گزارش پروژه داده کاوی – پیاده سازی SVM

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.

هدف اصلی در مسأله طبقه‌بندی دودویی، تخمین تابع (**Error!** *No text of specified style in document.*-1) با استفاده از داده‌های یادگیری نشان داده شده در (**Error!** *No text of specified style in document.*-2)، به طوری که تابع (**Error!** *No text of specified style in document.*-1) بتواند برچســـب داده‌های جدید را به درستی پیش‌بینی کند.

$$f : IR^n \to \{\pm 1\} \ or \ IR \qquad (\mathbf{1}\text{-}\textit{Error! No text of specified style in document.})$$

$$\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), ..., (\boldsymbol{x}_m, y_m)\} \in IR^n \times (\{\pm 1\} \ or \ IR) \qquad (\mathbf{2}\text{-}\textit{Error! No text of specified style in document.})$$

اگر فرض کنیم که $\boldsymbol{x}$ و y از یک تابع توزیع احتمال توأم $P(\boldsymbol{x},y)$ تولید شده‌اند، می‌توان ریسک مورد انتظار تابع (*Error! No text of specified style in document.*-1) را برای نمونه‌هایی که در فرایند یادگیری استفاده نشده‌اند را با استفاده از رابطه (**Error!** *No text of specified style in document.*-3) محاسبه کرد.

$$R(f) = \int \frac{1}{2} |f(\boldsymbol{x}) - y| dP(\boldsymbol{x}, y) \qquad (\mathbf{3}\text{-}\textit{Error! No text of specified style in document.})$$

به جز در مواردی محدود، معمولاً $P(\boldsymbol{x},y)$ را نداریم و مجبوریم از خطای تجربی (**Error!** *No text of specified style in document.*-4) به عنوان برآوردی از ریسک مورد انتظار استفاده کنیم.

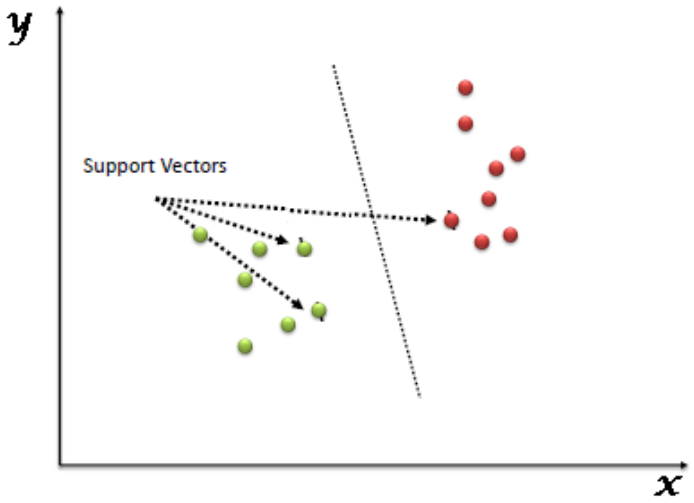
$$\widehat{R}(f) = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{2} |f(\boldsymbol{x}_i) - y_i| \qquad (\mathbf{4}\text{-}\textit{Error! No text of specified style in document.})$$

بنابراین طبقه‌بندی، که ریسـک سـاختاری را کمینه می‌کند، تابعی خطی اسـت که ریسـک تجربی را کمینه کند. SVM از این رهیافت اسـتفاده می‌کند و همواره در خانواده توابع خطی در جسـت‌وجوی تابعی با کمترین ریسـک تجربی است.

