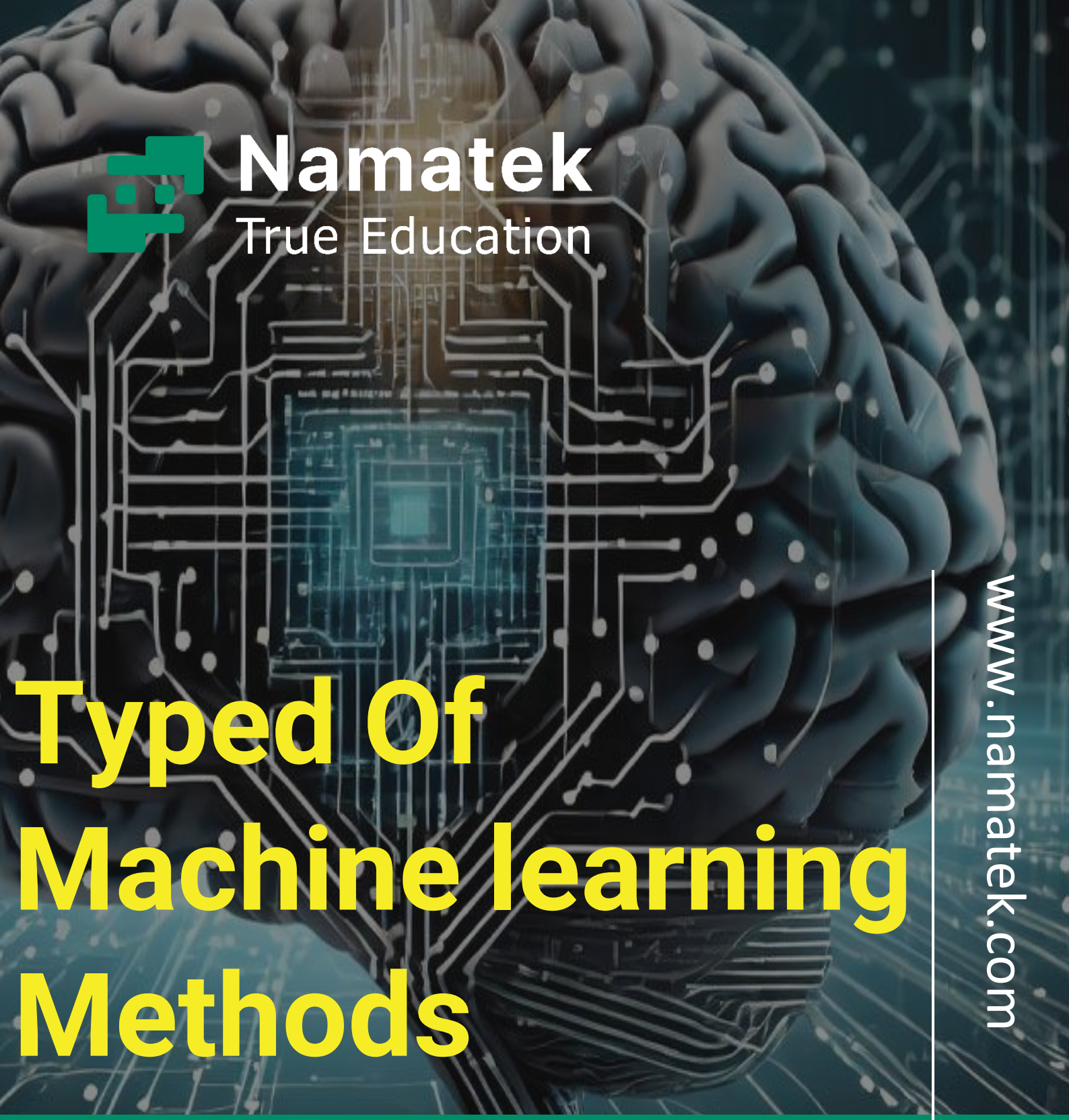


Namatek
True Education

Typed Of Machine learning Methods

www.namatek.com

الگوریتم های یادگیری ماشین

فهرست مطالب

www.namatek.com

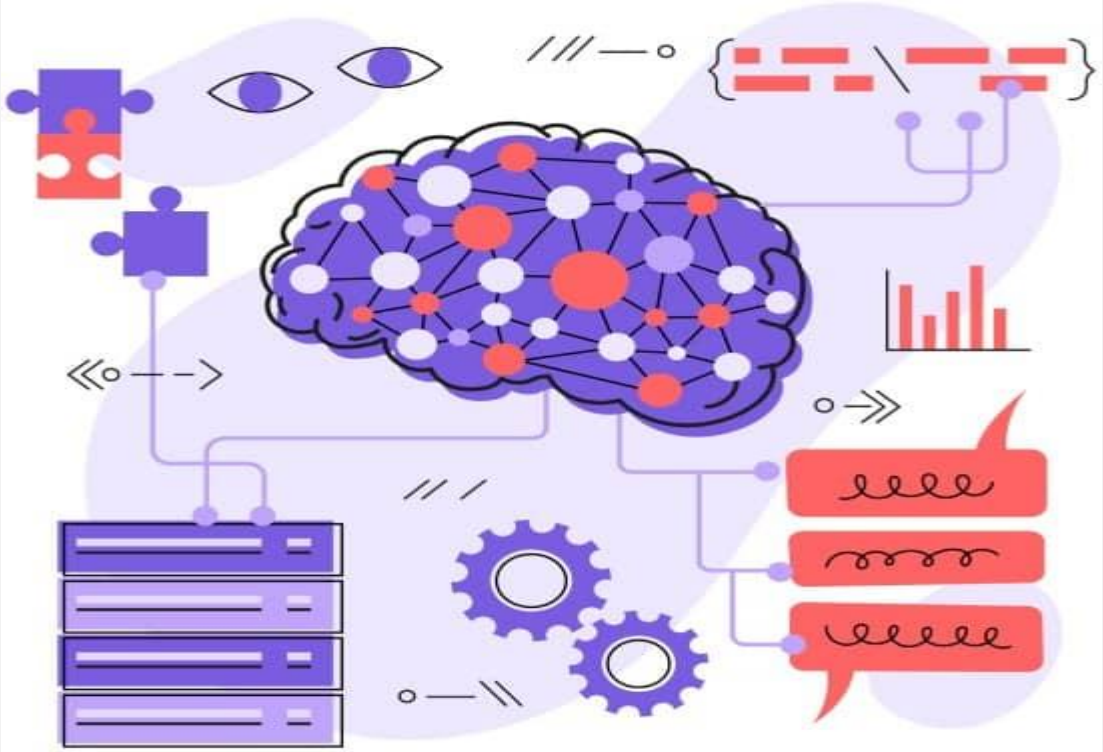
این روزها همه ما بارها اصطلاحاتی مانند هوش مصنوعی، یادگیری ماشین، الگوریتم های یادگیری ماشین و... به گوشمان خورده است و به خوبی می دانیم که دنیا در حال گسترش هوشمند شدن بسیاری از فرآیندها است.

