها و الگوریتم‌های مختلف یادگیری ماشین و برآورد پارامترهای آن استفاده می‌شود.

داده‌های آزمایشی (Test set): این قسمت از داده‌ها برای بررسی کارایی مدل‌ها و الگوریتم‌های مختلف یادگیری ماشین که در مرحله قبل آموزش دیده‌اند، استفاده می‌شود. اهمیت این بخش از داده‌ها در این نکته است که این مشاهدات شامل مقدارهای متغیرهای مستقل (xها) و پاسخی (y) هستند که در آموزش مدل‌های یادگیری ماشین به کار نرفته، ولی امکان مقایسه مقدار پیش‌بینی شده توسط مدل‌های یادگیری ماشین را با مقدار واقعی به ما می‌دهند؛ البته توجه داریم که این داده‌ها مدل را تحت تأثیر قرار نداده‌اند؛ پس در تعیین پارامترهای مدل نقشی نداشته و فقط برای ارزیابی مدل‌های یادگیری ماشین به کار می‌روند.

با توجه به تفکیکی که برای این دو گروه داده در نظر گرفته شد، مدل‌سازی فقط بر اساس بخش داده‌های آموزشی خواهد بود، ولی در روش اعتبارسنجی متقابل[15] که از این به بعد آن را به اختصار «CV» می‌نامیم، طی یک فرآیند تکرارشونده، قسمت داده‌های آموزشی (Training set) که به منظور مدل‌سازی به کار می‌رود، خود به دو بخش تفکیک می‌شود. در هر بار تکرار فرآیند CV، بخشی از داده‌ها برای آموزش و بخشی دیگر برای اعتبارسنجی[16] مدل به کار می‌رود. به این ترتیب این فرآیند یک روش بازنمونه‌گیری به منظور برآورد خطای مدل محسوب می‌شود.

باید توجه داشت که داده‌های آزمایشی در فرایند CV ممکن است در تکرار بعدی به عنوان داده‌های آموزشی به کار روند، در نتیجه، ماهیت آن‌ها با داده‌هایی که در قسمت قبل به عنوان داده‌های آزمایشی (Test set) معرفی شد، متفاوت است. شکل زیر به درک ماهیت داده‌های تست در فرآیند CV کمک می‌کند. مشخص است که داده‌های اعتبارسنجی بخشی از داده‌های آموزشی هستند و داده‌های آزمایشی نیز به عنوان بخشی مجزا از داده‌هایی

15. Cross Validation (CV)
16. Validation

آموزشی فرض شده‌اند. مراحل تکرار فرآیند CV نیز در تصویر به خوبی دیده می‌شود.

