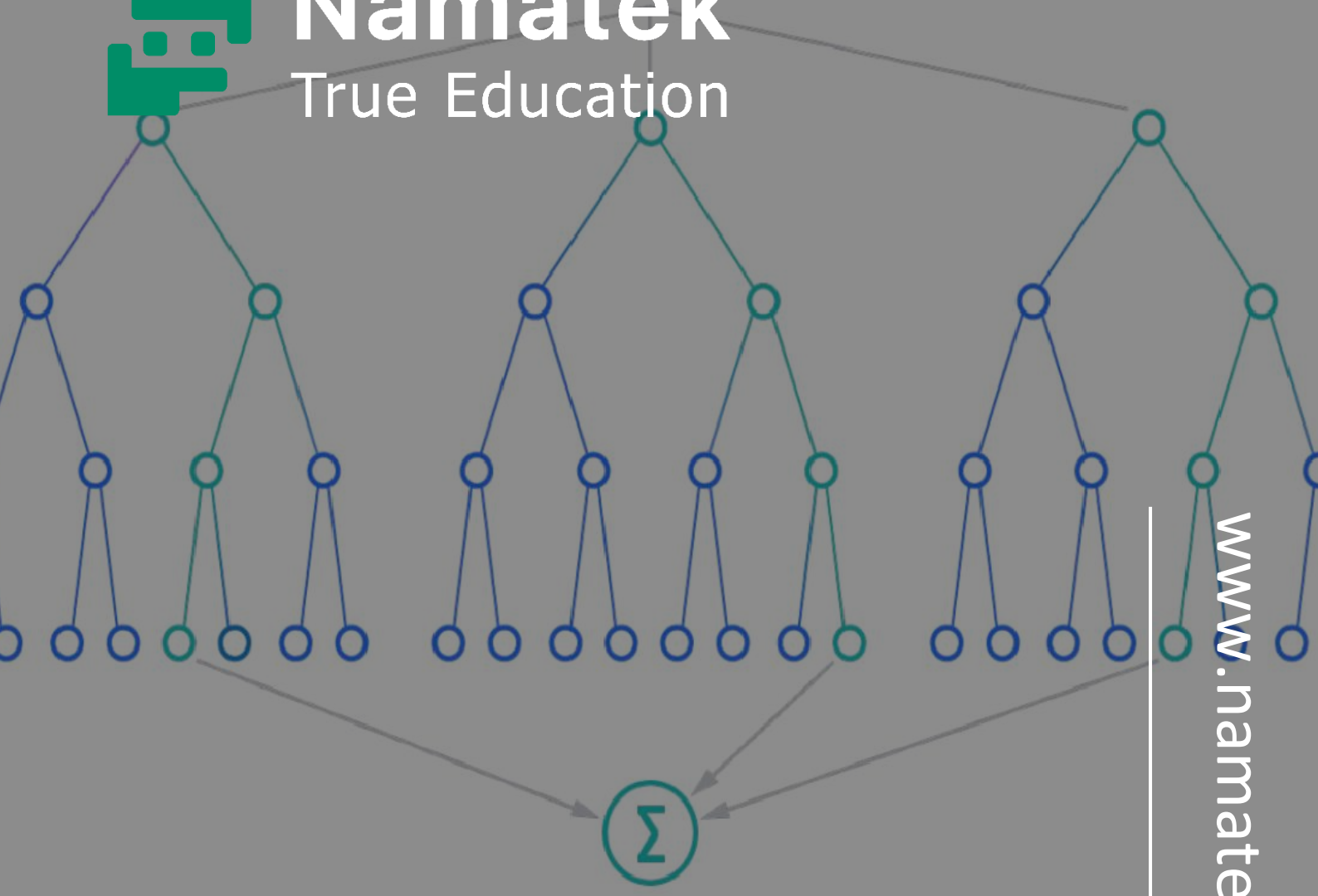
Namatek
True Education




# Random Forest

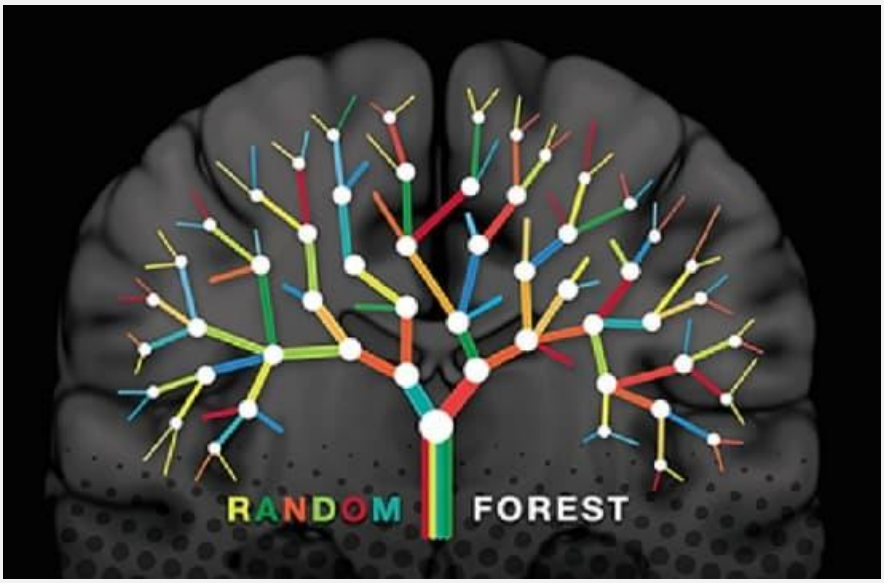
www.namatek.com

# جنگل تصادفی

# فهرست مطالب

یادگیری ماشین، ترکیبی شگفت‌انگیز از علم کامپیوتر و آمار، مرسوم‌ترین فناوری در تکنولوژی امروزی است که به عنوان زیرشاخه هوش مصنوعی در نظر گرفته می‌شود. با گذشت زمان، یادگیری ماشین شاهد پیشرفت‌های باورنکردنی بوده است. یکی از الگوریتم‌های برجسته در حوزه یادگیری ماشین، جنگل تصادفی است. جنگل‌های تصادفی یا درخت‌های تصمیم تصادفی تیمی مشترک از درخت‌های تصمیم‌گیری هستند که با هم کار می‌کنند تا یک خروجی واحد را ارائه دهند. Random Forest که در سال ۲۰۰۱ از طریق Leo Breiman آغاز شد، به سنگ بنای علاقه‌مندان به یادگیری ماشین تبدیل شده است. در این مقاله به بررسی اصول و روش پیاده‌سازی الگوریتم جنگل تصادفی می‌پردازیم.

## الگوریتم جنگل تصادفی چیست؟




الگوریتم جنگل تصادفی یک تکنیک قدرتمند یادگیری درختی در یادگیری ماشینی است که با ایجاد تعدادی درخت تصمیم در مرحله آموزش کار می‌کند. به عبارت دیگر جنگل تصادفی نوعی الگوریتم یادگیری گروهی است که با ترکیب چندین درخت تصمیم (Decision Tree) بر روی زیر

مجموعه‌های تصادفی از داده‌ها و ویژگی‌ها، عملکرد خود را بهبود می‌بخشد. یکی از مزایای اصلی این الگوریتم کاهش مشکل بیش‌برازش (Overfitting) است که معمولاً در درخت‌های تصمیم منفرد اتفاق می‌افتد. هر درخت با استفاده از یک زیرمجموعه تصادفی از مجموعه داده‌ها برای اندازه‌گیری