به همین دلیل بازار کار یادگیری ماشین بسیار داغ و پر مخاطب است و برای پا گذاشتن در این عرصه، باید با مفاهیم اولیه آن به خوبی آشنا بود.

در این مقاله قصد داریم به بررسی مهم ترین و محبوب ترین الگوریتم های یادگیری ماشین که امروزه مورداستفاده هستند، بپردازیم.

با ما همراه باشید تا به سادگی به همراه ذکر مثال، این الگوریتم ها را بشناسید.

الگوریتم یادگیری ماشین چیست؟



برای تعریف الگوریتم (Algorithm) در یک جمله ساده می توان گفت، یک مجموعه ای از قوانین و دستورالعمل ها است که برای حل یک مسئله یا محاسبه یک مقدار توسط کامپیوتر استفاده می شود.

الگوریتم های یادگیری ماشین (Machine Learning Algorithms) همانطور که از نامشان پیداست توسط ماشین هایی با قابلیت یادگیری یک موضوع استفاده می شوند و برنامه هایی هستند که بدون مشارکت و مداخله انسان ها می توانند از روی داده ها موضوعی را یادگرفته و با استفاده از تجربیات آن را بهبود ببخشند.

مواردی که در این حوزه برای یادگیری ماشین درنظر گرفته می شوند، برای مثال شامل توابع تبدیل یک مجموعه داده ورودی به خروجی های متناظر، تشخیص ساختار پنهان داخل یک مجموعه داده بدون برچسب و مشخصات یا یادگیری مبتنی بر نمونه ها و... است.

در دنیای امروز سیستم های هوشمند در حال جایگزینی سیستم های دستی هستند و تنوع بسیار زیادی در روش های پیاده سازی این سیستم ها و کاربردهای آن ها وجود دارد.

برای مثال امروزه با استفاده از الگوریتم های یادگیری ماشین می توان به یک کامپیوتر آموخت که چطور شطرنج بازی کند.

تا به امروز انواع بسیار گسترده ای از روش های یادگیری ماشین طراحی و ساخته شده اند تا در حل مشکلات پیچیده دنیای واقعی به ما کمک کنند.

در ادامه به معرفی دسته بندی اصلی و مهم ترین الگوریتم های موجود تا به امروز می پردازیم.

دسته بندی الگوریتم های یادگیری ماشین

روش های یادگیری ماشین به صورت کلی به سه دسته زیر تقسیم می شوند.

الگوریتم های یادگیری ماشین با نظارت (Supervised)

الگوریتم های با نظارت از داده های برچسب گذاری شده برای یادگیری توابع تبدیل ورودی ها به خروجی ها استفاده می کنند.

این روش دارای دو زیر شاخه اصلی و مهم با عناوین طبقه بندی (Classification) و رگرسیون (Regression) است. از خانواده الگوریتم های طبقه بندی زمانی استفاده می شود که خروجی تابع موردنظر از جنس دسته بندی ها (Categories) باشد.

برای مثال با استفاده از بررسی ورودی ها تشخیص می دهد یک فرد بیمار است یا سالم.

خانواده الگوریتم های رگرسیون برای توابعی استفاده می شوند که خروجی ها از جنس اعداد و ارزش های واقعی هستند.

برای مثال یک مدل رگرسیون ممکن است برای پیش بینی میزان بارش باران از یک مجموعه داده ورودی استفاده شود.

الگوریتم های یادگیری ماشین بدون نظارت (Unsupervised Learning)

الگوریتم های بدون نظارت زمانی استفاده می شوند که در مجموعه داده های موجود فقط ورودی های یک تابع را داشته باشیم و خروجی ای وجود نداشته باشد. در این حالت از روش یادگیری مبتنی بر داده های بدون برچسب گذاری استفاده می کنند تا ساختار داخلی داده ها را کشف کنند.

این روش دارای سه زیر شاخه اصلی با عناوین وابستگی (Association)، خوشه بندی (Clustering) و کاهش ابعاد (Dimensionality Reduction) است.

از خانواده الگوریتم های وابستگی زمانی استفاده می شود که قصد تعیین احتمال وقوع یک اتفاق در اثر وقوع همزمان یک اتفاق دیگر در همان مجموعه را داریم.

برای مثال این الگوریتم می تواند مشخص کند که یک مشتری با توجه به خرید نان از یک فروشگاه اینترنتی، ۸۰% احتمال خرید تخم مرغ نیز دارد.

خانواده مدل های خوشه بندی همانطور که از نامگذاری شان پیداست برای گروه بندی کردن داده ها بر اساس ویژگی های مشابه یکدیگر استفاده می شوند.

برای مثال این الگوریتم می تواند یک مجموعه مقاله را بررسی کرده و آن ها را بر اساس تشابه ساختار کلی به دسته های علمی، ادبی، هنری و... تقسیم کند.