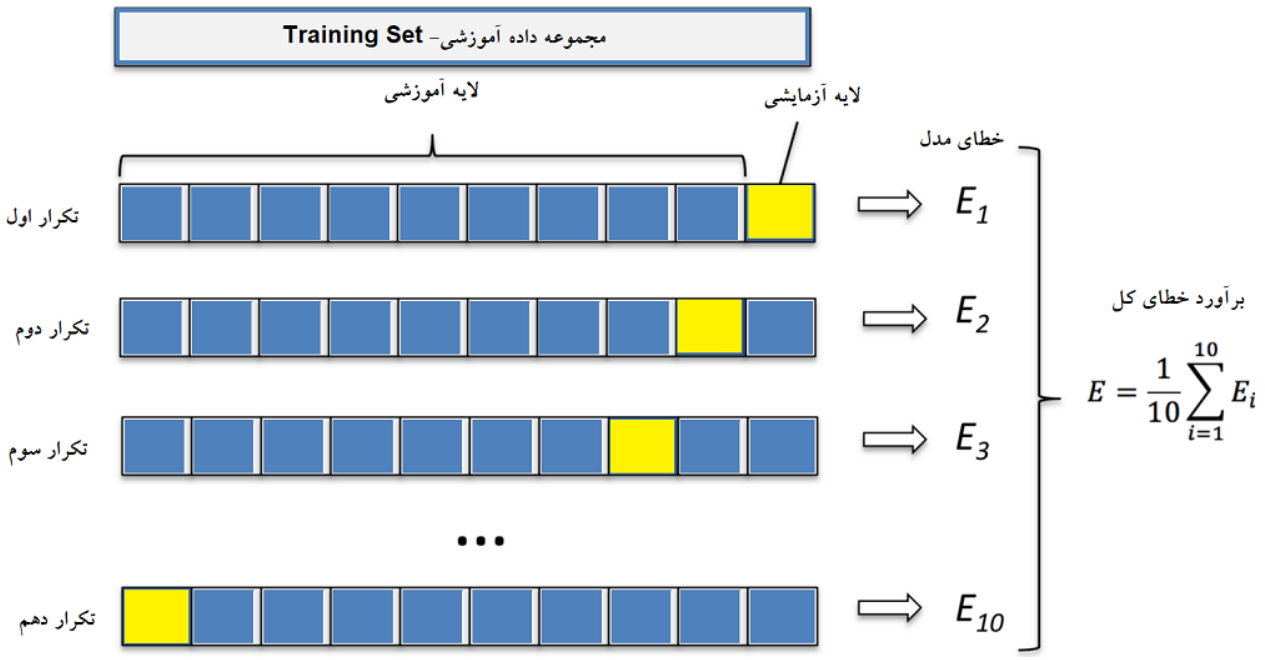
نکته دیگری که در شکل زیر مشخص است، مکمل بودن مجموعه داده‌های آموزشی و اعتبارسنجی است. با انتخاب بخشی از داده‌ها برای انجام فرایند CV، بقیه داده‌ها برای آموزش به کار گرفته می‌شوند. در هر مرحله از فرایند CV، مدل به دست آمده توسط داده‌های آزمایشی برای پیش‌بینی داده‌های CV به کار گرفته و «خطا» (Error) یا «دقت» (Accuracy) حاصل از برازش مدل روی داده‌های CV محاسبه می‌شود. معمولاً میانگین این خطاها (دقت‌ها) به عنوان خطای (دقت) کلی مدل در نظر گرفته می‌شود؛ البته بهتر است انحراف معیار خطاها (دقت‌ها) نیز گزارش شود. به این ترتیب با توجه به تعداد پارامترهای مختلف (پیچیدگی مدل)، می‌توان مدل‌های متفاوتی تولید و خطای برآورد آن‌ها را به کمک روش CV اندازه‌گیری کرد. در انتها مدلی را به عنوان مدل مناسب انتخاب خواهیم کرد که دارای کمترین برآورد خطا باشد.

روش اعتبارسنجی متقابل (Cross Validation)

روش اعتبار سنجی متقابل K-Fold Cross Validation

بر اساس شیوه و روش انتخاب مجموعه داده‌های اعتبارسنجی، گونه‌های مختلفی از روش‌های CV معرفی شده‌اند که در اینجا به آن نمی‌پردازیم. روش اعتبار سنجی متقابلی که در این تحقیق برای آموزش مدل‌های یادگیری

ماشین استفاده شده است، روش اعتبارسنجی متقابل K لایه‌ای[17] می‌باشد. اگر مجموعه داده‌های آموزشی را به‌طور تصادفی به k زیرنمونه یا لایه[18] با حجم یکسان تفکیک کنیم، می‌توان در هر مرحله از فرایند CV، تعداد k-1 از این لایه‌ها را به عنوان مجموعه داده آموزشی و یکی را به عنوان مجموعه داده اعتبارسنجی در نظر گرفت. شکل زیر، مراحل روش k-Fold را به خوبی نشان می‌دهد. مشخص است که با انتخاب k=10، تعداد تکرارهای فرآیند CV برابر با ۱۰ خواهد بود و دستیابی به مدل مناسب به سرعت امکان‌پذیر می‌شود. در این تحقیق، تعداد لایه‌ها یا فولدها برابر با 4 (K=4)، در نظر گرفته شده است.



روش اعتبار سنجی متقابل K-Fold Cross Validation

نرمال سازی داده‌ها

قبل از شروع مدل‌سازی ابتدا بایستی ورودی‌ها و در بعضی از موارد خروجی‌ها را نیز نرمال کرد زیرا وارد کردن داده ها به صورت خام باعث کاهش سرعت و دقت مدل KNN می شود.

17. K-Fold Cross Validation
18. Fold

برای نرمال کردن داده های ورودی از فرمول زیر اســـتفاده می کنیم، این فرمول داده ها را در بازه a و b نرمال می‌کند.

$$XN = a + \frac{X - X_{\min}}{X_{\max} - X_{\min}} \times (b - a)$$

در این رابطه $X_{\min}$، $X_{\max}$، XN به ترتیب مقدار مینیمم و ماکزیمم داده های ورودی و داده نرمالایز شده است. همچنین a و b نیز به ترتیب برابر با حد پایین و بالای بازه مورد نظر برای نرمالیزه کردن می‌باشد که در اینجا به ترتیب برابر با 0 و 1 می‌باشند.