زیرمجموعه تصادفی از ویژگی‌ها در هر پارتیشن ساخته می‌شود. این حالت تصادفی، تنوع را در بین درختان منفرد معرفی می‌کند و عملکرد کلی پیش‌بینی را بهبود می‌بخشد. در پیش‌بینی، الگوریتم نتایج همه درخت‌ها را با رأی دادن برای وظایف طبقه‌بندی یا با میانگین‌گیری برای وظایف رگرسیون تجمیع می‌کند. جنگل‌های تصادفی به طور گسترده برای طبقه‌بندی و توابع رگرسیون استفاده می‌شوند که به دلیل توانایی آن‌ها در مدیریت داده‌های پیچیده و ارائه پیش‌بینی‌های قابل‌اعتماد در محیط های مختلف شناخته شده است.

## روند انجام الگوریتم جنگل تصادفی

الگوریتم جنگل تصادفی در چندین مرحله کار می‌کند که در زیر مورد بحث قرار می‌گیرد:

## تعریف و به وجود آوردن درختان تصمیم‌گیری

Random Forest از قدرت یادگیری گروهی با ساخت مجموعه‌ای از درختان تصمیم استفاده می‌کند. این درختان مانند افراد متخصصی هستند که هر کدام در جنبه خاصی از داده‌ها تخصص دارند. مهم‌تر از همه، آن‌ها به طور مستقل عمل می‌کنند و خطر تحت تاثیر قرار گرفتن بیش از حد مدل توسط تفاوت‌های ظریف یک درخت را به حداقل می‌رسانند.

## انتخاب ویژگی تصادفی

برای اطمینان از این‌که هر درخت تصمیم در مجموعه چه چشم‌انداز منحصربه‌فردی به ارمغان می‌آورد، Random Forest از انتخاب ویژگی تصادفی استفاده می‌کند. در طول آموزش هر درخت، یک زیرمجموعه تصادفی از ویژگی‌ها انتخاب می‌شود و این تصادفی بودن تضمین می‌کند که هر درخت روی جنبه‌های مختلف داده‌ها تمرکز می‌کند و مجموعه‌ای از پیش‌بینی‌کننده‌ها را در مجموعه ایجاد می‌کند.