1-2 ماشین بردار پشتیبان (SVM)

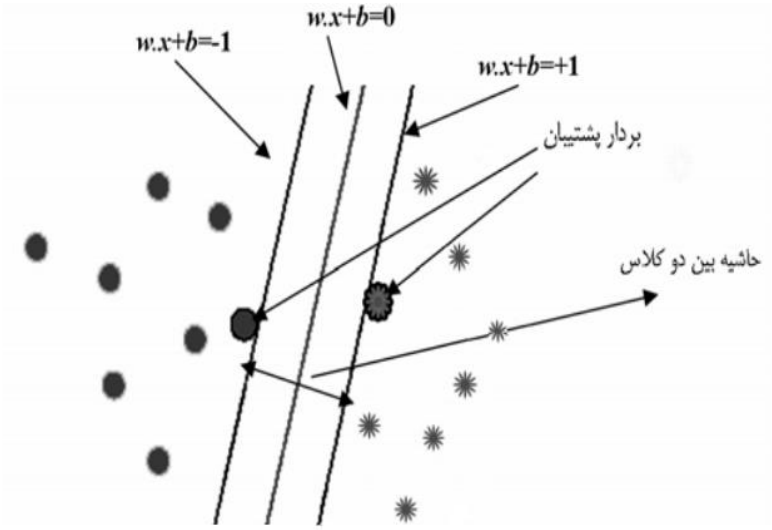
ماشـین بردار پشـتیبان یا SVM یک نوع الگوریتم نظارت شـده یادگیری ماشـین اسـت که در سـال 1979 توسـط

وپنیک Vapnik ارائه گردید که دارای کاربردهایی برای دسته‌بندی داده‌های ورودی (Classification) و نیز برای تخمین و برآورد تابع برازش داده‌ها (Regression) به کار می‌رود، به طوری که در دسته بندی و برازش داده‌ها، کمترین خطا رخ دهد. در حالت کلی داده‌ها به سه دسته داده‌های آموزشی، داده‌های صحت‌سنجی و داده‌های آزمون تقسیم بندی می‌شوند به طوری که داده‌های آموزشی باعث آموزش ماشین بردار پشتیبان می‌شوند، داده‌های صحت‌سنجی به واسنجی پارامترهای ماشین می‌پردازد و در نهایت از این ماشین برای طبقه‌بندی یا برآورد داده‌های آزمون استفاده می‌شود که این داده‌ها در مراحل قبل به الگوریتم داده نشده‌اند و الگوریتم باید برچسب یا مقدار خروجی متناظر با این داده‌ها را تولید کند. این روش بر مبنای تئوری بهینه‌سازی مقید است که از اصل کمینه‌سازی خطای ساختاری استفاده کرده و منجر به یک جواب بهینه کلی می‌گردد. با این حال از الگوریتم SVM بیشتر در مسائل طبقه‌بندی استفاده می‌شود. در الگوریتم SVM، هر نمونه داده را به عنوان یک نقطه در فضای n-بعدی روی نمودار پراکندگی داده‌ها ترسیم کرده (n تعداد ویژگی‌هایی است که یک نمونه داده دارد) و مقدار هر ویژگی مربوط به داده‌ها، یکی از مؤلفه‌های مختصات نقطه روی نمودار را مشخص می‌کند. سپس، با ترسیم یک خط راست، داده‌های مختلف و متمایز از یکدیگر را دسته‌بندی می‌کند (مطابق شکل **1-Error! No text of specified style in document.**). به بیان ساده، بردارهای پشتیبان در واقع مختصات یک مشاهده منفرد هستند. ماشین بردار پشتیبان ایجاد کننده‌ی مرزی است که به بهترین شکل دسته‌های داده‌ها را از یکدیگر جدا می‌کند.



شکل 1-Error! No text of specified style in document.**: ترسیم داده‌ها در فضای n بعدی در الگوریتم ماشین بردار پشتیبان**

ماشین بردار پشتیبان مبتنی بر مینیمم سازی ریسک ساختاری می‌باشد که از تئوری آموزش آماری گرفته شده است. ماشین بردار پشتیبان در واقع یک طبقه بندی کننده دودویی است. در مورد دو کلاس، روش SVM سعی دارد یک ابر صفحه ایجاد نماید که فاصله هر کلاس را تا ابرصفحه حداکثر نماید. داده‌های نقطه‌ای که به ابرصفحه نزدیک تر هستند، برای اندازه‌گیری این فاصله به کار می‌روند. از این رو، این داده‌های نقطه‌ای را بردارهای پشتیبان می‌نامند. در شکل **Error! No text of specified style in document.**-2 دو کلاس و بردارهای پشتیبان مربوط به آن‌ها نشان داده شده است.