معیارهای ارزیابی و اعتبارسنجی

در این تحقیق، به منظور ارزیابی کارآیی مدل‌ها، از معیارهای معتبر به شرح زیر استفاده شده است.

نرخ طبقه بندی صحیح (Correct Classification Rate)

نرخ طبقه بندی صحیح، برای مدل های طبقه بندی، به نسبت ردیف هایی که به درستی طبقه بندی شده‌اند به تعداد کل ردیف‌ها در دیتاست گفته می‌شود.

به عنوان مثال، یک نرخ طبقه بندی $0/82$ به این معنی اســـت که $82٪$ از ردیف‌های مجموعه داده های آموزش به درستی بر اساس مدل طبقه بندی شده اند.

ماتریس کانفیوژن (Confusion Matrix)

در بحث «دسـته‌بندی» (Classification) یک «مجموعه داده» (Data Set) با اسـتفاده از روش‌های دسـته‌بندی، هدف دسـتیابی به بالاترین دقت ممکن در دسـته‌بندی و تشـخیص دسـته‌ها اسـت. در برخی از مسـائل، تشـخیص صـــحیح نمونه‌های مربوط به یکی از دســـته‌ها برای ما اهمیت بیشـــتری دارد. به عنوان مثال، تحقیقی را در نظر

بگیرید که در آن، هدف شناسایی افراد مبتلا به یک نوع خاص از یک بیماری خطرناک است. فرض کنید برای افرادی که مبتلا به این بیماری هستند، خطر مرگ وجود دارد و جهت رفع این خطر، نیاز به دریافت نوعی داروی خاص دارند. در این شرایط، تشخیص درست بیماران دارای اهمیت بسیار زیادی است.

به این معنا که خطا در تشخیص افراد سالم قابل چشم پوشی است اما برای شناسایی افراد بیمار نمی‌توان این احتمال را به جان خرید. به عبارت دیگر، انتظار ما تشخیص تمام افراد بیمار است، بدون جا انداختن، حتی اگر فرد سالمی به اشتباه جز افراد بیمار دسته‌بندی شود. در چنین مواقعی، که دقت تشخیص یک دسته در مقایسه با دقت تشخیص کلی، اهمیت بیشتری دارد، مفهوم «ماتریس درهم‌ریختگی» (Confusion Matrix)، به کمک ما می‌آید.

بر اساس مثالی که پیش‌تر بیان شد، فرض کنید تعلق به دسته افراد بیمار را مثبت بودن (Positive) و عدم تعلق به این دسته را منفی بودن (Negative) در نظر بگیریم. هر نمونه یا فردی در واقعیت، متعلق به یکی از کلاسهای مثبت یا منفی است و از سوی دیگر، از هر الگوریتمی که برای دسته‌بندی داده‌ها استفاده شود، در نهایت هر نمونه عضو یکی از این دو «دسته» (Class) دسته‌بندی خواهد شد. بنابراین برای هر نمونه داده، یکی از چهار حالتی که در ادامه بیان شده، ممکن است اتفاق بیفتد.

- نمونه عضو دسته مثبت باشد و عضو همین کلاس تشخیص داده شود (مثبت صحیح یا True Positive)
- نمونه عضو کلاس مثبت باشد و عضو کلاس منفی تشخیص داده شود (منفی کاذب یا False Negative)
- نمونه عضو کلاس منفی باشد و عضو همین کلاس تشخیص داده شود (منفی صحیح یا True Negative)
- و در نهایت، نمونه عضو کلاس منفی باشد و عضو کلاس مثبت تشخیص داده شود (مثبت کاذب یا False Positive)

پس از اجرای الگوریتم دسته‌بندی، با توجه به توضیحات و تعاریف ذکر شده، می‌توان عملکرد یک طبقه‌بند را به کمک جدولی به شکل زیر بررسی کرد.