- **Bagging یا Bootstrap Aggregating**

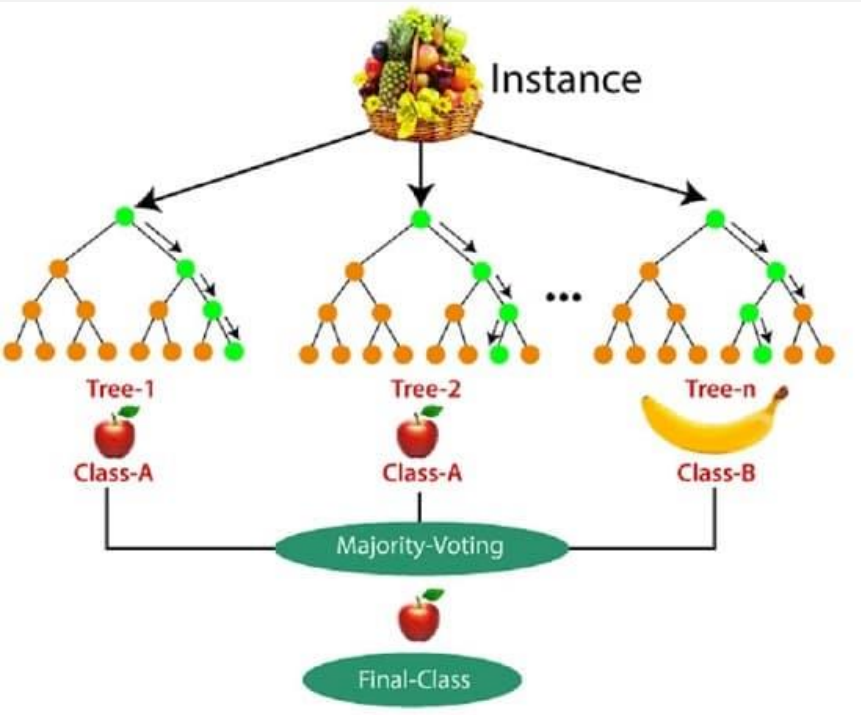
تکنیک bagging سنگ بنای استراتژی آموزشی الگوریتم جنگل تصادفی است که شامل ایجاد چندین نمونه بوت استرپ از مجموعه داده اصلی می‌باشد که امکان نمونه‌برداری از نمونه‌ها را با جایگزینی فراهم می‌کند. این تکنیک منجر به زیرمجموعه‌های متفاوتی از داده‌ها برای هر درخت تصمیم می‌شود که تنوع را در فرآیند آموزش معرفی می‌کند و مدل را قوی‌تر می‌کند.

## تصمیم گیری و رای گیری

وقتی نوبت به پیش‌بینی می‌رسد، هر درخت تصمیم در جنگل تصادفی رای خود را می‌دهد. برای کارهای طبقه‌بندی، پیش‌بینی نهایی با حالت متداول‌ترین پیش‌بینی در همه درختان تعیین می‌شود. در فرآیند رگرسیون، میانگین پیش‌بینی‌های درختی فردی گرفته می‌شود.

این مکانیسم رأی‌گیری داخلی، فرآیند تصمیم‌گیری متوازن و جمعی را تضمین می‌کند. با مثال زیر عملکرد الگوریتم جنگل تصادفی را می‌توان بهتر درک کرد: فرض کنید مجموعه داده‌ای وجود دارد که حاوی چندین تصویر

میوه است و این مجموعه داده به طبقه‌بندی‌کننده جنگل تصادفی داده می‌شود. مجموعه داده مفروض به زیر مجموعه‌ها تقسیم می‌شود و به هر درخت تصمیم داده می‌شود. در طول مرحله آموزش، هر درخت تصمیم یک نتیجه پیش‌بینی ایجاد می‌کند و زمانی که یک نقطه داده جدید رخ می‌دهد، بر اساس اکثر نتایج، طبقه‌بندی جنگل تصادفی تصمیم نهایی را پیش‌بینی می‌کند. در این راستا می‌توانید تصویر زیر را در نظر بگیرید:




## ویژگی‌های کلیدی جنگل تصادفی

فرآیند الگوریتم جنگل تصادفی برای طبقه‌بندی ویژگی‌های متفاوتی را ارائه می‌دهد. برخی از ویژگی‌های کلیدی Random Forest در زیر مورد بحث قرار گرفته‌اند:

## دقت پیش‌بینی بالا

الگوریتم جنگل تصادفی را به‌عنوان تیمی از جادوگران تصمیم‌گیری تصور کنید.هر درخت تصمیم به بخشی از مشکل نگاه می‌کند و با هم بینش‌های خود را در یک تابلوی پیش‌بینی قدرتمند می‌بافند. این کار تیمی اغلب منجر به مدل دقیق‌تری می‌شود.

