پردیس دانشکده های فنی دانشگاه تهران

دانشکده برق و کامپیوتر

# هوش مصنوعی

# پروژه چهارم - یادگیری ماشین

مهلت تحویل: ۱۹ آذر

طراحان: مهسا اسکندری - آنیتا قندهاری - دیار محمدی

## هدف پروژه:

هدف این پروژه آشنایی با روشهای یادگیری ماشین[1] به کمک کتابخانه SciKit-Learn است. این پروژه چهار در فاز تعریف شده است. در فاز صفر، به بررسی مجموعه داده ها و تجزیه و تحلیل داده های اکتشافی[2] میپردازید. در فاز اول و پیش پردازش[3] را خواهید داشت. در فاز دوم، با استفاده از چند تخمین زن[4] تعریف شده در کتابخانه SciKit-Learn مدلهایی را پیاده سازی و بهینه سازی خواهیدکرد. نهایتا در فاز سوم، با استفاده از مدلهای بهینه فاز دوم، به پیاده سازی چند روش یادگیری گروهی[5] و تحلیل نتایج حاصل میپردازید.

- **در طول این متن با علامت [Number]مواجه می شوید که به معنای شماره ی لینک مربوط به آن موضوع می باشد. لینک ها در انتها آمده اند.**
- **در طی این پروژه نیاز است که اثر تصمیمات خود را در نتیجه‌ی نهایی بررسی کنید. لذا بهتر است به نحوی کد بزنید که اعمال تغییر در هر مرحله‌ی آن و بررسی نتیجه آسان باشد.**
- **در این پروژه تحلیل شما از تک تک مراحل کار، بیان مزایا و معایب هر روش و به طور کلی توانایی تشخیص استفاده از ابزار مناسب در هر مرحله از این پروژه از اهمیت بسیاری برخوردار است و شامل 80 درصد نمره پروژه خواهد بود. بنابراین در این پروژه کد شما تنها بخش محدودی از نمره را شامل می شود و بخش عمده نمره به توضیحات شما در حین ارائه و همچنین تحلیل های شما از هر بخش و ارائه نمودارهای مناسب در گزارش کار اختصاص دارد.**

---

[1] Machine Learning
[2] Exploratory Data Analysis
[3] Preprocessing
[4] Regressor
[5] Ensemble Learning

## معرفی مجموعه داده[6]:

مجموعه دادهاي که در اختیار شما قرار دارد House Prices است که یک مجموعه داده بارگذاری شده در سایت Kaggle است. این مجموعه داده شامل ویژگی[7] هاي خانه های مختلف شامل خیابان، محله، خدمات، تعداد اتاق ها، نوع ساختمان، سال ساخت و ... می باشد و داده هدف[8]، قیمت خانه است.

این مجموعه داده شامل ۴ فایل است که به ترتیب توضیح داده خواهند شد.

train.csv:

دارای ۸۳ ستون است که یکی از آن‌ها قیمت فروش هر خانه است و مابقی ویژگی‌های دیگر هر خانه‌ی فروخته شده است. این داده‌ها برای فرآیند یادگیری استفاده می‌شوند.

test.csv:

دارای ۸۲ ستون است و کاملا مشابه train.csv است با این تفاوت که ستون قیمت فروش خانه‌ها در آن وجود ندارد.

data_description.txt:

توضیحات کامل هر ستون مانند نام ستون، مفهوم آن و مقادیر آن است. توصیه میکنیم برای آشنایی با داده‌ها این فایل را حتما ملاحظه فرمایید.

sample_submission.csv:

یک نمونه فایل خروجی برای شرکت در مسابقه‌ی مربوط به مجموعه داده در سایت Kaggle است.

---

[6] Dataset
[7] Feature
[8] Target

## فاز صفر: Visualization and EDA

در این فاز داده های خام ورودي را بررسی خواهید کرد. در آمار، EDA رویکردی برای تجزیه و تحلیل مجموعه داده ها برای جمع بندی خصوصیات اصلی آنها است که اغلب با روش های Visualization انجام می شود. چه از مدل های آماری استفاده بشود چه نشود در درجه اول EDA برای دیدن آنچه داده ها می توانند فراتر از وظیفه مدلسازی به ما بگویند، انجام می شود.

