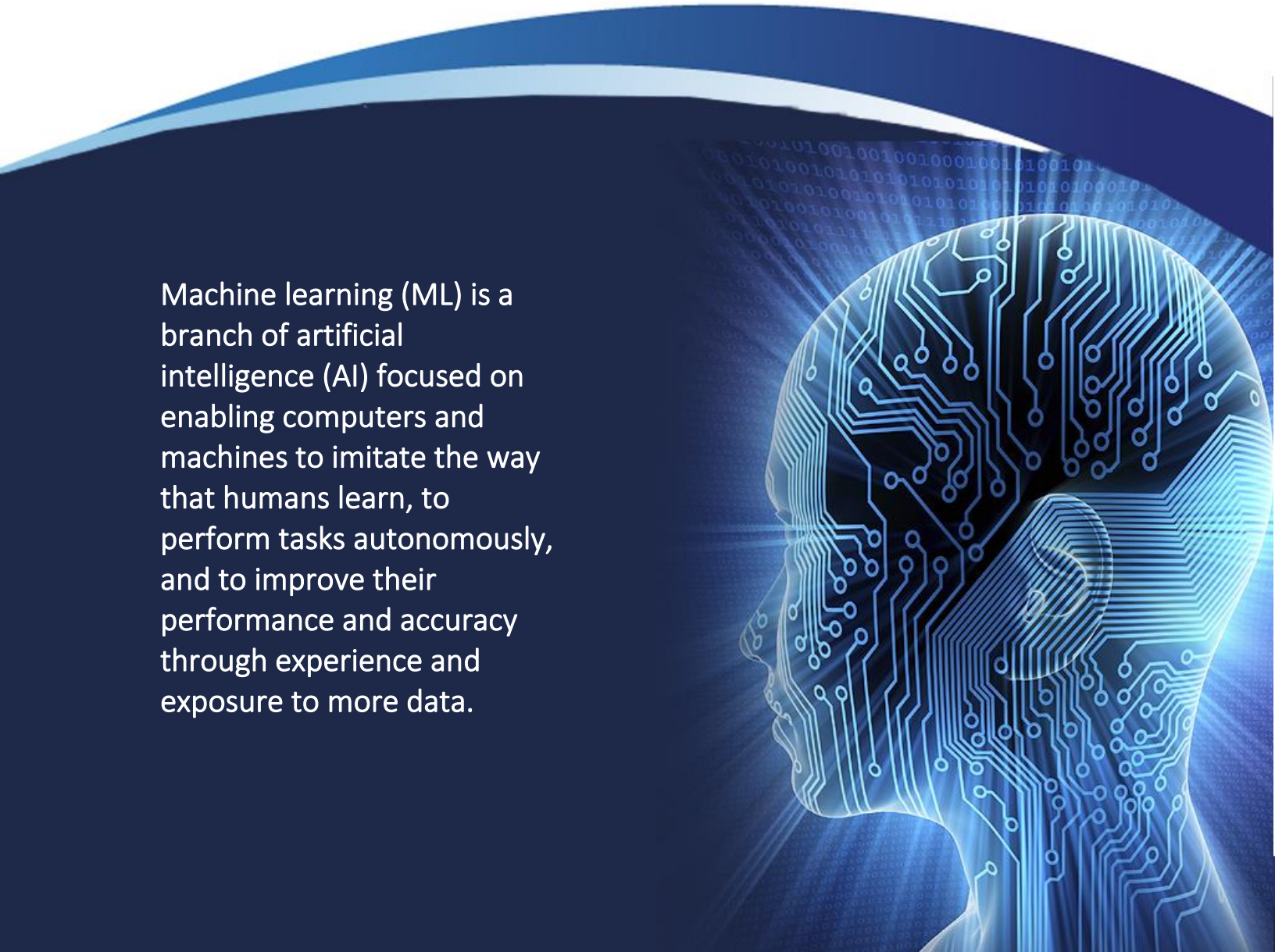
دانشگاه صنعتی اصفهان

ترم دوم سال تحصیلی ۱۴۰۳_۱۴۰۴

مبانی یادگیری ماشین

پروژه عملی

Machine learning (ML) is a branch of artificial intelligence (AI) focused on enabling computers and machines to imitate the way that humans learn, to perform tasks autonomously, and to improve their performance and accuracy through experience and exposure to more data.