از مدل کاهش ابعاد برای کاهش تعداد متغیرهای یک مجموعه داده استفاده می شود که در عین حال از انتقال اطلاعات مهم آن ها مطمئن است.

الگوریتم های یادگیری ماشین تقویتی (Reinforcement Learning)

الگوریتم های یادگیری تقویتی به یک عامل یا نماینده از مجموعه ای از داده ها این اجازه را می دهد که برای اقدام بعدی خود بهترین تصمیم را بگیرد و این کار را بر اساس یادگیری رفتارهایی که بیشینه پاداش را دارند، انجام می دهد.

این مدل ها معمولا یادگیری را از طریق سعی و خطا پیش می برند تا در نهایت به بهینه ترین اقدام ها دست پیدا کنند.

برای مثال تصور کنید که در یک بازی کامپیوتری یک بازیکن نیاز دارد تا در زمان های دقیقی در یک محل مشخص باشد تا امتیاز کسب کند.

روش الگوریتم های یادگیری ماشین تقویتی برای این بازی شروع به تست کردن تمامی حرکات شخصیت بازی به صورت تصادفی و رندوم می کند و در نهایت بر اساس تمام آزمون و خطاهای انجام شده بهترین و بهینه ترین حرکت های موردنیاز شخصیت داخل بازی را برای رسیدن به بیشترین امتیاز تعیین می کند.

هر یک از این دسته بندی ها شامل چندین نوع الگوریتم با کاربردها و ویژگی های متفاوت هستند.

در ادامه مقاله به معرفی ۱۰ مورد از برترین و پرکاربردترین الگوریتم های یادگیری ماشین می پردازیم.

الگوریتم های یادگیری ماشین

از جمله مهم ترین و پرکاربردترین این الگوریتم ها می توان به موارد زیر اشاره کرد:

رگرسیون خطی (Linear Regression)

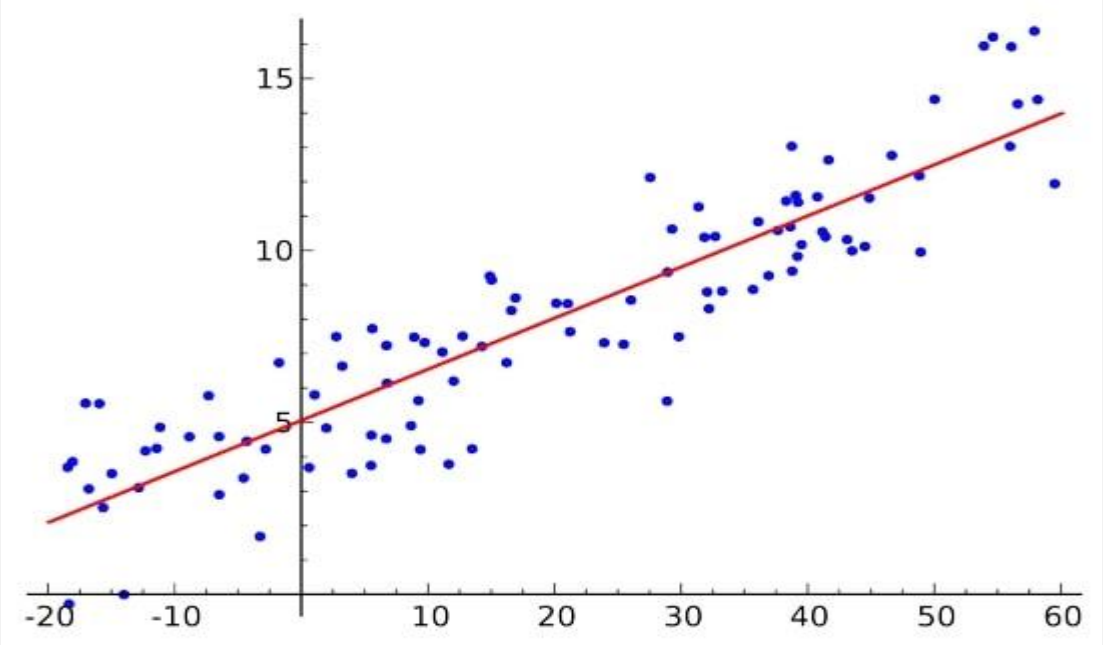
در یادگیری ماشین ما مجموعه ای از داده های ورودی (x) را داریم که استفاده می شوند تا یک متغیر خروجی (y) را مشخص کنند.

همواره یک رابطه میان متغیرهای ورودی و خروجی وجود دارد و هدف اصلی یادگیری ماشین پیدا کردن این رابطه و کمیت بخشیدن به آن است.

در رگرسیون خطی رابطه بین ورودی ها و خروجی با یک معادله خطی ساده به صورت زیر بیان می شود.

$$y = a + bx$$

بنابراین هدف اصلی الگوریتم رگرسیون خطی پیدا کردن مقادیر a و b است.



در واقع این روش سعی می کند بین تمام x و y های موجود در مجموعه دیتاها، یک خط مستقیم رسم کند که کمترین میزان خطای (Error) ممکن را داشته باشد.

متغیرهایی که با استفاده از رگرسیون خطی می توان به آن ها دست پیدا کرد مقادیر پیوسته هستند؛ برای مثال مقدار بارش باران بر حسب سانتی متر. این روش جزو الگوریتم های یادگیری ماشین با نظارت است.