شکل .Error! No text of specified style in document**-2: مرز خطی بهینه برای حالتی که دو کلاس کامل از یکدیگر جدا هستند**

فرض کنید داده‌ها از دو کلاس تشکیل شده و کلاس‌ها در مجموعه دارای X_i , $i=1,2,3,\ldots,L$ نقطه آموزشی باشند که X_i یک بردار است. این دو کلاس با $Y_i=\pm 1$ برچسب زده می‌شوند. برای محاسبه مرز تصمیم‌گیری دو کلاس کاملا جدا از هم، از حاشیه بهینه استفاده می‌شود. در این روش مرز خطی بین دو کلاس به گونه‌ای محاسبه می‌شود که:

1- تمام نمونه‌های کلاس 1+ در یک طرف مرز و تمام نمونه‌های کلاس 1- در طرف دیگر مرز واقع شوند.

2- مرز تصمیم گیری به گونه‌ای باشد که فاصله نزدیک ترین نمونه‌های آموزشی هر دو کلاس از یکدیگر در

راستای عمود بر مرز تصمیم‌گیری تا جایی که ممکن است حداکثر شود.

یک مرز تصمیم گیری خطی را در حالت کلی می‌توان به صورت زیر نوشت:

$$w.x + b = 0 \qquad \textbf{(5-}\textit{Error! No text of specified style in document.}\textbf{)}$$

x یک نقطه روی مرز تصمیم گیری و w یک بردار n بعدی عمود بر مرز تصمیم گیری است. b فاصله مبدا تا مرز تصمیم گیری و $w.x$ بیانگر ضرب داخلی دو بردار w و x است. از آن جا که با ضرب یک ثابت در دو طرف رابطه (5-***Error!*** *No text of specified style in document.*) باز هم تساوی برقرار است، برای تعریف یکتای مقدار b و w شرایط زیر روی آن‌ها اعمال می‌شود:

$$\Rightarrow Y_i(W.X_i + b) = 1 \quad \text{اگر } X_i \text{ یک بردار پشتیبان باشد}$$

$$\Rightarrow Y_i(W.X_i + b) > 1 \quad \text{اگر } X_i \text{ یک بردار پشتیبان نباشد}$$

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

اولین مرحله برای محاسبه مرز تصمیم گیری بهینه، پیدا کردن نزدیک ترین نمونه‌های آموزشی دو کلاس است. در مرحله بعد فاصله آن نقاط از هم در راستای عمود بر مرزهایی که دو کلاس را به طور کامل جدا می‌کنند محاسبه می‌شود. مرز تصمیم گیری بهینه، مرزی است که حداکثر حاشیه را داشته باشد. مرز تصمیم گیری بهینه با حل مسئله بهینه سازی زیر محاسبه می‌شود:

$$\max\left[y_i \frac{(W.X_i + b)}{|W|} \right] \quad \text{i=1,2,3,...,L}$$

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

با توجه به رابطه (**7**-*Error! No text of specified style in document.*) و انجام یک سری عملیات ریاضی، رابطه بالا به رابطه زیر تبدیل می‌شود:

$$\min \frac{1}{2}|W|^2 \ , \ y_i(W.X_i + b) \geq 0 \quad \text{i=1,2,3,...,L}$$

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

حل کردن مسئله بهینه سازی (**8**-*Error! No text of specified style in document.*) کار مشکلی است. برای ساده تر کردن آن با استفاده از روش ضرایب نامعین لاگرانژ این مسئله بهینه سازی را می‌توان به فرم زیر تبدیل

کرد که λ_i ها ضرایب لاگرانژ می‌باشند.

$$\max\left[-\frac{1}{2}\sum_{i=1}^{L}\sum_{j=1}^{L}\lambda_i y_i(X_i,X_j)y_i\lambda_i+\sum_{i=1}^{L}\lambda_i\right] \qquad \sum_{i=1}^{L}\lambda_i y_i=0$$
$$\lambda_i \geq 0 \quad \text{i=1,2,3,...,L}$$

