دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

دانشکدهٔ ریاضی و علوم کامپیوتر

پروژهٔ دوم درس داده‌کاوی

پیش‌بینی بارش به کمک تحلیل داده‌های اقلیمی

استاد درس: دکتر فاطمه شاکری

طراحان پروژه: مریم صادقی، مهدی رجالی

زمستان ۱۴۰۲

معرفی مجموعه داده

مجموعه دادهٔ آورده‌شده حاوی اطلاعات آب و هوایی ۱۰ سال اخیر در یک کشور می‌باشد که به صورت روزانه و در مکان‌های مختلف جمع‌آوری شده است.
مجموعه داده شامل ۱۴۵۴۶۰ نمونه با ۲۲ ویژگی می‌باشد که از انواع اسمی و عددی هستند. توضیحی مختصر از ویژگی‌ها در زیر آورده شده است:

- Date: تاریخ ثبت مشاهدات
- Weather Station: کد مکان اندازه‌گیری و ثبت مشاهدات آب و هوایی
- Minimum/Maximum Temperature: حداقل یا حداکثر دمای ثبت‌شده در طی یک روز (بر حسب سلسیوس)
- Rainfall: میزان بارش در طی آن روز (بر حسب میلی‌متر)
- Evaporation: میزان تبخیر در آن روز (بر حسب میلی‌متر)
- Sunshine: تعداد ساعات پردرخشش آفتاب در آن روز
- Gust Trajectory: جهت قوی‌ترین باد در طی آن روز (بر حسب مقیاس 16 Compass Points)
- Air Velocity: سرعت قوی‌ترین باد در طی آن روز (بر حسب کیلومتر بر ساعت)
- Gust Trajectory at 9 AM/3 PM: جهت باد برای از ۱۰ دقیقه قبل از زمان مذکور (بر حسب مقیاس Compass Points)
- Air Velocity at 9 AM/3 PM: سرعت باد از ۱۰ قبل از زمان مذکور (بر حسب کیلومتر بر ساعت)
- Moisture Level at 9 AM/3 PM: میزان رطوبت هوا در زمان مذکور (بر حسب درصد)
- Atmospheric Pressure at 9 AM/3 PM: فشار هوا در زمان مذکور (بر حسب هکتوپاسکال)
- Cloudiness at 9 AM/3 PM: میزان مسدودیت آسمان توسط ابرها (یک‌هشتم یک‌هشتم)
- Recorded Temperature at 9 AM/3 PM: دما در زمان مذکور (بر حسب سلسیوس)
- Rain that day: وقوع یا عدم وقوع بارش در آن روز

هدف

از شما خواسته شده است تا بر اساس مجموعه دادهٔ فوق، یک مدل یادگیری ماشین برای پیش‌بینی بارش ارائه دهید که با گرفتن اطلاعات آب و هوایی یک روز مشخص، بتواند با دقت خوبی وضعیت بارش در روز آتی را مشخص کند.

چالش‌ها

مجموعه دادهٔ ارائه‌شده به نحوی انتخاب شده است که شما را با طیفی از مشکلات و چالش‌هایی که ممکن است در هنگام تحلیل انواع داده آشنا کند. در این مجموعه داده، به مسائل زیر ممکن است بر بخورید:

- **ناقص بودن تعداد قابل‌توجهی از داده‌ها:** برای سازگاری داده‌ها با مدل‌های مختلف، نیاز به تکمیل یا حذف داده‌های ناقص خواهید داشت.
- **تنوع ویژگی‌ها:** برای استفاده از اغلب مدل‌ها، نیاز است که داده‌های اسمی را به نحوی به داده‌های عددی تبدیل کنید. راه‌های متنوعی برای این کار از جمله Label Encoding و One-Hot Encoding وجود دارند که هر کدام مناسب نوع دادهٔ مشخصی هستند. برای کارکرد بهتر مدل می‌بایست روش مناسب تبدیل هر ویژگی را تشخیص داده و آنها را تبدیل کنید.
- **همبستگی ویژگی‌ها:** برخی از ویژگی‌ها ممکن است ارتباط و تشابه زیادی با یک‌دیگر داشته‌باشند و نیاز به نگه‌داری همگی آنها نباشد. برای یافتن این وابستگی‌ها و ارتباطات، از ابزار متنوعی مانند Correlation Heatmap استفاده کنید. یافتن ویژگی‌های بسیار مرتبط و حذف یا ترکیب آنها اغلب سبب بهبود کارایی مدل و سرعت آموزش آن می‌شود.
- **داده‌های پرت و نویز:** به دلیل تعداد بسیار زیاد نمونه‌ها و ماهیت این مجموعه داده، انتظار می‌رود قسمتی از نمونه‌ها، داده‌های پرت یا نویز باشند. برای کارکرد بهتر مدل، نیاز است تا این داده‌ها را شناسایی کرده و حذف کنید.
- **عدم توازن کلاس‌ها:** نسبت داده‌های دو کلاس تقریباً به صورت ۳ به ۱ است، که ممکن است باعث ایجاد بایاس [۱] در مدل شود و از دقت آن بکاهد. در مورد راهکارهای حل مشکلات مجموعه داده‌های غیرمتوازن [۲] تحقیق کنید و به دلخواه از یکی از آنها استفاده کنید و بهبود یا عدم بهبود کارایی را گزارش کنید.