رگرسیون لجستیک (Logistic Regression)

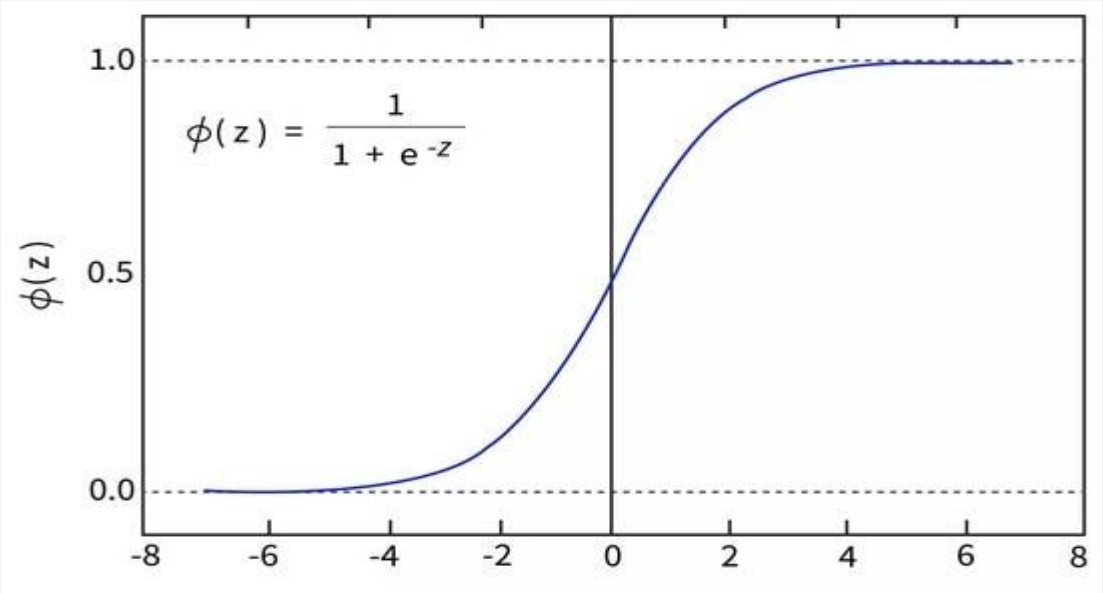
از رگرسیون لجستیک برای داده های گسسته استفاده می شود؛ برای مثال اینکه یک دانش آموزش واحد درسی را قبول می شود یا خیر.

این روش برای داده های باینری بسیار کمک کننده است.

علت نامگذاری رگرسیون لجستیک استفاده آن از تابع تبدیل لجستیک برای تعیین خروجی است.

$$h(x) = 1/(1+e^{-x})$$

بر خلاف رگرسیون خطی که خروجی یک مقدار تعیین شده است، خروجی رگرسیون لجستیک به فرم احتمالات است و به همین دلیل خروجی یک مقدار بین ۰ تا 1 است.



برای مثال فرض کنید که قصد داریم از این الگوریتم برای احتمال بیماری یک شخص استفاده کنیم و مقدار ۱ را برای شخص کاملا بیمار در نظر گرفته ایم.

اگر خروجی این الگوریتم ۰/۹۸ باشد به این معنی است که شخص موردنظر به احتمال قوی بیمار است. این روش هم از نوع الگوریتم های یادگیری ماشین نظارت شده است.

درخت تصمیم گیری (Decision Tree)

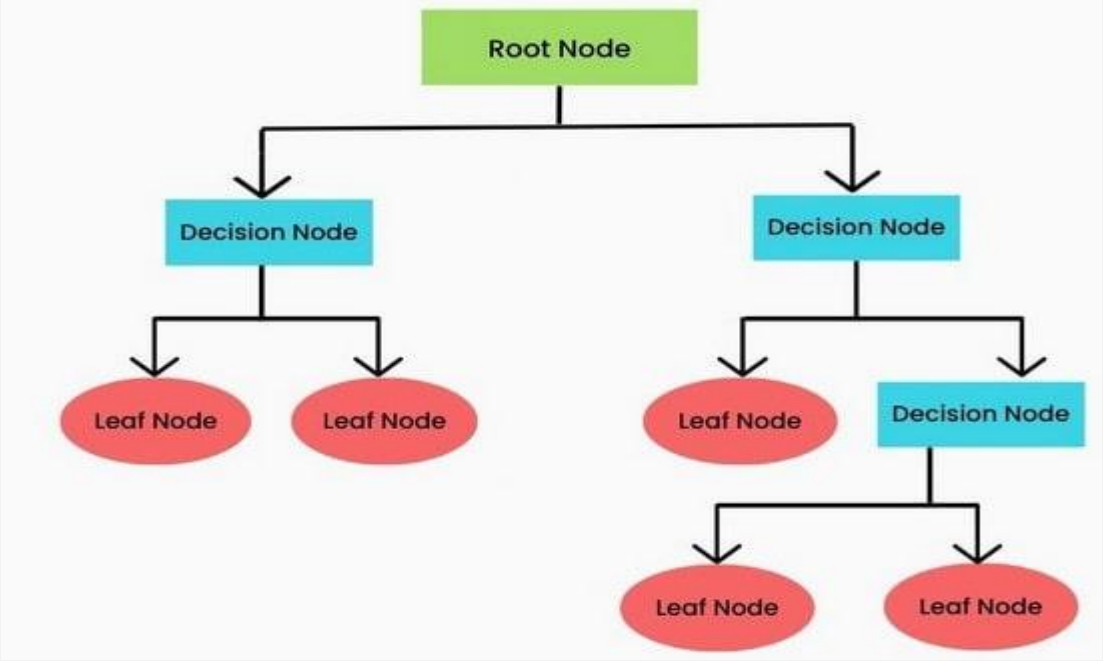
الگوریتم درخت تصمیم گیری یکی از روش های یادگیری ماشین نظارت شده است که برای هر دو نوع داده های پیوسته و گسسته مورداستفاده قرار می گیرد.

در این روش متغیرها بر اساس ویژگی هایشان به صورت شاخ و برگ های یک درخت تقسیم می شوند.

به این صورت که متغیرهای ابتدایی شامل گره های اصلی (Root Node) و گره های داخلی (Internal Node) هستند که در واقع همان ورودی های ما را تشکیل می دهند.

به گره های داخلی، گره تصمیم گیری (Decision Node) نیز گفته می شود.