## مقاومت در برابر تطبیق بیش از حد

جنگل تصادفی مانند یک مربی خونسرد است که شاگردان خود (درختان تصمیم) را راهنمایی می‌کند. به‌جای اینکه به هر شاگرد اجازه دهد تمام جزئیات آموزش خود را به‌خاطر بسپارند، درک جامع‌تری را در بین شاگردان ایجاد می‌نماید که این رویکرد به جلوگیری از درگیرشدن بیش از حد با داده‌های آموزشی کمک می‌کند و باعث می‌شود مدل کمتر مستعد بیش از حد برازش شود.

## مدیریت مجموعه داده‌های بزرگ

اگر با کوهی از داده‌ها سروکار دارید، جنگل تصادفی مانند یک کاوشگر کارکشته با تیمی از یاران (درختان تصمیم) با آن مقابله می‌کند. هر کمک‌کننده بخشی از مجموعه داده را بر عهده می‌گیرد و اطمینان می‌دهد که این فرآیند نه‌تنها کامل است؛ بلکه به طرز شگفت‌آوری سریع است.

## ارزیابی اهمیت متغیر

جنگل تصادفی را به‌عنوان یک کارآگاه در صحنه جرم در نظر بگیرید که می‌فهمد کدام سرنخ‌ها و ویژگی‌ها بیشترین اهمیت را دارند. این الگوریتم، اهمیت هر سرنخ را در حل پرونده ارزیابی می‌کند و به شما کمک می‌کند روی عناصر کلیدی که پیش‌بینی‌ها را هدایت می‌کنند، تمرکز کنید.

## مقیاس‌بندی و عادی‌سازی

در حالی که جنگل تصادفی به مقیاس‌بندی ویژگی حساس نیست، عادی‌سازی ویژگی‌های عددی همچنان می‌تواند به فرآیند آموزشی کارآمدتر و بهبود همگرایی کمک کند.

## اعتبارسنجی داخلی

جنگل تصادفی مانند داشتن یک مربی شخصی است که شما را کنترل می‌کند و همانطور که هر درخت تصمیم را آموزش می‌دهد، یک گروه مخفی از موارد را نیز برای آزمایش کنار می‌گذارد که این مورد اعتبارسنجی داخلی را تضمین می‌نماید. در نتیجه مدل شما نه‌تنها در آموزش کوتاهی نمی‌کند؛ بلکه در چالش‌های جدید نیز عملکرد خوبی دارد.

## مدیریت ارزش‌های گمشده

زندگی درست مانند مجموعه داده‌هایی با مقادیر گمشده جنگل تصادفی پر از عدم قطعیت است. این الگوریتم، دوستی است که با موقعیت وفق داده می‌شود و با استفاده از اطلاعات موجود پیش‌بینی می‌کند. با قطعات

ازدست‌رفته دچار آشفتگی نمی‌شود و در عوض، بر آنچه می‌تواند با اطمینان به ما بگوید، تمرکز می‌کند.

## موازی‌سازی برای سرعت

الگوریتم جنگل تصادفی دوست شماست که در زمان صرفه‌جویی می‌کند. هر درخت تصمیم را به‌عنوان کارگری تصور کنید که به طور همزمان با تکه‌ای از یک پازل برخورد می‌کند. این رویکرد موازی از قدرت فناوری مدرن بهره می‌برد و کل فرآیند را برای انجام پروژه‌های در مقیاس بزرگ سریع‌تر و کارآمدتر می‌کند.