(*Error! No text of specified style in document.*-**9**)

پس از حل مسئله بهینه سازی (*Error! No text of specified style in document.*-**9**) و یافتن ضرایب لاگرانژ، W با استفاده از رابطه زیر محاسبه می‌شود:

$$W=\sum_{i=1}^{L}\lambda_i y_i X_i$$

(*Error! No text of specified style in document.*-**10**)

λ_i مربوط به بردارهای پشتیبان بزرگتر از صفر و λ_i مربوط به سایر نقاط صفر خواهد بود. بنابراین با توجه به رابطه (*Error! No text of specified style in document.*-**10**) و صفر بودن λ_i مربوط به X_i هایی که بردار پشتیبان نیستند، برای بدست آوردن مرز تصمیم گیری فقط نیاز به تعدادی محدود از نقاط آموزشی که همان بردارهای پشتیبان هستند می‌باشد و همه آن‌ها لازم نیستند. در نتیجه پس از یافتن W با استفاده از رابطه (*Error! No text of specified style in document.*-**11**) مقدار b به ازای بردارهای پشتیبان مختلف محاسبه شده و b نهایی با میانگین گیری از b های حاصل بدست می‌آید.

$$[Y_i(W.X_i+\mathrm{b})-1]=0$$

(*Error! No text of specified style in document.*-**11**)

در نتیجه طبقه بندی کننده‌های نهایی از طریق رابطه زیر بدست می‌آیند:

$$f(X,W,b)=\mathrm{sgn}(W.X+\mathrm{b})$$

(*Error! No text of specified style in document.*-**12**)

الگوریتم بالا مرز خطی دو کلاس کاملا جدا از هم را نشان می‌دهد، اما در حالتی که کلاس‌ها با هم همپوشانی داشته باشند جدا کردن کلاس‌ها به وسیله مرز تصمیم گیری خطی همواره با خطای زیادی همراه خواهد بود. برای حل این مشکل می‌توان ابتدا داده‌ها را از فضای اولیه R^n با استفاده از یک تبدیل غیرخطی ϕ، به فضای R^m با ابعاد بیشتر منتقل کرد که در فضای جدید کلاس‌ها تداخل کمتری با یکدیگر داشته باشند. سپس در فضای جدید با استفاده از معادلات قبلی و جایگزینی X_i با $\phi(X_i)$ و در نظر گرفتن مقداری خطا مرز تصمیم گیری بهینه محاسبه می‌شود. با توجه به این امر و رابطه (*Error! No text of specified style in document.*-**9**)

در این حالت یافتن مرز تصمیم‌گیری بهینه از حل مسئله بهینه سازی (*Error! No text of specified style in document.*-**13**) بدست می‌آید.

$$\max\left[-\frac{1}{2}\sum_{i=1}^{L}\sum_{j=1}^{L}\lambda_i y_i(\phi(X_i),\phi(X_j))y_i\lambda_i+\sum_{i=1}^{L}\lambda_i\right] \qquad \sum_{i=1}^{L}\lambda_i y_i=0$$

(*Error! No text of specified style in document.*-**13**)

$$C\geq\lambda_i\geq 0 \quad \text{i=1,2,3,...,L}$$

در این مسئله مقدار C یک عدد ثابت است. اگر $C\to\infty$، مسئله بهینه سازی به سمت یافتن یک مرز برای کلاس‌های با تداخل بسیار زیادتر پیش می‌رود. از طرفی اگر $C\to 0$، مسئله بهینه سازی به سمت یافتن مرز بهینه جدا کننده کلاس‌های با تداخل بسیار کم پیش خواهد رفت. در رابطه (*Error! No text of specified style in document.*-**13**) به جای استفاده از ϕ، از یک تابع کرنل که به صورت زیر تعریف می‌شود، استفاده می‌گردد.

