به نام خدا

آشنایی با مفاهیم پایه یادگیری ماشین با استفاده از داده‌های خواص عناصر (علم مواد)

اهداف آموزشی پروژه

- آشنایی با خواندن و پردازش دادهها در پایتون
- انجام مراحل پیشپردازش دادهها (نرمالسازی، حذف مقادیر گمشده و ...).
- تقسیم دیتاست به دادههای آموزشی و تست
- پیادهسازی سه الگوریتم:
  - Support Vector Machine (SVM)
  - Decision Tree
  - Random Forest
- ارزیابی و مقایسه عملکرد مدلها با استفاده از معیارهایی مانند دقت (Accuracy)
- نمایش نتایج به صورت نمودارهای مقایسهای

مراحل پروژه

---

1. انتخاب دیتاست (Elements Dataset):

دیتاست خواص عناصر یکی از مجموعه‌داده‌های پرکاربرد در حوزه علم داده و یادگیری ماشین است که برای آموزش مفاهیم پایه تحلیل داده مورد استفاده قرار می‌گیرد. این دیتاست شامل اطلاعات مربوط به عناصر جدول تناوبی و ویژگی‌های فیزیکی و شیمیایی آن‌ها می‌باشد.

1.2. ویژگیهای اصلی دیتاست:

دیتاست مورد استفاده شامل اطلاعات مربوط به عناصر جدول تناوبی است. مهم‌ترین ویژگی‌های این دیتاست عبارت‌اند از:

- **عدد اتمی(atomicNumber)**
- **نام عنصر(name)**
- **نماد شیمیایی(symbol)**
- **جرم اتمی(atomicMass)**
- **الکترونگاتیوی(electronegativity)**
- **شعاع اتمی(atomicRadius)**
- **شعاع یونی(ionRadius)**
- **شعاع واندروالسی(vanDelWaalsRadius)**
- **انرژی یونش(ionizationEnergy)**
- **میل الکترونی(electronAffinity)**
- **حالت‌های عدد اکسایش(oxidationStates)**
- **حالت فیزیکی استاندارد(standardState)**
- **نوع پیوند شیمیایی(bondingType)**
- **نقطه ذوب(meltingPoint)**
- **نقطه جوش(boilingPoint)**
- **چگالی(density)**
- **گروه شیمیایی(groupBlock)**
- **سال کشف(yearDiscovered)**

---

2. خواندن و بررسی اولیه داده ها:

- بارگذاری دیتاست با استفاده ازpandas
- بررسی شکل کلی دادهها (تعداد ردیفها و ستونها)
- نمایش چند ردیف ابتدایی برای آشنایی اولیه با ساختار
- بررسی نوع دادهها (Data Types) و اطلاعات کلی ستونها با استفاده از متد یا تابع info
- محاسبه آمار توصیفی ویژگیهای عددی (مثل میانگین، میانه، انحراف معیار) با استفاده از متد describe در پانداس
- بررسی تعداد مقادیر یکتا در هر ستون
- بررسی وجود مقادیر گمشده در هر ستون (Missing Value)

---

3. پیش پردازش داده ها:

پیش‌پردازش داده‌ها مرحله‌ای است که طی آن داده‌های خام برای استفاده در مدل‌های یادگیری ماشین آماده می‌شوند. در دیتاست خواص عناصر، این مرحله می‌تواند شامل حذف یا تکمیل مقادیر گمشده، تبدیل داده‌های متنی به عددی، حذف برخی ستون‌های غیرضروری و در صورت نیاز نرمال‌سازی ویژگی‌های عددی باشد.

برای مثال، مقادیر گمشده در برخی ستون‌ها می‌توانند با مقدار میانگین یا مد جایگزین شوند و ستون‌هایی مانند groupBlock یا standardState به صورت عددی تبدیل گردند. همچنین ستون‌هایی مانند nameو symbol که اطلاعات توصیفی دارند، می‌توانند از مجموعه داده حذف شوند.

به طور کلی، مرحله پیش‌پردازش شامل مجموعه‌ای از عملیات مختلف است و در این پروژه لازم نیست تمامی این مراحل انجام شود.