		برچسب پیش‌بینی شده	
		مثبت	منفی
برچسب شناخته شده	مثبت	TP	FN
	منفی	FP	TN

این جدول را اصطلاحا ماتریس درهم ریختگی می‌گویند. جدول یا ماتریس درهم ریختگی، نتایج حاصل از طبقه‌بندی را بر اساس اطلاعات واقعی موجود، نمایش می‌دهد. حال بر اساس این مقادیر می‌توان معیارهای مختلف ارزیابی دسته بند و اندازه‌گیری دقت را تعریف کرد. پارامتر دقت (Accuracy)، متداول‌ترین، اساسی‌ترین و ساده‌ترین معیار اندازه‌گیری کیفیت یک دسته‌بند است و عبارت است از میزان تشخیص صحیح دسته‌بند در مجموع دو دسته. این پارامتر در واقع نشان‌گر میزان الگوهایی است که درست تشخیص داده شده‌اند و بر اساس ماتریس ارائه شده در بالا، به شکل زیر فرموله و تعریف می‌شود:

$$Accuracy = (TP+TN) / (TP+FN+FP+TN)$$

البته، پارامتر دقت معمولا به صورت درصد بیان می‌شود. اما پارامترهای دیگری نیز علاوه بر معیار دقت وجود دارند که می‌توان به سادگی از این ماتریس استخراج کرد. یکی از متداول‌ترین آن‌ها، معیار حساسیت (Sensitivity) است که آن را «نرخ پاسخ‌های مثبت درست» (True Positive Rate) نیز می‌گویند. حساسیت به معنی نسبتی از موارد مثبت است که آزمایش آن‌ها را به درستی به عنوان نمونه مثبت تشخیص داده است. این پارامتر به صورت زیر محاسبه می‌شود:

$$Sensitivity\ (TPR) = TP / (TP+FN)$$

در واقع، «حساسیت» همان معیار بحث شده در مورد مثال بالا است. معیاری که مشخص می‌کند دسته‌بند، به چه اندازه در تشخیص تمام افراد مبتلا به بیماری موفق بوده‌است. همانگونه که از رابطه فوق مشخص است، تعداد افراد سالمی که توسط دسته‌بند به اشتباه به عنوان فرد بیمار تشخیص داده شده‌اند، هیچ تاثیری در محاسبه این پارامتر ندارد و در واقع زمانی که پژوهشگر از این پارامتر به عنوان پارامتر ارزیابی برای دسته‌بند خود استفاده

می‌کند، هدفش دستیابی به نهایت دقت در تشخیص نمونه‌های کلاس مثبت است.

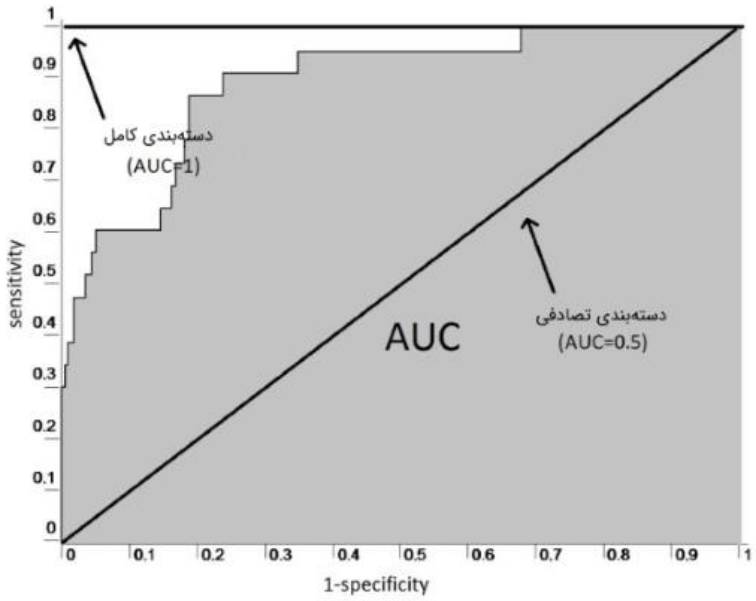
در نقطه مقابل این پارامتر، ممکن است در مواقعی دقت تشخیص کلاس منفی حائز اهمیت باشد. از متداول‌ترین پارامترها که معمولا در کنار حساسیت بررسی می‌شود، پارامتر خاصیت (Specificity)، است که به آن «نرخ پاسخ‌های منفی درست» (True Negative Rate) نیز می‌گویند. خاصیت به معنی نسبتی از موارد منفی است که آزمایش آن‌ها را به درستی به عنوان نمونه منفی تشخیص داده است. این پارامتر به صورت زیر محاسبه می‌شود:

$$\text{Specificity (TNR)} = TN / (TN+FP)$$

این دو پارامتر (حساسیت و خاصیت) نیز مشابه معیار دقت، معمولا به صورت درصد بیان می‌شوند. واضح است که پیش‌بینی عالی، پیش‌بینی است که مقادیر Sensitivity و Specificity مربوط به آن، هر دو صد درصد باشند؛ اما احتمال وقوع این اتفاق در واقعیت بسیار کم است و همیشه یک حداقل خطایی وجود دارد. پارامترهای حساسیت و خاصیت، بنابر ماهیتی که دارند همواره در رقابت با یکدیگر هستند. یعنی افزایش یکی با کاهش دیگری همراه است و برعکس. همین وضعیت منجر به تولید ابزاری دیگر برای ارزیابی کیفیت دسته‌بندها شده است.