متغیرهای نهایی که گره های برگ (Leaf Node) نامیده می شوند، متغیرهای خروجی هستند و همان پاسخ سیستم به ورودی مدنظر می باشند.



عملکرد کلی این الگوریتم به این شکل است که ورودی ها را بر اساس گره های مشخص شده دنبال می کند تا به یک گره برگ دست پیدا کند.

بیز ساده (Naive Bayes)

برای محاسبه امکان رخ دادن یک اتفاق با توجه به اینکه یک اتفاق دیگر پیش از آن رخ داده از این روش الگوریتم یادگیری ماشین استفاده می شود.

برای محاسبه احتمال وقوع فرضیه (h) با توجه به دانش قبلی (d) رابطه این الگوریتم به صورت زیر است که در آن P(x) به معنای احتمال رخ دادن پدیده x است.

$$P(h|d) = (P(d|h) P(h)) / P(d)$$

این الگوریتم، ساده نامگذاری شده است؛ چون برای رسیدن به پاسخ این فرضیه ای را در نظر می گیرد که همه متغیرها نسبت به یکدیگر مستقل هستند؛ درحالیکه در مثال های دنیای واقعی این طرز فکر ساده لوحانه است.

برای مثال در جدول زیر بر اساس متغیرهای مشخص شده اگر بخواهیم احتمال بازی کردن (play = yes) در یک روز آفتابی (sunny) را محاسبه کنیم داریم که:

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

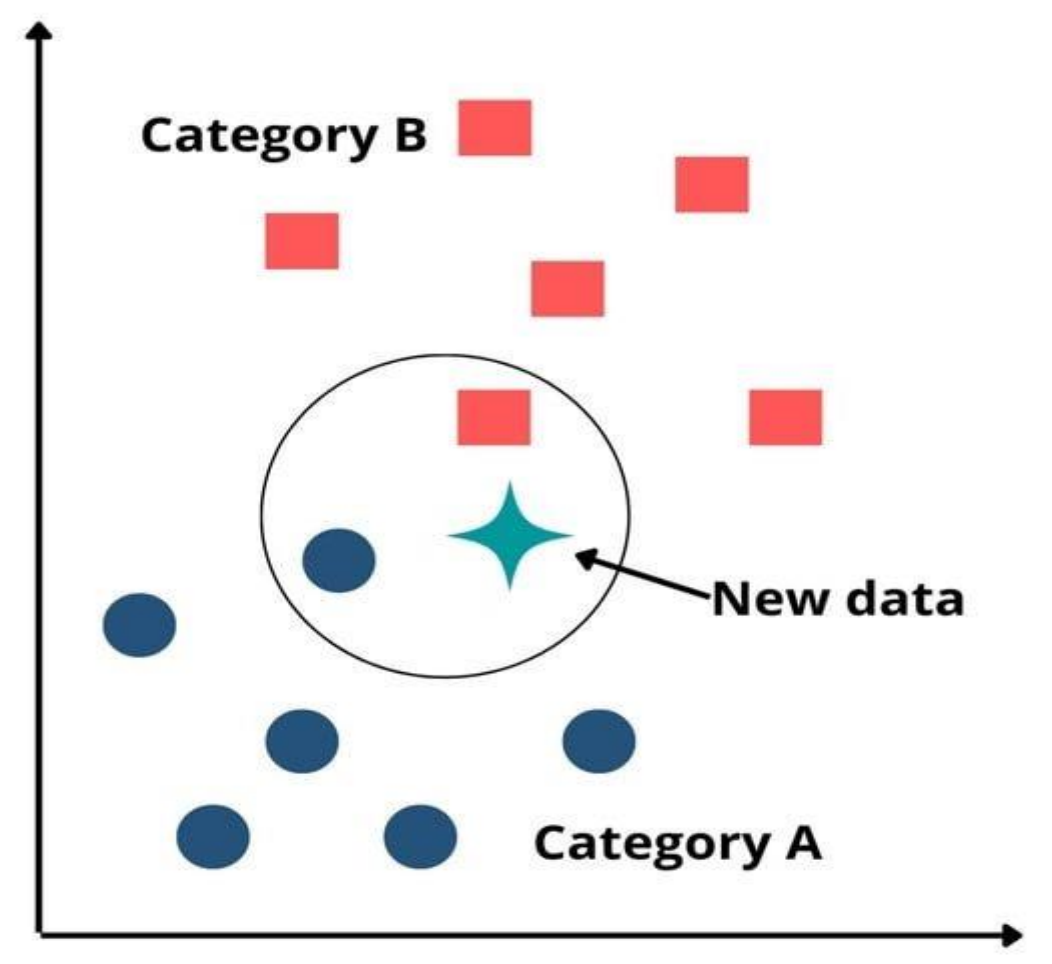
Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

P(yes|sunny) = (P(sunny|yes) * P(yes)) / P(sunny) = (3/9 * 9/14) / (5/14) = 0.60

بنابراین اگر هوا آفتابی باشد خروجی مثبت برای بازی کردن محتمل تر است.

این نوع الگوریتم یادگیری ماشین هم جزو دسته با نظارت است.

نزدیک ترین همسایه کی (K-Nearest Neighbors)



این الگوریتم به جای تقسیم بندی داده ها به دو مجموعه تست و یادگیری، از تمامی مجموعه داده ها به عنوان دیتاهای آموزشی استفاده می کند.