3.1. **حذف یا پر کردن دادههای گمشده (Missing Value) :**

مقادیر ناموجود در برخی ستون‌ها باید به صورت زیر مدیریت شوند:

- **پر کردن مقادیر خالی در ستون‌های عددی مانند atomicMass یا density با مقدار میانگین**
- **پر کردن مقادیر خالی در ستون‌های متنی مانند groupBlock یا standardState با مقدار مد**

3.2. **حذف ستونهای غیرضروری:**

- **نام عنصر : (name)** شامل نام عنصر است و معمولاً اطلاعات قابل استفاده و ساختاری برای مدل‌سازی فراهم نمی‌کند.
- **نماد شیمیایی : (symbol)** نماد عنصر بوده و صرفاً جنبه توصیفی دارد و برای استفاده در مدل‌های یادگیری ماشین مناسب نیست.
- **پیکربندی الکترونی : (electronicConfiguration)** به دلیل متنی بودن و پیچیدگی ساختار، قابلیت استفاده مستقیم در مدل را ندارد و حذف می‌شود.
- **حالت‌های عدد اکسایش : (oxidationStates)** شامل مقادیر متنی و چندحالته است و پردازش آن در این پروژه انجام نمی‌شود، بنابراین حذف می‌گردد.

---

4. تقسیم داده ها به دو بخش آموزش و تست:

در یادگیری ماشین، دادههای آموزش به دو بخش اصلی تقسیم میشوند :ترین (Train)و تست (Test). بخش ترین برای آموزش مدل استفاده میشود تا بتواند الگوها و روابط موجود در دادهها را یاد بگیرد. پس از آموزش، از دادههای تست استفاده میشود تا عملکرد مدل را ارزیابی کنیم؛ یعنی بررسی کنیم که مدل تا چه حد میتواند دادههایی را که قبلاً ندیده، به درستی پیشبینی کند. این تقسیم بندی از آن جهت اهمیت دارد که بتوان از بیش پردازش (Overfitting)جلوگیری کرد و اطمینان حاصل نمود که مدل در شرایط واقعی نیز عملکرد خوبی دارد .معمولا از 70% دادهها برای آموزش و 30% برای تست استفاده میکنند. شما نیز این تقسیم بندی استفاده کنید.

---

5. آموزش مدل ها با استفاده از سه الگوریتم یادگیری ماشین:

•اس.وی.ام(SVM): الگوریتم SVM یک روش ناظر برای طبقه بندی یا کلاسیفیکیشن است که هدف آن یافتن بهترین مرز تصمیمگیری بین کلاسهای مختلف داده ها است. این الگوریتم تلاش میکند فاصله بین مرز تصمیم و نزدیکترین نقاط از هر کلاس (که به آنها بردار پشتیبان گفته میشود) را بیشینه کند تا تفکیکپذیری مدل افزایش یابدSVM .به ویژه برای مسائل با داده های کم بعد و همچنین وقتی داده ها به صورت خطی یا نزدیک به خطی جداشدنی هستند، بسیار مؤثراست.

•درخت تصمیم (Decision Tree): این الگوریتم مدل تصمیمگیری را به صورت یک ساختار درختی نمایش میدهد که در آن هر گره یک ویژگی از داده را بررسی میکند، هر شاخه نشان دهنده یک نتیجه ممکن از آن بررسی است، و برگ ها نمایانگر خروجی نهایی یا کلاس پیشبینی شده هستند. درخت تصمیم

به دلیل ساده فهم بودن و توانایی در کار با دادههای عددی و طبقهای، یکی از محبوبترین روشها برای مدلسازی مسائل طبقهبندی (کلاسیفیکیشن)و رگرسیون است. فرق بین کلاسیفیکیشن و رگرسیون این است که دراولی خروجی به صورت گسسته و دومی به صورت پیوسته است.

•رندوم فارست (Random Forest): این الگوریتم ترکیبی از چندین درخت تصمیم است که به صورت تصادفی ساخته شده اند. هر درخت روی یک زیرمجموعه تصادفی از داده ها آموزش می بیند و پیشبینی نهایی مدل با رایگیری (در طبقهبندی) یا میانگین گیری (در رگرسیون) از خروجی تمام درختها به دست می آید.Random Forest نسبت به یک درخت تصمیم تکی پایدارتر است و دقت بالاتری دارد، چرا که از میانگینگیری چندین مدل برای کاهش نوسانات و جلوگیری از بیشبرازش استفاده میکند.

---

6. ارزیابی مدل ها:

نمایش نتایج به صورت نمودار در این مرحله دقت به دست آمده باید با استفاده از نمودار میله ای نمایش داده شود.

---

7. خروجی های مورد انتظار:

- گزارش خلاصه ای از مراحل کار
- کد کامل در یک فایل ipynb
- نمودار نهایی مقایسه دقت مدل ها (مثل نمودار میله ای)
- تحلیل اینکه کدام الگوریتم بهتر عمل کرده است و چرا به نظر شما بهتر عملکرد است.