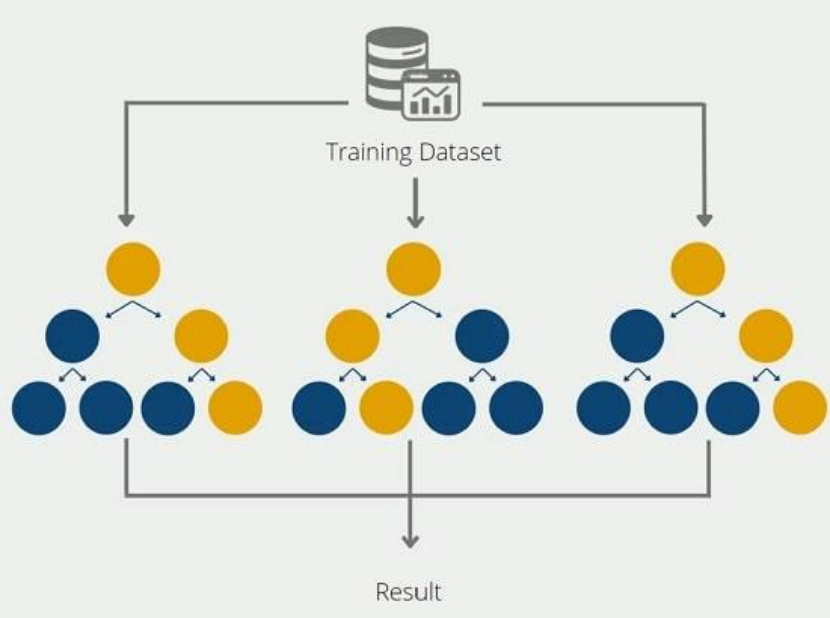
## پرداختن به داده‌های نامتعادل

اگر با کلاس‌های نامتعادل سروکار دارید، تکنیک‌هایی مانند تنظیم وزن کلاس یا استفاده از روش‌های نمونه‌گیری مجدد برای اطمینان از نمایش متعادل در طول فرآیند آموزش الگوریتم اجرا می‌گردد.

## رمزگذاری متغیرهای طبقه‌بندی

جنگل تصادفی به ورودی‌های عددی نیاز دارد؛ بنابراین متغیرهای طبقه‌بندی باید کدگذاری شوند. تکنیک‌هایی مانند رمزگذاری، ویژگی‌های دسته‌بندی را به قالبی مناسب برای الگوریتم تبدیل می‌کنند.

## مزایای جنگل تصادفی




تعدادی از مزایا و چالش‌های کلیدی وجود دارد که الگوریتم جنگل تصادفی هنگام استفاده برای مسائل طبقه‌بندی یا رگرسیون ارائه می‌کند. برخی از آن‌ها عبارت‌اند از:

### کاهش میزان بیش برازش (overfitting)

درخت‌های تصمیم‌گیری امکان وقوع برازش بیش از حد را دارند؛ زیرا تمایل دارند تمام نمونه‌ها را در داده‌های آموزشی کاملاً منطبق کنند. با این حال، هنگامی که تعداد زیادی درخت تصمیم در یک جنگل تصادفی وجود دارد، وجود طبقه‌بندی‌کننده بیش از حد با مدل سازگاری ندارد؛ زیرا میانگین‌گیری درخت‌های همبسته، واریانس کلی و خطای پیش‌بینی را کاهش می‌دهد.

### انعطاف‌پذیری

از آنجایی که جنگل تصادفی می‌تواند هر دو وظایف رگرسیون و طبقه‌بندی را با درجه بالایی از دقت انجام دهد، روشی محبوب در میان دانشمندان

داده است. بسته‌بندی ویژگی همچنین طبقه‌بندی‌کننده جنگل تصادفی را به ابزاری مؤثر برای تخمین مقادیر گمشده تبدیل می‌کند؛ زیرا دقت را در مواقعی که بخشی از داده‌ها از دست می‌دهند، حفظ می‌کند.

## تعیین آسان اهمیت ویژگی

جنگل تصادفی ارزیابی اهمیت متغیر یا مشارکت در مدل را آسان می‌کند. چند راه برای ارزیابی اهمیت ویژگی وجود دارد و شاخص جینی و کاهش میانگین ناخالصی (MDI) معمولاً برای اندازه‌گیری میزان کاهش دقت مدل زمانی که یک متغیر معین حذف می‌شود، استفاده می‌شود. با این حال، شاخص جایگشت، (همچنین به‌عنوان دقت کاهش میانگین (MDA) شناخته می‌شود) یکی دیگر از معیارهای بااهمیت است.

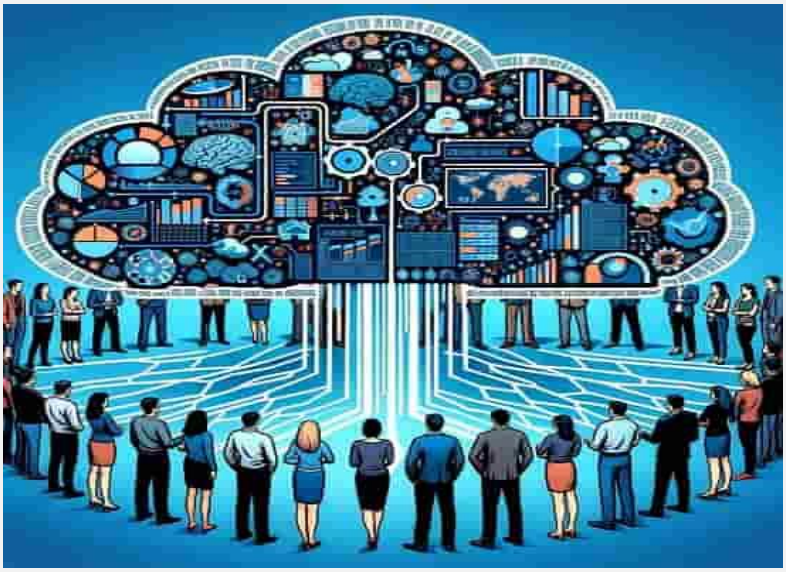
MDA میانگین کاهش دقت را با تغییر تصادفی مقادیر ویژگی در نمونه‌های oob شناسایی می‌کند.