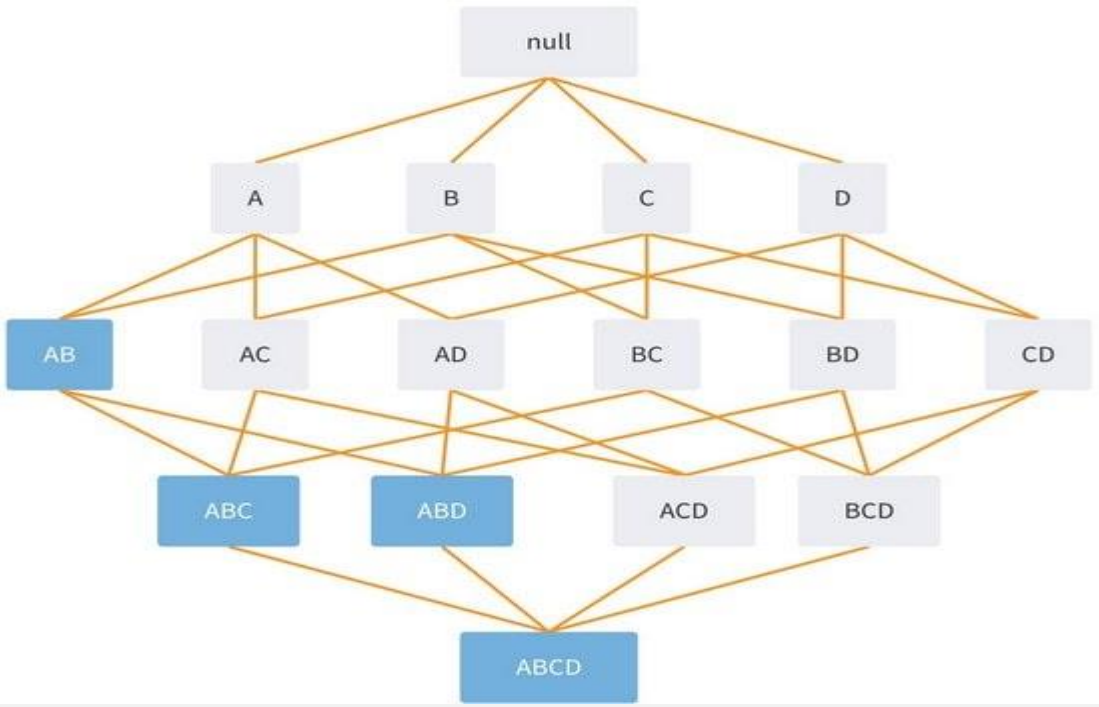
زمانیکه برای یک نمونه داده جدید یک خروجی نیاز است، الگوریتم KNN تمامی داده های موجود را بررسی می کند تا k مورد از نزدیک ترین یا شبیه

ترین نمونه ها به نمونه جدیدمورد نظر را به دست بیاورد و سپس میانگین یا مد آن ها را به عنوان خروجی، معرفی می کند.

مقدار K در این مدل توسط کاربر مشخص می شود. الگوهای سنجش مقدار شباهت بین نمونه ها روش هایی مانند فاصله اقلیدسی و فاصله همینگ هستند.

این روش هم جزو دسته بندی الگوریتم های یادگیری ماشین با نظارت شمرده می شود.

اپریوری (Apriori)



اپریوری یکی از انواع الگوریتم های یادگیری ماشین است که برای دیتابیس های تراکنشی استفاده می شود تا مجموعه داده های متداول و تکراری را شناسایی کرده و سپس قواعدی برای وابستگی تولید کند.

از این روش معمولا برای تحلیل های سبد فروش و بازار استفاده می شود.

برای نمایش قاعده وابستگی اینکه اگر یک شخص آیتم X را خرید کرده باشد در نتیجه آیتم Y را نیز تهیه می کند از علامت X -> Y استفاده می شود.

برای مثال اگر یک شخص شکر (sugar) و شیر (milk) بخرد به احتمال زیاد پودر قهوه (coffee powder) نیز می خرد و قاعده آن به این صورت نوشته می شود که:

{milk,sugar} -> coffee powder

همانطور که احتمالا حدس زده اید این روش جزو خانواده الگوریتم های یادگیری ماشین بدون نظارت است.

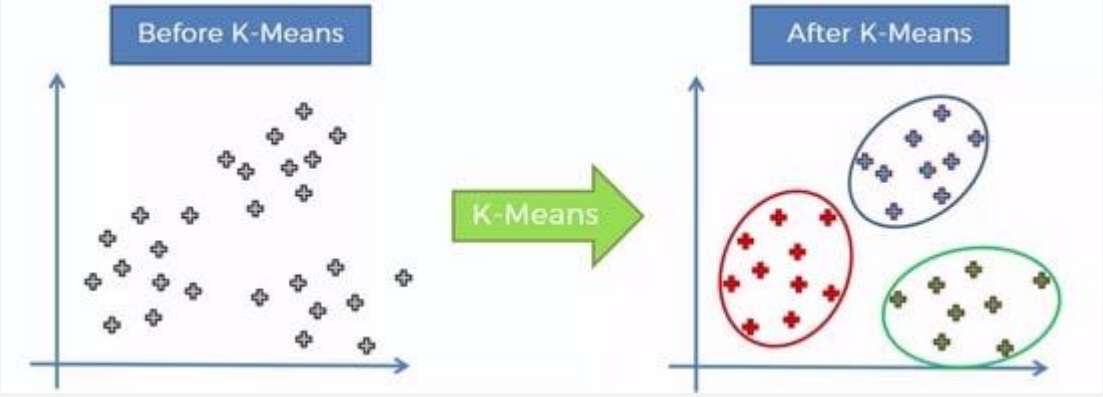
میانگین کی (K-means)

الگوریتم میانگین K یک مدل تکرارشونده است که داده های مشابه را به خوشه ها گروه بندی می کند.

برای این خوشه بندی یک مقدار مرکزی k را مشخص می کند و سپس داده هایی که در کمترین فاصله از این مقدار هستند را جزو یک خوشه در نظر می گیرد.

روش کلی عملکرد این الگوریتم بدون نظارت به این صورت است که ابتدا یک عدد برای k در نظر گرفته می شود و سپس برای هر شاخه یک عدد رندوم مرکزی محاسبه می شود.