در هر مرحله نتایجی را که بدست می آورید در گزارشتان ذکر کنید

1. ساختار کلی داده ها را با متدهای describe و info از پکیج Pandas بررسی کنید.
2. درصد داده‌های از دست رفته‌ی هر ویژگی [9] را بدست آورید.
3. نمودار وابستگی ویژگی ها به هم را رسم کنید (می توانید از برای این کار می توانید از Heatmap استفاده کنید.) به نظرتان کدام ویژگی‌ها در تخمین قیمت خانه‌ها می توانند مفید باشند؟ حداقل چهار ویژگی را نام ببرید.
4. از قیمت خانه‌ها لگاریتم گرفته و دوباره نمودار وابستگی ویژگی‌ها را رسم کنید. آیا تفاوت وجود دارد؟ چرا؟
5. آیا سوال ۳ برای بررسی تمام ویژگی‌ها کافی است؟ چه ایرادی وجود دارد؟
6. ارتباط ویژگی‌هایی که در سوال سه نام بردید را با قیمت خانه‌ها دقیق‌تر بررسی کنید. می‌توانید از نمودارهای Hexbin و Scatter استفاده کنید.
7. نگاهی به فایل data_description.txt انداخته و سه ویژگی دسته‌ای که به نظرتان در پیش‌بینی قیمت مفید هستند را انتخاب کنید سپس با نمودارهای مناسب آن‌ها را نمایش دهید.
8. شما می‌توانید هر بررسی دیگری به هر طریقی که می خواهید بر روی داده‌ها داشته باشید تا دید مناسب‌تری نسبت به ویژگی‌ها داشته باشید و در فازهای بعدی بهتر عمل کنید.

[9] Missing Values

## فاز اول: Preprocessing

پیش پردازش داده ها [1] مرحله مهمی در فرآیند داده کاوی است. روش های جمع آوری اطلاعات اغلب به راحتی کنترل نمی شوند، و در نتیجه مقادیر خارج از محدوده (به عنوان مثال، تعداد فرزندان: 100) ، ترکیب داده های "غیر ممکن" (به عنوان مثال، جنسیت: مرد ، باردار: بله)، مقادیر از دست رفته و غیره ممکن است در مجموعه داده‌ها وجود داشته باشند و می توانند نتایج گمراه کننده ای ایجاد کنند. بنابراین، قبل از هر کاری باید اینگونه داده ها را اصلاح کنیم. غالباً، پیش پردازش داده ها مهمترین مرحله پروژه یادگیری ماشین است، پس بهتر است کاملا روی این موضوع مسلط باشید.

برای هر کدام از موارد زیر ممکن است روش های مختلفی وجود داشته باشد، تفاوت روش ها را در گزارش توضیح داده و دلیل استفاده از روشی که انتخاب کردید را ذکر کنید.

1. دو روش برای برطرف کردن مشکل داده‌های گمشده، ۱- حذف کل ستون و ۲- پر کردن مقادیر گمشده با آماره‌ها(مثلا میانگین) هستند. دیگر روش‌های موجود برای هندل کردن داده‌های گمشده را در صورت وجود مختصرا توضیح دهید و این روش‌ها را مقایسه کنید.
2. با توجه به نتایج سوال دوم فاز صفر، کدام ویژگی‌ها بیشترین میزان داده‌ی گمشده را دارند؟ آیا باید ویژگی‌های با درصد بالای داده‌ی گمشده را از مجموعه داده‌ها حذف کرد؟ با روش‌های مناسب داده‌های گم‌شده را هندل کنید.
3. در ویژگی‌های عددی Normalizing یا Standardizing داده‌های عددی به چه منظور انجام می‌شود؟ به نظر شما آیا نیاز است که داده‌ها خود را در این پروژه نرمالیزه کنیم؟
4. برای این که مدل ما بتواند با داده های دسته ای کار کند چه روش هایی وجود دارد؟ آیا همه ی داده های دسته ای را باید با یک روش هندل کنیم؟ [2]
5. خوب است این جا فکر کنید که آیا باید همه ی ستون ها را نگه داریم؟ کدام ستون ها را می توانیم حذف کنیم؟ چرا؟(در پاسخ‌گویی به این پرسش سوال سوم از فاز صفر نیز می‌تواند مفید باشد.)
6. نیاز است که داده‌ها را به دو دسته‌ی train و test تقسیم کنیم. یک روش این است که P درصد اول داده‌ها را برای یادگیری و مابقی را برای ارزیابی اختصاص دهیم. روی مقدار مناسب P بحث کنید. آیا روش‌های دیگری وجود دارد؟ نام ببرید. آیا نیاز است تقسیم داده‌ها به صورت تصادفی باشد؟ چرا؟

## فاز دوم: Model Training, Evaluation and Hyperparameter Tuning

در این فاز از پروژه، سه مدل بر پایه K-Nearest-Neighbours و Decision Tree و Linear Regression با کمک کتابخانه SciKit-Learn پیاده‌سازی می‌کنید. سپس باید به بهینه‌سازی هایپرپارمترهای هر مدل بپردازید.تعداد همسایه‌ها برای الگوریتم KNN و عمق ماکزیمم برای الگوریتم Decision Tree کفایت میکند.