## چالش‌های جنگل تصادفی

- **فرآیند زمان‌بر**: ازآنجایی‌که الگوریتم‌های جنگل تصادفی می‌توانند مجموعه داده‌های بزرگ را مدیریت کنند و پیش‌بینی‌های دقیق‌تری ارائه کنند، ممکن است در پردازش داده‌ها کند باشند؛ زیرا در حال محاسبه داده‌ها برای هر درخت تصمیم هستند.
- **داشتن نیاز به منابع بیشتر**: با توجه به این که جنگل‌های تصادفی مجموعه داده‌های بزرگ‌تری را پردازش می‌کنند، به منابع بیشتری برای ذخیره آن داده‌ها نیاز دارند.

- **پیچیده بودن فرآیند**: تفسیر یک درخت تصمیم منفرد در مقایسه با نوع جنگلی آسان‌تر است.

## کاربردهای الگوریتم جنگل تصادفی



امروزه الگوریتم جنگل تصادفی در صنایع مختلفی اعمال شده است و به آن‌ها اجازه می‌دهد تصمیمات تجاری بهتری بگیرند. برخی از موارد استفاده عبارت‌اند از:

- **مالی:** این یک الگوریتم ترجیحی نسبت به سایرین است؛ زیرا زمان صرف شده برای مدیریت داده و وظایف پیش‌پردازش را کاهش می‌دهد. می‌توان از آن برای ارزیابی مشتریان با ریسک اعتباری بالا، کشف تقلب و مشکلات قیمت‌گذاری استفاده کرد.
- **مراقبت‌های بهداشتی:** الگوریتم جنگل تصادفی دارای کاربردهایی در زیست‌شناسی است و با مشکلاتی مانند طبقه‌بندی بیان ژن، کشف نشانگرهای زیستی و حاشیه‌نویسی توالی مقابله می‌کند. همچنین پزشکان می‌توانند در مورد تخمین پاسخ دارویی به داروهای خاص از این الگوریتم استفاده کنند.

- **تجارت الکترونیک:** از الگوریتم جنگل تصادفی می‌توان برای موتورهای توصیه برای اهداف فروش متقابل استفاده کرد.

## جنگل تصادفی در مقابل سایر الگوریتم‌های یادگیری ماشین

جنگل تصادفی در مقایسه با سایر الگوریتم‌های یادگیری ماشین برتری‌هایی را ارائه می‌دهد. برخی از تفاوت‌های کلیدی در زیر مورد بحث قرار می‌گیرند.

### جنگل تصادفی

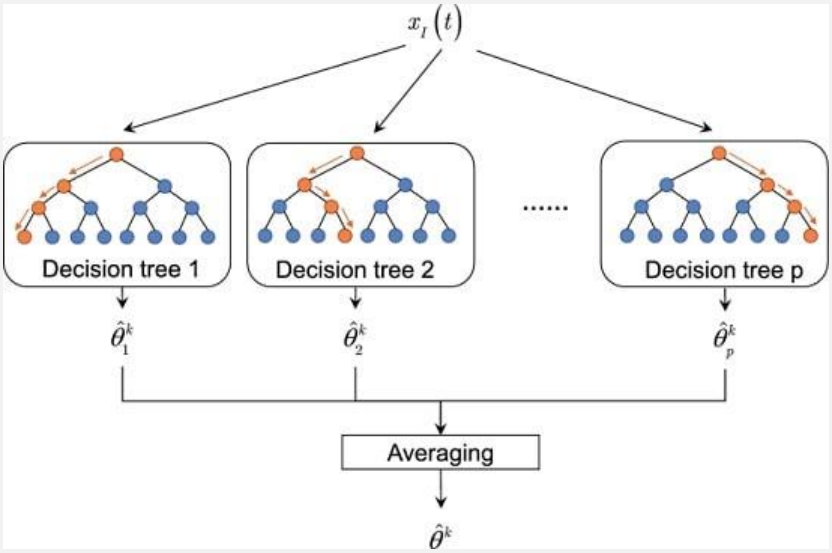
- **رویکرد گروهی:** از مجموعه‌ای از درختان تصمیم استفاده می‌کند و خروجی‌های آن‌ها را برای پیش‌بینی‌ها ترکیب می‌کند، استحکام و دقت را تقویت می‌نماید.
- **مقاومت در برابر برازش بیش از حد:** به دلیل تجمع درختان تصمیم‌گیری متنوع، از به‌خاطر سپردن داده‌های آموزشی در برابر برازش بیش از حد، مقاوم است.
- **رسیدگی به داده‌های ازدست‌رفته:** با استفاده از ویژگی‌های موجود برای پیش‌بینی‌ها، انعطاف‌پذیری را در مدیریت مقادیر ازدست‌رفته نشان می‌دهد و به عملی‌بودن در سناریوهای دنیای واقعی کمک می‌کند.
- **اهمیت متغیر:** یک مکانیسم داخلی برای ارزیابی اهمیت متغیر، کمک به انتخاب ویژگی و تفسیر عوامل تأثیرگذار ارائه می‌کند.
- **افزایش پتانسیل موازی‌سازی:** جنگل تصادفی از موازی‌سازی استفاده می‌کند و آموزش هم‌زمان درخت‌های تصمیم را امکان‌پذیر می‌کند که