در ادامه دیتاهای نزدیک به این اعداد مرکزی مشخص شده و یک خوشه را تشکیل می دهند و سپس برای خوشه تشکیل شده یک عدد مرکزی نزدیک تر به آن ها در نظر گرفته می شود و این دو مرحله به قدری تکرار می شود که تغییر حالتی در خوشه های ایجاد شده به وجود نیاید.

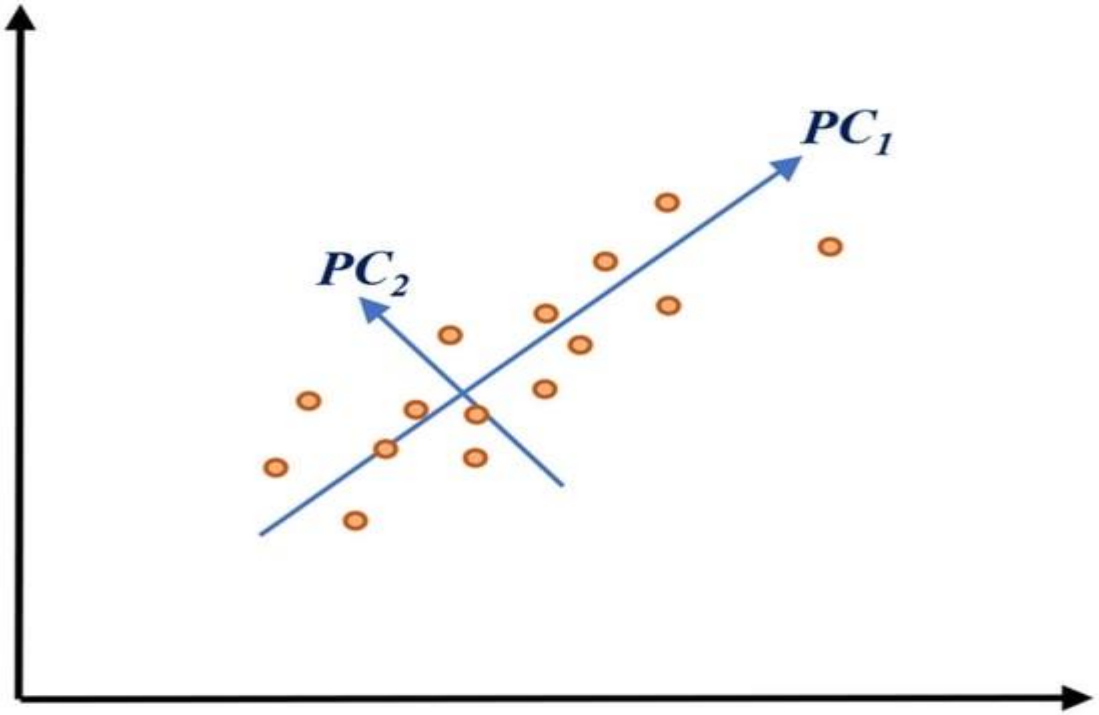


این روش هم جزو خانواده الگوریتم های یادگیری ماشین بدون نظارت است.

تحلیل مؤلفه های اصلی (Principal Component Analysis)

الگوریتم PCA به این منظور استفاده می شود که داده ها را با استفاده از کاهش تعداد متغیرها، برای بررسی ساده تر و قابل تجسم کند. این مدل برای این منظور حداکثر واریانس موجود در داده ها را در یک دستگاه مختصات جدید با محورهای مؤلفه های اصلی ثبت می کند.

هر مؤلفه یک ترکیب خطی از متغیرهای اصلی مجموعه داده هاست و بر دیگر مؤلفه ها عمود است که این تعامد نمایشگر صفر بودن مقدار همبستگی داده ها بین این دو مؤلفه است.



مؤلفه اول اصلی جهت بیشینه تغییرپذیری موجود در داده ها را ثبت می کند.

مؤلفه دوم اصلی واریانس باقی مانده در داده ها را ثبت می کند؛ اما متغیرها در آن با مؤلفه اول نامرتبط هستند.

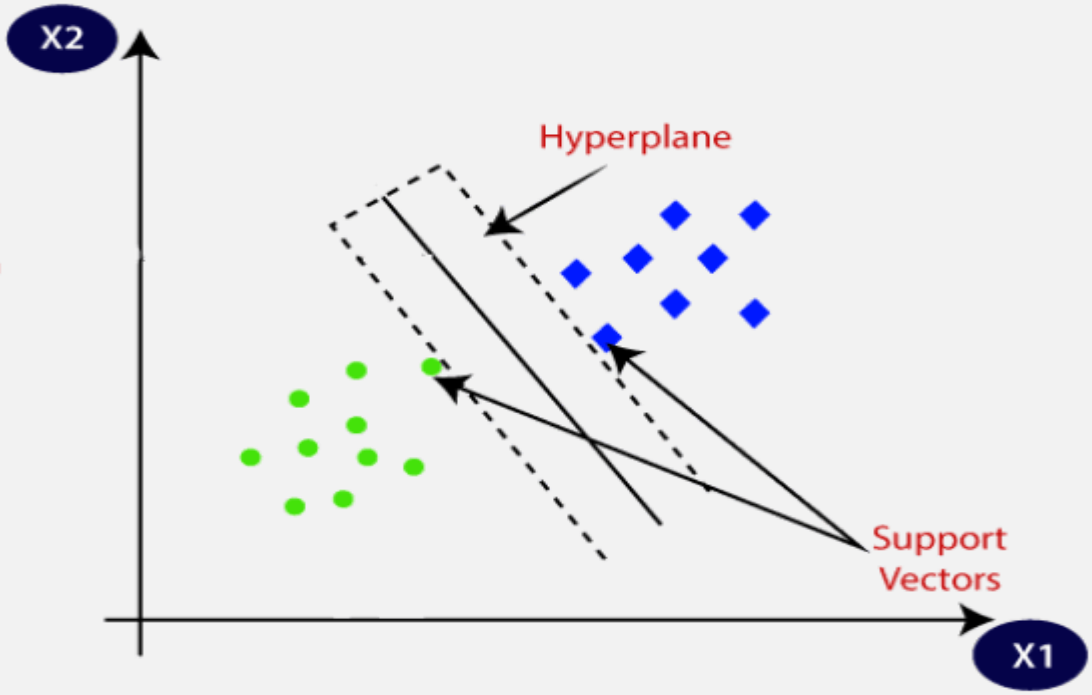
این روش نیز از خانواده الگوریتم های بدون نظارت است.