نکات تکمیلی

۱. پاسخ ها در یکتا بارگذاری شوند و ارسال از روش های دیگر مورد قبول نیست.
۲. تحویل تکلیف با تاخیر تا ۷ روز امکان پذیر است و به ازای هر روز،۷ درصد از نمره آن تکلیف کسر میشود که به این نکته **حتما** توجه شود.
۳. ساختار نامگذاری تکلیف ارسالی باید به این صورت باشد.Project_ Programming_LastName_StudentID
studentID شماره دانشجویی وLastName نام خانوادگی شماست.
٤. انجام تکالیف به صورت تک نفره هست و در صورت مشاهده تقلب نمرات هم مبدا و هم مقصد آن صفر خواهد شد.
٥. براي انجام اين تكليف استفاده از زبان پايتون الزامي است و استفاده از توابعی جز pandas، numpy و matplotlib درصورتی که در سوال ذکر شود، مجاز نمیباشد.
٦. تکالیف را در محیط jupyter notebook یا google colab پیاده سازی کنید و فایل ipynb. را ارسال کنید.
۷. توضیح کدی که نوشته اید، بررسي و تحليل نتايج آن و بيان علت نتايج و نيز مقايسه نتيجه با آنچه مورد انتظارتان بوده است، از اهميت بالايي برخوردار است. شما مي توانيد گزارش پروژه را در همان محيطjupyter notebook بنويسيد و نيازي به فايل pdf جداگانه نيست. هم چنين اگر براي حل سوال فرضيات خاصي مدنظر داريد حتما آن را در متن گزارش قيد كنيد.
۸. در صورت هرگونه ابهامی می توانید سوالات خود را در گروه تلگرام بپرسید. و همچنین میتوانید با دستیاران آموزشی از طریق تلگرام در ارتباط باشید.

آیدی تلگرام:

@amirrezagholizade

تکلیف عملی دوم

سؤال اول: درخت تصمیم (۵۰ نمره)

مجموعه داده **تایتانیک** شامل اطلاعات مسافران از جمله سن، جنسیت، کلاس، قیمت بلیط و وضعیت زنده ماندن آن‌ها است. **هدف** این مجموعه داده پیش‌بینی این است که **آیا یک مسافر زنده می‌ماند یا خیر**.

برای استفاده از مجموعه داده تایتانیک به این لینک مراجعه کنید.

ویژگی‌های مجموعه داده:

Pclass - کلاس بلیط (عددی: ۱، ۲، ۳)

Sex - جنسیت مسافر (دسته‌بندی: مرد / زن)

Age - سن مسافر (عددی)

Fare - قیمت بلیط (عددی)

SibSp - تعداد خواهر، برادر یا همسر همراه مسافر (عددی)

Parch - تعداد والدین یا فرزندان همراه مسافر (عددی)

Embarked - بندر سوار شدن (دسته‌بندی: Q, C, S)

Survived (y) - آیا مسافر زنده ماند؟ (بله = ۱، خیر = ۰)

الف) پیش‌پردازش داده‌ها (۲۰)

۱. داده‌های موجود را از titanic.csv بخوانید. (۱)
۲. بررسی کنید که کدام ستون ها دارای مقادیر گمشده هستند و با انتخاب روش مناسب از نظر خودتان، این مقادیر را جایگزین کنید. در گزارشتان توضیح دهید که چرا این روش را انتخاب کردید و چرا فکر می‌کنید برای این ستون مناسب است. (۵)
۳. ویژگی‌هایی که غیر عددی هستند را به داده‌های عددی تبدیل کنید. (۵)
٤. داده‌ها را به دو بخش تقسیم کنید، ۸۰ درصد برای آموزش و ارزیابی و باقی آن‌ها برای تست در نظر بگیرید.(٤)
۵. با استفاده از روش StandardScaler داده ها را نرمال کنید. سپس داده های نرمال‌شده را بر روی نمودار نشان دهید. همچنین حداقل از ۳ نمودار برای نمایش داده ها و تفسیر آن‌ها به دلخواه انتخاب کنید و رسم کنید. (۵)

ب) آموزش با استفاده از درخت تصمیم (۲۰)