منحنی ROC و سطح زیر آن AUC

«منحنی مشخصه عملکرد سیستم» (Receiver Operating Characteristic | ROC)، عبارت است از منحنی که ارتباط بین دو پارامتر حساسیت و خاصیت را بیان می‌کند. چنانکه در شکل زیر مشاهده می‌کنید، محور عمودی این نمودار نشان‌دهنده نرخ مثبت صحیح (Sensitivity)، و محور افقی نشان‌دهنده مقدار نرخ مثبت غلط (One-Specificity) است. نتایج مختلف دسته‌بندی نشانگر نقاط مختلف بر روی این نمودار هستند و در نهایت یک منحنی را تشکیل می‌دهند. با توجه به شکل زیر، در بهترین حالت و با فرض طبقه‌بندی صد درصد صحیح در هر دو دسته، نقطه مربوطه عبارت است از نقطه گوشه بالای سمت چپ، یعنی نقطه $(0,1)$ و نیز با فرض دسته‌بندی به صورت تصادفی، نقطه متناظر در منحنی، یکی از نقاط موجود روی خط واصل نقطه $(0,0)$ و نقطه $(1,1)$ خواهد بود. در واقعیت، منحنی حاصل از یک دسته‌بندی، منحنی بین این دو حالت است.



مساحت زیر این نمودار (Area Under Curve)، به عنوان یک معیار برای ارزیابی عملکرد دسته‌بند مورد استفاده قرار می‌گیرد. با توجه به توضیحاتی که پیش‌تر ارائه شد، بدیهی است که در حالت ایده‌آل، مساحت زیر منحنی برابر با بیشترین مقدار خود، یعنی یک است. بنابراین، هر چه مساحت زیر نمودار به عدد یک نزدیکتر باشد، به معنای بهتر بودن عملکرد دسته‌بند است. علاوه بر دو پارامتر حساسیت و خاصیت، پارامترهای دیگری هم از ماتریس درهم‌ریختگی استخراج می‌شوند که هر یک بیان کننده مفهومی هستند و کاربردهای متفاوتی دارند.

پارامتر مهم دیگری به نام «معیار اف» (F-Measure) وجود دارد که برای ارزیابی عملکرد دسته‌بندها بسیار مورد استفاده قرار می گیرد و از ترکیب دو پارامتر حساسیت و ارزش اخباری مثبت حاصل می شود. با این توضیح که پارامتر ارزش اخباری مثبت را اصطلاحا دقت (Precision)، و حساسیت را اصطلاحا صحت (Recall) می نامند، «معیار اف» به دو صورت زیر تعریف می‌شود:

$$\text{F-measure= 2 * (Recall * Precision) / (Recall + Precision)}$$

$$\mathrm{F_1} = 2 \cdot \frac{\mathrm{PPV} \cdot \mathrm{TPR}}{\mathrm{PPV} + \mathrm{TPR}} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

ماتریس درهم ریختگی، با وجود منطق و ساختار ساده ای که دارد، مفهومی قدرتمند است که در انواع تحقیقات، می‌تواند به تنهایی اطلاعاتی جامع از نحوه عملکرد دسته‌بند ارائه کند.

میانگین مربعات خطا (Mean Squared Error)

روشی برای برآورد میزان خطاست که در واقع تفاوت بین مقادیر تخمینی و آنچه تخمین زده شده، است. MSE به دو دلیل تقریباً همه جا مثبت است (صفر نیست) یک اینکه تصادفی است و دوم به این دلیل که تخمین‌گر اطلاعاتی که قابلیت تولید تخمین دقیق تری دارد را حساب نمی‌کند. پس این شاخص که مقداری همواره نامنفی دارد، هرچقدر مقدار آن به صفر نزدیکتر باشد، نشان دهنده میزان کمتر خطاست. مقدار این شاخص به صورت زیر بیان می شود:

$$MSE = \frac{1}{n} \times \sum_{i=1}^{n} [(x_{imeas} - x_{ipred})^2]$$

x_{imeas}, x_{ipred}, n به ترتیب برابر با تعداد متغیر اندازه‌گیری شده، مقدار متغیر پیش‌بینی شده و مقدار متغیر اندازه‌گیری شده می‌باشد.