ماشین بردار پشتیبان Vector (Support Machine)

SVM یک روش الگوریتم با نظارت دیگر است که عموما برای مسائل طبقه بندی استفاده می شود.

برای استفاده از SVM داده های خام در یک فضای n بعدی به صورت نقاط پراکنده تعریف می شوند n) .تعداد ویژگی های موجود برای داده ها است(

هدف اصلی این الگوریتم ایجاد یک ابر صفحه (Hyperplane) یا مرز تصمیم گیری (Decision Boundary) است که بتوان به وسیله آن داده های مختلف یک مجموعه دیتا را تفکیک کرد.



نقاط داده ای که برای تعیین ابر صفحه کمک کننده هستند به نام بردارهای پشتیبانی (Support Vectors) شناخته می شوند.

برای مثال از کاربردهای این مدل در دنیای واقعی می توان به تشخیص چهره، طبقه بندی تصاویر، دسته بندی متن و... اشاره کرد.

جنگل تصادفی (Random Forest)

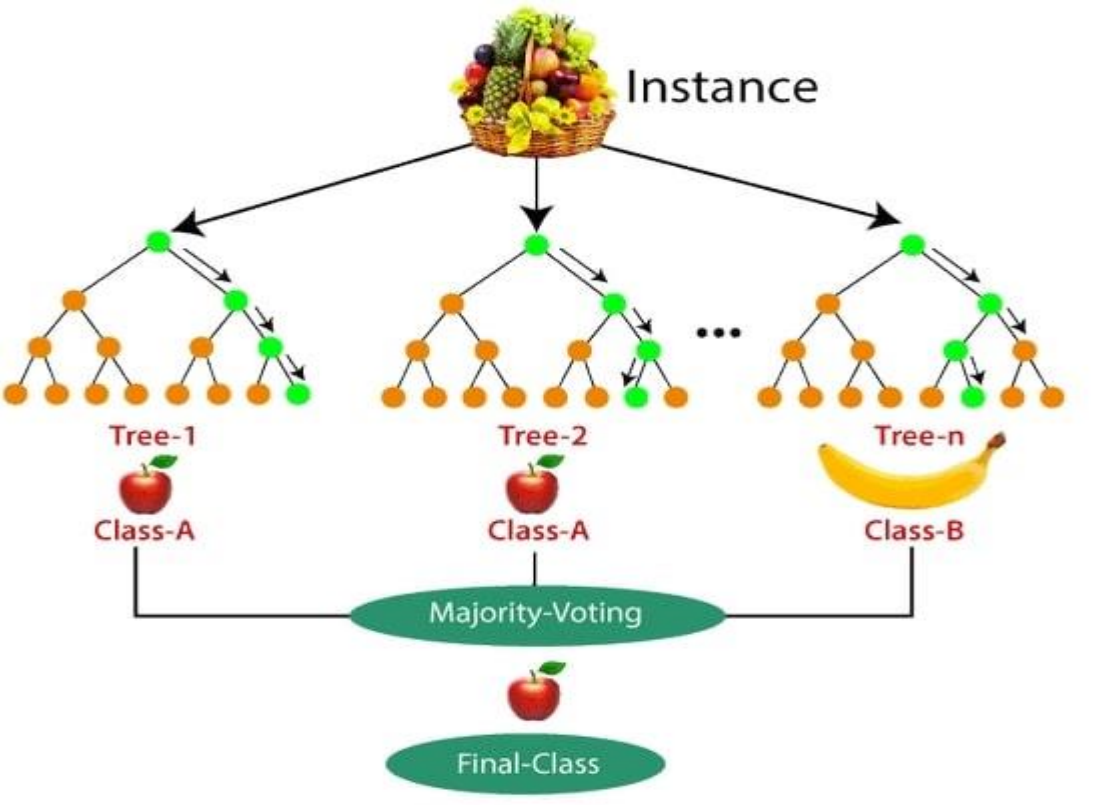
جنگل تصادفی یکی از انواع الگوریتم های یادگیری ماشین نظارت شده است که می تواند برای هر دو بخش یادگیری ماشین طبقه بندی و رگرسیون استفاده شود.

این روش یک تکنیک یادگیری گروهی است که با استفاده از ترکیب طبقه بندی کننده ها و بهبود عملکرد مدل، پیش بینی می کند.

الگوریتم جنگل تصادفی شامل چندین درخت تصادفی برای زیرمجموعه های داده های مشخص شده است و با استفاده از میانگین گیری به بهبود دقت پیش بینی کمک می کند.

هر جنگل تصادفی باید شامل ٦٤ تا ١٢٨ درخت باشد؛ هرچه تعداد درخت ها بیشتر باشد، دقت خروجی بالاتر خواهد بود.

برای طبقه بندی یک مجموعه داده یا شی جدید، هر درخت یک نتیجه می سازد و بر اساس اکثریت آرا الگوریتم جنگل تصادفی خروجی نهایی را پیش بینی می کند.



سخن پایانی

در این مقاله سعی کردیم به بیان ساده شما را با مهم ترین الگوریتم های یادگیری ماشین که در دنیای هوشمندسازی امروز در سراسر دنیا استفاده می شوند، آشنا کنیم.

بکارگیری این ابزارهای حرفه ای به بشر در علوم و صنایع مختلفی کمک کرده است، حتی بشر به این توانایی دست پیدا کرده که برای تشخیص بیماری یک فرد از روی علائم و نتایج آزمایشات آن از سیستم های یادگیری ماشین استفاده کند.