در نتیجه محاسبات سریع‌تر برای مجموعه داده‌های بزرگ انجام می‌شود.

## سایر الگوریتم‌های یادگیری ماشین

- **رویکرد گروهی:** معمولاً به یک مدل واحد (مثلاً رگرسیون خطی، ماشین بردار پشتیبان) بدون رویکرد مجموعه متکی هستند که به طور بالقوه منجر به انعطاف‌پذیری کمتر در برابر نویز می‌شوند.
- **مقاومت در برابر برازش بیش از حد:** ممکن است به‌دلیل سازگاری بیش از حد با نویز آموزشی، برخی از الگوریتم‌ها مستعد برازش بیش از حد باشند. (به‌ویژه زمانی که با مجموعه داده‌های پیچیده سروکار دارند).
- **رسیدگی به داده‌های ازدست‌رفته:** سایر الگوریتم‌ها ممکن است نیاز به انتساب یا حذف داده‌های ازدست‌رفته داشته باشند که به طور بالقوه بر آموزش و عملکرد مدل تأثیر می‌گذارد.
- **اهمیت متغیر:** بسیاری از الگوریتم‌ها ممکن است فاقد ارزیابی صریح اهمیت ویژگی باشند که این مسئله، شناسایی متغیرهای حیاتی برای پیش‌بینی‌ها را به چالش می‌کشد.
- **افزایش پتانسیل موازی‌سازی:** برخی از الگوریتم‌ها ممکن است قابلیت‌های موازی‌سازی محدودی داشته باشند که این مورد، به طور بالقوه منجر به زمان‌های آموزشی طولانی‌تر برای مجموعه داده‌های گسترده می‌شود.

# غلبه بر چالش‌ها در مدل‌سازی تصادفی جنگل




برای استفاده بسیار کارآمد از الگوریتم جنگل تصادفی در برنامه‌های کاربردی دنیای واقعی، باید بر برخی چالش‌های بالقوه غلبه کنیم که در زیر مورد بحث قرار می‌گیرند:

## حل مشکل برازش بیش از حد

کم کردن تمایل درختان تصمیم‌گیری فردی به برازش بیش از حد یک چالش مهم است که برای این منظور، استراتژی‌هایی مانند تنظیم فراپارامترها، تنظیم عمق درخت و اجرای تکنیک‌های انتخاب ویژگی برای ایجاد تعادل مناسب بین پیچیدگی و تعمیم انجام می‌پذیرد.

## بهینه‌سازی منابع محاسباتی

کارایی Random Forest در مدیریت مجموعه داده‌های بزرگ گاهی اوقات می‌تواند یک شمشیر دو لبه باشد که به منابع محاسباتی قابل توجهی نیاز دارد.

پیاده‌سازی تکنیک‌های موازی‌سازی و کاوش الگوریتم‌های بهینه‌شده، گام‌های کلیدی در غلبه بر چالش‌های محاسباتی و اطمینان از مقیاس‌پذیری هستند.

## برخورد با داده‌های نامتعادل

وقتی با مجموعه داده‌های نامتعادل مواجه می‌شویم، جایی که یک کلاس به طور قابل‌توجهی بر دیگری برتری دارد، جنگل تصادفی ممکن است به سمت گروه اکثریت منحرف شود. کاهش این سوگیری شامل استراتژی‌هایی مانند تنظیم وزن کلاس، نمونه‌برداری بیش از حد از زیر کلاس‌ها یا استفاده از الگوریتم‌های ویژه برای مقابله با موقعیت‌های نامتعادل است.