$$K(X_i,X_j)=\phi(X_i)\phi(X_j)$$

(*Error! No text of specified style in document.*-**14**)

پس از تعیین یک $K(X_i,X_j)$ در رابطه (*Error! No text of specified style in document.*-**13**) به جای $\phi(X_i)\phi(X_j)$، تابع $K(X_i,X_j)$ قرار داده شده و مسئله بهینه‌سازی حل می‌شود. $K(X_i,X_j)$ در واقع یک تابع در فضای اولیه می‌باشد که برابر با ضرب داخلی دو بردار در فضای ویژگی است. برای معادل بودن تابع $K(X_i,X_j)$ با ضرب داخلی دو بردار در فضای ویژگی، باید $K(X_i,X_j)$ یک تابع معین و مثبت متقارن بوده و در شرایط مرسر (Mercer Condition) صدق کند. برخی از مهم‌ترین توابع کرنل یا هسته که در این شرط صدق می‌کنند، عبارتند از:

جدول Error! No text of specified style in document.-**1**: **توابع کرنل رایج برای ماشین بردار پشتیبان**

کرنل خطی (Linear)	$K(X_i,X_j)=(X_i,X_j)$
کرنل چند جمله‌ای (Polynomial)	$K(X_i,X_j)=((X_i,X_j)+1)^d$
کرنل گوسین (Gaussian)	$K(X_i,X_j)=\exp(-\frac{\|X_i-X_j\|^2}{2\sigma^2})$

کرنل سیگموئید (Sigmoid)	$K(X_i, X_j) = \tanh(-\alpha(X_i, X_j) + c)$

میزان کارایی ماشـــین بردار پشـــتیبان به ازای هر نوع تابع کرنل متفاوت می‌باشـــد. اینکه کدام تابع کرنل بهرین نتیجه را برای یک سری داده ارائه می‌دهد، به درستی معلوم نیست وباید از طریق آزمون و خطا مشخص شود.

ماشین بردار پشتیبان یک طبقه بندی کننده دودویی است. بنابراین در حالتی که بیش از دو کلاس وجود داشته باشـد نمی‌توان مسـتقیما از آن اسـتفاده کرد. در حالت کلی برای اسـتفاده از طبقه بندی کننده‌های دودویی در حالت چند کلاسـه باید ابتدا چند طبقه بندی کننده دودویی طراحی شـود. طبقه بندی نهایی با اسـتفاده از ادغام اطلاعات طبقه بندی کننده‌های دودویی انجام می‌گیرد.

آماده‌سازی داده‌ها

جهت مدل سازی با ماشین بردار پشتیبان SVM، 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 ماشین بردار پشتیبان و 20 درصد باقیمانده برای آزمایش ماشین بردار پشتیبان 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[1] 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[2]، شاخص توده بدنی[3]، گلوکز[4]، انسولین[5]، مدل ارزیابی هموستاتیک[6]، لپتین[7]، ادیپونکتین[8]، رزیستین[9]، پروتئین کموتاکسی مونوسیت یک[10].

متغیرهای ورودی به مدل ماشین بردار پشتیبان SVM در جدول زیر ارائه شده است.

متغیرهای ورودی و خروجی مدل SVM

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

[1] https://archive.ics.uci.edu/ml/datasets.php
[2] Age
[3] BMI
[4] Glucose
[5] Insulin
[6] homeostatic model assessment (HOMA)
[7] Leptin
[8] Adiponectin
[9] Resistin
[10] 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[11] 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[12] مدل به کار می‌رود. به این ترتیب این فرآیند یک روش بازنمونه‌گیری به منظور برآورد خطای مدل محسوب می‌شود.