مجذور میانگین مربعات خطا (Root Mean Square Error)

ریشهٔ میانگین مربعات خطا (RMSE) نیز یک تابع تناسب یا تابع هدف است و در واقع مجذور شاخص میانگین مربعات خطاست. این شاخص به عنوان معیاری از خطای مطلق بین متغیر شبیه سازی و مشاهده‌ای است. مقدار این شاخص آماری بین صفر تا بی نهایت متغیر است. هر چه مقدار این شاخص کمتر باشد شبیه سازی بهتری صورت گرفته است و مقدار بهینهٔ آن صفر است. مقدار این شاخص به صورت زیر بیان می شود:

$$RMSE = \sqrt{\frac{1}{n} \times \sum_{i=1}^{n} [(x_{imeas} - x_{ipred})^2]}$$

x_{imeas}, x_{ipred}, n به ترتیب برابر با تعداد متغیر اندازه‌گیری شده، مقدار متغیر پیش‌بینی شده و مقدار متغیر اندازه‌گیری شده می‌باشد.

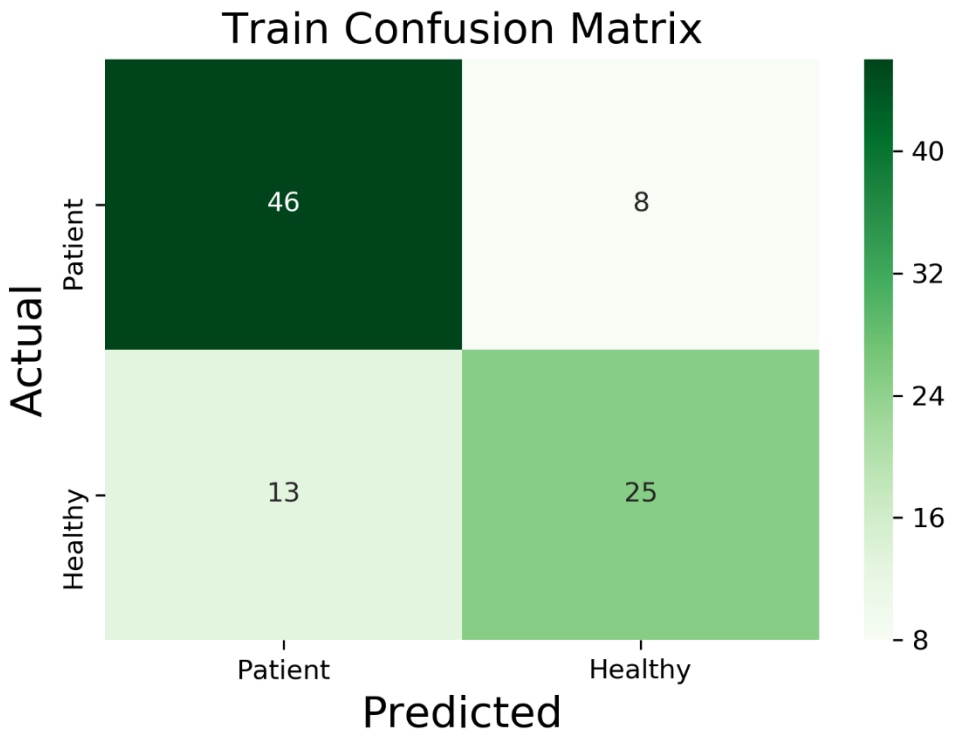
نتایج مدل‌سازی و طبقه‌بندی توسط مدل KNN

نتایج مدل KNN برای طبقه بندی و تشخیص بیماری سرطان سینه

AUC	F-score	Specificity	Sensitivity	RMSE	ACC (%)	مرحله
0.75	**0.77**	**0.66**	**0.85**	**0.48**	**0.77**	**TR**
0.84	**0.83**	**0.79**	**0.9**	**0.41**	**0.83**	**TS**

در ادامه ماتریس‌های کانفیوژن برای دو حالت آموزش، آزمایش ارائه می‌شود

- حالت آموزش (Train)



- حالت آزمایش (Test)