۱. با استفاده از کتابخانه scikit-learn، یک مدل درخت تصمیم را استفاده کنید و به کمک Grid Search مقادیر بهینه max_depth را بیابید. گزارش کنید کدام مقدار بهترین عملکرد را دارد. (۱۰)
۲. نمودار دقت برحسب عمق درخت را برای بررسی عملکرد درخت تصمیم رسم کنید. (۵)
۳. با توجه به بهترین عمقی که برای درخت تصمیم پیدا کردید نموداری رسم کنید که نشان دهد کدام ویژگی‌ها (Features) تأثیر بیشتری در تصمیم‌گیری دارند. کمی درباره این که نمودار چه چیزی را نشان می‌دهد توضیح دهید. (از feature_importances_ برای رسم نمودار استفاده کنید) (۵)

پ) بررسی مدل بر روی داده های تست (۱۰)

۱. عملکرد مدل را بر روی داده های آزمایشی ارزیابی کنید و معیار های **دقت (Accuracy)**، **یادآوری (Recall)**، **دقت پیش‌پینی (Precision)**، **F1-score** را برای مدل گزارش دهید. (۵)
۲. ساختار درخت را با استفاده از plot_tree() از sklearn.tree رسم کنید.(برای رسم از عمق ٤ استفاده کنید.) (۵)

تکلیف عملی دوم

سؤال دوم: خوشه‌بندی با KMeans(۵۰ نمره)

مجموعه داده Mall Customers شامل اطلاعات جمعیتی و رفتاری ۲۰۰ مشتری یک مرکز خرید است. این داده‌ها برای تحلیل رفتار خرید و بخش‌بندی مشتریان بسیار مناسب‌اند. در این تمرین، هدف شناسایی گروه‌های رفتاری مختلف بین مشتریان با استفاده از الگوریتم KMeans است.

برای استفاده از مجموعه داده به فایل داخل پروژه مراجعه کنید.

ویژگی‌های مجموعه داده:

Gender - جنسیت مشتری (دسته‌بندی: مرد / زن)

Age - سن مشتری به صورت یک مقدار عددی پیوسته ثبت شده است. (عددی)

Annual Income (k$) - میزان درآمد سالیانه مشتری، بر حسب هزار دلار. این ویژگی برای تحلیل وضعیت اقتصادی مشتری بسیار مفید است. (عددی)

Spending Score (1-100) - امتیاز مصرف مشتری که توسط مرکز خرید و بر اساس رفتار خرید او تعیین شده است؛ این امتیاز عددی بین ۱ تا ۱۰۰ است و نشان می‌دهد که مشتری تا چه اندازه اهل خرید و هزینه‌کردن است. (عددی)

الف) پیش‌پردازش داده‌ها (۲۰)

٦. داده‌های موجود را از Mall_Customers.csv بخوانید. (۱)
۷. ساختار و آماره‌های اولیه (توضیحات کلی، مقادیر گمشده، نوع داده‌ها) را بررسی و گزارش دهید. (۵)
۸. آماره‌های توصیفی مانند میانگین، انحراف معیار، کمینه و بیشینه را برای ویژگی‌های عددی نمایش دهید. (۳)
۹. تنها از ویژگی‌های Annual Income و Spending Score برای خوشه‌بندی استفاده کنید و این دو ویژگی را با استفاده از روش StandardScaler نرمال‌سازی کنید.(٦)
۱۰. یک نمودار پراکندگی (scatter plot) از این دو ویژگی بکشید تا توزیع مشتریان را ببینید. (۵)

ب) اجرای الگوریتم KMeans و تحلیل خوشه‌ها (۲۵)

٤. الگوریتم KMeans را برای تعداد خوشه‌های ۱ تا ۱۰ اجرا کنید و مقدار خطای درون‌خوشه‌ای (Inertia) را ذخیره کنید. نمودار Elbow را رسم کرده و عدد مناسب خوشه‌ها را انتخاب کنید. (۱۰)

٥. با استفاده از تعداد خوشه‌های انتخاب‌شده، KMeans را مجدداً اجرا کنید و برای هر مشتری خوشه مربوطه را تعیین کنید. (٥)

٦. سپس نمودار scatter دسته‌بندی شده با رنگ‌بندی خوشه‌ها رسم کنید. مراکز خوشه را هم در نمودار مشخص کنید. (٥)

٧. تعداد اعضای هر خوشه را گزارش دهید و ویژگی‌های عمومی هر خوشه را توصیف کنید (مثلاً: "مشتریانی با درآمد بالا و Spending Scoreپایین"). (٥)

پ) تحلیل خوشه‌بندی (٥)

٣. از معیار **Silhouette Score** برای ارزیابی کیفیت خوشه‌بندی استفاده کرده و مقدار آن را گزارش کنید. (٥)