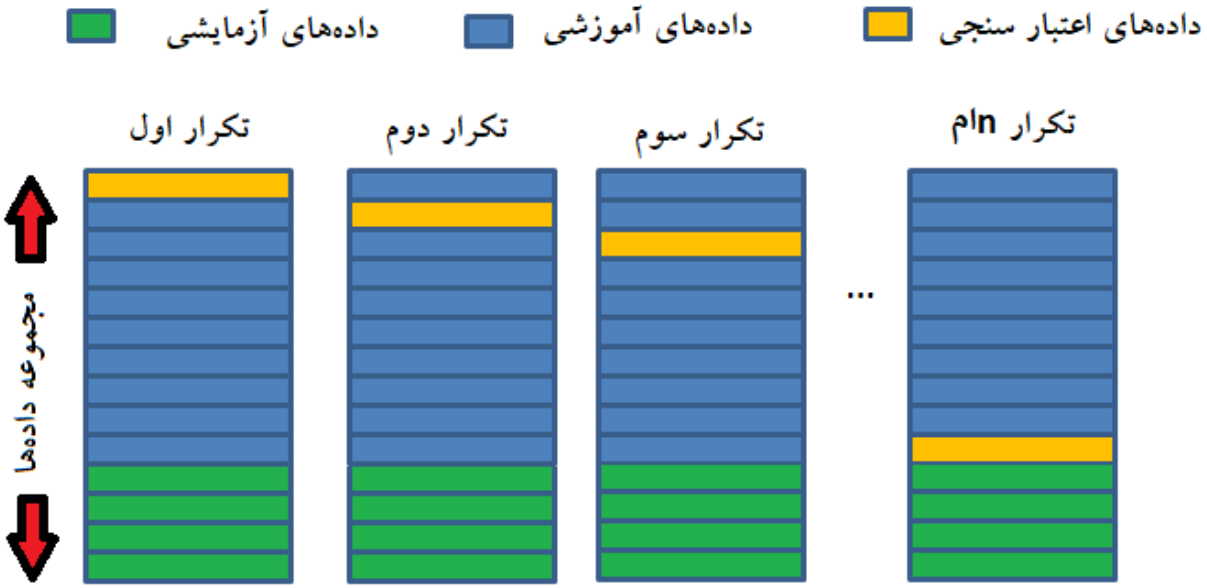
باید توجه داشت که داده‌های آزمایشی در فرایند CV ممکن است در تکرار بعدی به عنوان داده‌های آموزشی به کار روند، در نتیجه، ماهیت آن‌ها با داده‌هایی که در قسمت قبل به عنوان داده‌های آزمایشی (Test set) معرفی شد، متفاوت است. شکل زیر به درک ماهیت داده‌های تست در فرآیند CV کمک می‌کند. مشخص است که داده‌های اعتبارسنجی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

---

11. Cross Validation (CV)
12. Validation

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.



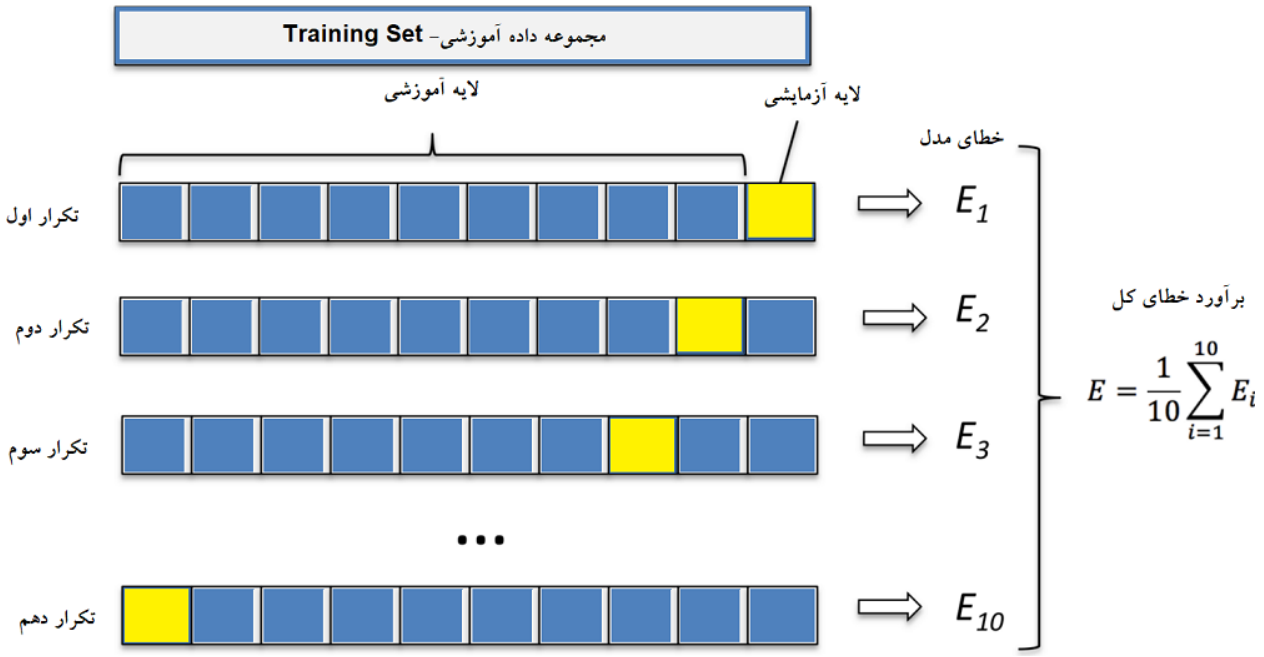
روش اعتبارسنجی متقابل (Cross Validation)