جزئیات هر بخش

- شما در این پروژه موظف به پیاده‌سازی و مقایسهٔ عملکرد سه مدل SVM، KNN و Decision Tree هستید.
- شما می‌بایست پیش‌پردازش‌های مورد نیاز (مثلاً برای آماده‌سازی ورودی یک مدل) را با ذکر دلیل انجام دهید. مواردی همچون نرمال‌سازی [۳]، استانداردسازی [۴] و کاهش بعد [۵] برخی از پیش‌پردازش‌های احتمالی هستند که می‌توانند مفید باشند. دقت داشته باشید که پیش‌پردازش اشتباه می‌تواند باعث کمتر شدن کارایی مدل شود.
- برای پویش در داده‌ها، می‌بایست مجموعه داده‌های آموزشی را به کمک حداقل دو مدل نمودار بصری‌سازی کرده و به تحلیل آنها بپردازید.
- پس از آموزش مدل‌ها، برای ارزیابی آنها از مجموعه دادهٔ آزمایشی استفاده کرده و در کنار تشکیل ماتریس آشفتگی [۶]، با گزارش معیارهای Accuracy، Precision، Recall و F1 Score به تحلیل عملکرد مدل‌های مختلف بپردازید. در نظر داشته باشید که برای پیاده‌سازی معیارهای فوق می‌توانید از توابع موجود در کتابخانهٔ `sklearn` استفاده کنید.
- در صورت بهینه‌سازی مدل‌ها با انتخاب ابرپارامتر های [۷] بهتر برای هر مدل، دلیل یا نحوهٔ انتخاب این ابرپارامترها را توضیح دهید.

[1] Bias
[2] Imbalanced Dataset
[3] Normalization
[4] Standardization
[5] Dimensionality Reduction
[6] Confusion Matrix
[7] Hyperparameter

نحوهٔ ارزیابی

این پروژه دارای ۱۰۰ نمرهٔ اصلی و ۱۰ نمرهٔ امتیازی است، که تقسیم‌بندی ۱۰۰ نمرهٔ اصلی به شکل زیر می‌باشد:

بخش	نمره
بصری‌سازی	۱۰
پیش‌پردازش داده‌ها	۳۰
ساخت و آموزش مدل‌ها	۳۰
پیاده‌سازی معیارهای ارزیابی و مقایسهٔ عملکرد مدل‌ها	۳۰

نمرهٔ امتیازی به تحلیل‌های کامل‌تر و پیش‌پردازش‌های بهتر تعلق خواهد گرفت.

نحوه و مهلت ارسال

مهلت ارسال

شما برای ارسال این پروژه، تا پایان ۲۵ام اردیبهشت (یعنی ساعت ۲۳:۵۹ روز مذکور) فرصت دارید.

قالب ارسالی

فایل ارسالی شما می‌بایست به فرمت `zip`. با نام‌گذاری `project_2_<GroupID>.zip` باشد که در آن می‌بایست به جای `<GroupID>`، شمارهٔ گروه خود را قرار دهید. فایل فشردهٔ ارسالی می‌بایست شامل **هر سهٔ** مورد زیر باشد:

۱. یک فایل `pdf`. حاوی گزارش پروژه

۲. یک فایل `ipynb`. حاوی کد **کامنت‌گذاری‌شدهٔ** مرتبط با پروژه

۳. یک فایل `txt`. حاوی لینک گوگل کولب [8] منطبق با ژوپیتر نوتبوک [9] بخش قبل

[8] Google Colaboratory
[9] Jupyter Notebook