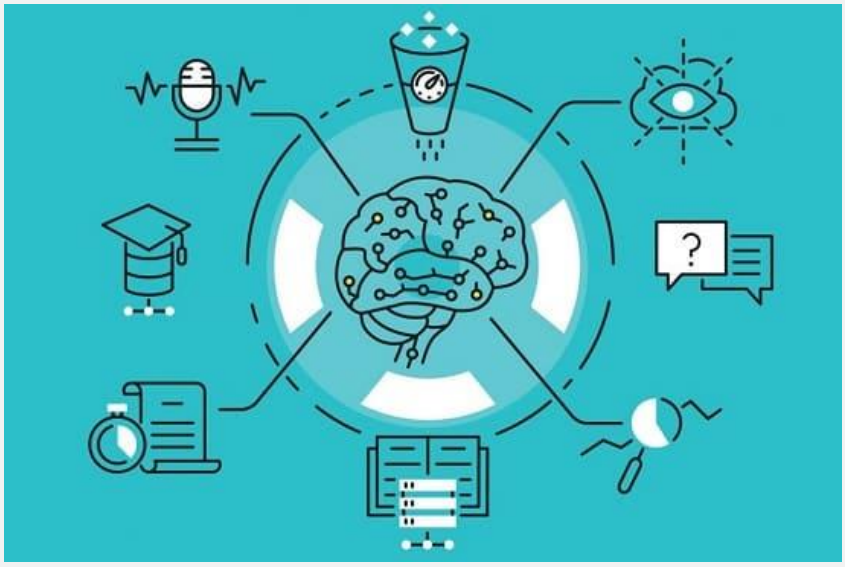
## تعریف مدل‌های پیچیده

اگرچه جنگل‌های تصادفی پیش‌بینی‌های قوی ارائه می‌دهند، تفسیر فرآیند تصمیم‌گیری مدل می‌تواند به دلیل ماهیت خوشه‌ای آن پیچیده باشد.
روش‌هایی مانند تحلیل اهمیت ویژگی، نمودارهای وابستگی جزئی و روش‌های تفسیرپذیری مدل - آگنوستیک برای بهبود تفسیر مدل استفاده می‌شوند.

## مدیریت استفاده از حافظه

از آنجایی که جنگل تصادفی درخت‌های تصمیم‌گیری متعددی را در طول آموزش می‌سازد، مدیریت استفاده از حافظه بسیار مهم می‌شود. پارامترهای تنظیم دقیق مانند تعداد درختان، عمق درخت و اندازه زیرمجموعه‌های ویژگی می‌تواند به ایجاد تعادل بین عملکرد مدل و کارایی حافظه کمک کند.

## روندهای آینده در جنگل تصادفی



با نگاهی به آینده، جنگل تصادفی و یادگیری ماشین در حال شکل‌گیری روندی بسیار جذاب است. تصور کنید که با جهش فناوری، پروسه‌ای وجود دارد تا جادویی که در داخل این مدل‌ها اتفاق می‌افتد قابل‌درک‌تر شود.

کارشناسان در حال کار بر روی نوعی هوش مصنوعی قابل توضیح (XAI) هستند که با هدف ابهام‌زدایی از پیچیدگی این مدل‌ها و با یادگیری عمیق، مانند کنار هم قرار دادن دو ابرقهرمان، با جنگی تصادفی همکاری می‌کنند.

همه چیز در مورد ترکیب قابلیت اطمینان جنگل تصادفی با قدرت محض شبکه‌های عصبی است. همچنین کارشناسان در حال بررسی یادگیری ماشین، که کار خود را در محاسبات لبه انجام می‌دهند و دستگاه‌های ما را در لحظه هوشمندتر می‌کنند، فعالیت می‌نمایند.

## سخن آخر

جنگل تصادفی یک الگوریتم عالی برای آموزش در مراحل اولیه توسعه مدل است تا ببینیم چگونه کار می‌کند. سادگی آن، ساختن یک جنگل تصادفی را

به یک پیشنهاد تبدیل می‌کند.این الگوریتم همچنین یک انتخاب عالی برای هر کسی است که نیاز به توسعه سریع یک مدل دارد که شاخص بسیار خوبی از اهمیتی که به ویژگی‌های شما می‌دهد، ارائه می‌کند. شکست جنگل‌های تصادفی از نظر عملکرد نیز بسیار سخت است. البته احتمالاً همیشه می‌توانید مدلی پیدا کنید که می‌تواند عملکرد بهتری داشته باشد؛ اما توسعه این مدل‌ها معمولاً زمان بیشتری می‌برد، اگرچه می‌توانند انواع مختلفی از ویژگی‌ها مانند باینری، دسته‌بندی و عددی را مدیریت کنند. به طور کلی، جنگل تصادفی یک ابزار عمدتاً سریع، ساده و قابل‌انعطاف است؛ اما بدون محدودیت نیست.