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[۱۳] می‌باشد. اگر مجموعه داده‌های آموزشی را به‌طور

13. K-Fold Cross Validation

تصادفی به k زیرنمونه یا لایه[14] 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 ماشین بردار پشتیبان می شود.

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

14. Fold

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

در این رابطه $X_{\min}$، $X_{\max}$، XN به ترتیب مقدار مینیمم و ماکزیمم داده های ورودی و داده نرمالایز شـــده اســت. همچنین a و b نیز به ترتیب برابر با حد پایین و بالای بازه مورد نظر برای نرمالیزه کردن می‌باشــد 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 اســت 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 صــحیح نمونه‌های مربوط به یکی از دســته‌ها برای ما اهمیت بیشــتری دارد. به عنوان مثال، تحقیقی را در نظر بگیرید که در آن، هدف شـناسـایی افراد مبتلا به یک نوع خاص از یک بیماری خطرناک اسـت. 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 اصـــطلاحا ماتریس درهم ریختگی می‌گویند. جدول یا ماتریس درهم ریختگی، نتایج حاصـــل از طبقه‌بندی را بر اســـاس اطلاعات واقعی موجود، نمایش می‌دهد. حال بر اســـاس 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 صـورت درصـد بیان می‌شـود. اما پارامترهای دیگری نیز علاوه بر معیار دقت وجود دارند که می‌توان به ســـادگی از این ماتریس اســـتخراج کرد. یکی از متداول‌ترین آن‌ها، معیار حســـاســـیت (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 پژوشـهگر از این پارامتر به عنوان پارامتر ارزیابی برای دسـته‌بند خود اسـتفاده می‌کند، هدفش دستیابی به نهایت دقت در تشخیص نمونه‌های کلاس مثبت است.

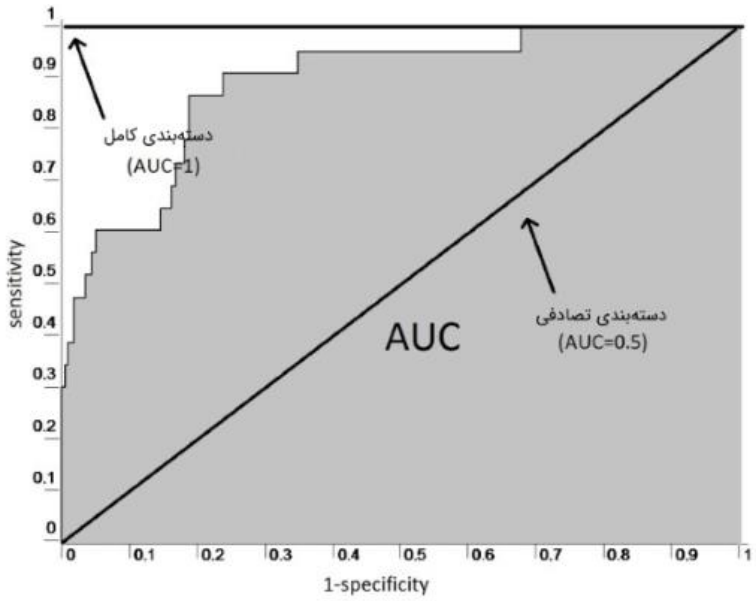
در نقطه مقابل این پارامتر، ممکن است در مواقعی دقت تشخیص کلاس منفی حائز اهمیت باشد. از متداول‌ترین پارامترها که معمولا در کنار حساسیت بررسی می‌شود، پارامتر خاصیت (Specificity)، است که به آن «نرخ پاسخ‌های منفی درست» (True Negative Rate) نیز می‌گویند. خاصیت به معنی نسبتی از موارد منفی است که آزمایش آن‌ها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

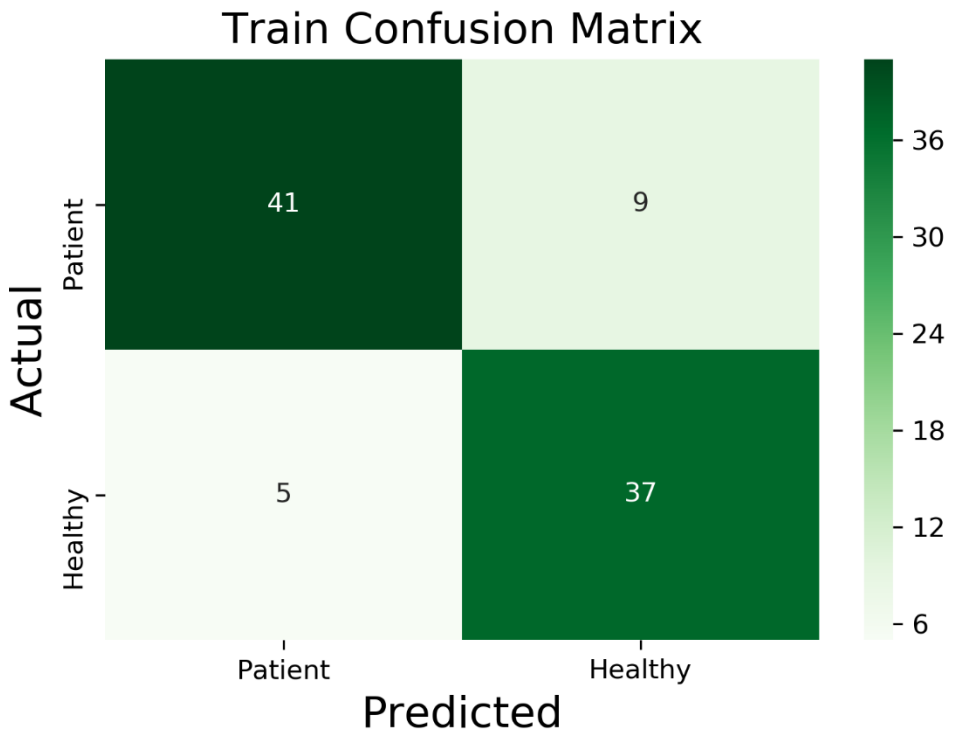
نتایج مدل‌سازی و طبقه‌بندی توسط ماشین بردار پشتیبان SVM

نتایج مدل ماشین بردار پشتیبان SVM برای طبقه بندی و تشخیص بیماری سرطان سینه

AUC	F-score	Specificity	Sensitivity	RMSE	ACC (%)	مرحله
0.85	**0.85**	**0.88**	**0.82**	**0.39**	**0.85**	**TR**
0.84	**0.83**	**0.9**	**0.79**	**0.41**	**0.83**	**TS**

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

- حالت آموزش (Train)



- حالت آزمایش (Test)