1. دقت هر مدل را بر اساس معیارهای MAE و RMSE روی داده های test و train اندازه‌گیری کنید.
2. در مدل‌های که نیاز به بهینه کردن هایپر پارامتر است نمودار تغییرات معیارهای بالا را به ازای مقادیر مختلف هایپر پارامتر رسم کنید.(در صورتی که تعداد هایپرپارامترها بالا برود برای سهولت کار می‌توانید از ابزار Grid Search در کتابخانه SciKit-Learn استفاده کنید.)
3. در مورد underfitting و overfitting تحقیق کنید و بررسی کنید که آیا در مدل‌های شما underfitting یا overfitting اتفاق افتاده است؟
4. تاثیر پیش‌پردازش‌هایی که روی داده‌ها انجام دادید را به طور کامل بررسی کنید.(مثلا تاثیر روش‌های مختلف هندل کردن مقادیر گمشده را روی معیارهای نهایی ببینید.)

## فاز سوم: Ensemble Methods

یادگیری گروهی به این معناست که از تجمیع نتایج حاصل از تعدادی مدل، پیش‌بینی نهایی را انجام دهیم. در این فاز به پیاده سازي و تحلیل چند روش یادگیری گروهی می‌پردازیم.

1. Random Forest روشی است که تعدادی Decision Tree با ویژگی‌های مختلف و دیتاهای مختلف را در کنار هم قرار داده و هر کدام جداگانه یادگیری را انجام می‌دهند و در نهایت میان آن‌ها به طریقی رای‌گیری می‌شود. با کمک کتابخانه SciKit-learn این مدل را پیاده‌سازی کنید. تاثیر حداقل دوتا از hyperparameter ها را بر مدل بررسی کرده و معیارهاي دقت را محاسبه کنید.
2. یکی دیگر از روشهاي یادگیری گروهی Voting Regression است. در این روش، پیش‌بینی نهایی، میانگینی از خروجی همه‌ی مدل‌ها است. با استفاده از سه مدل بهینه شده در فاز اول، یک مدل بر این پایه پیاده سازي کنید و معیارهاي دقت را اندازه‌گیری کنید.
3. در مورد Voting Regression علت شکست یا برتری آن‌را نسبت به روش‌های یادگیری فردی بررسی کنید.(برای مثال می توانید تفاوت جواب‌های هر کدام از سه مدل فاز قبل را نسبت به جواب‌های دیگر مدل‌ها بررسی کنید.)

## منابع:


:Data Preprocessing [1]

Towardsdatascience.com: Data Preprocessing Concepts

:Encoding Categorical Data [2]

Towardsdatascience.com: Categorical Encoding Techniques

:Hyperparameter Tuning [3]

Towardsdatascience.com: Hyperparameter Tuning

Towardsdatascience.com: Hyperparameters Optimization

:Overfitting [4]

Towardsdatascience.com: What Are Overfitting and Underfitting in Machine Learning

Towardsdatascience.com: Overfitting vs Underfitting a Complete Example

## نکات پایانی:

- مقدار مطلوب برای MSE و RMSE حداکثر به ترتیب $20000 و $30000 است. توجه داشته باشید که کسب این دقت، تنها توسط یکی از مدلهایی که در پروژه پیاده سازی کردید کافی است و سایر مدل ها باید صرفا دقت معقولی داشته باشند.
- همچنین به شدت توصیه میکنیم برای مقایسه روش‌های مورد استفاده و نتایج خودتان با دیگران، نوتبوک‌های موجود در رابطه با این مجموعه داده را ببینید و حتما در این مسابقه‌ی Kaggle شرکت کنید و نتایج به دست آمده برای فایل test.csv را در مسابقه آپلود کنید.(توجه داشته باشید که این مجموعه‌داده، به صورت کامل بر روی وب قرار دارد و تعدادی افراد با سوءاستفاده از این موضوع اقدام به شرکت در مسابقه و گرفتن امتیازهای عالی و میزان خطای نزدیک به صفر می‌کنند، نتایج خود را با چنین افرادی مقایسه نکنید. کسب خطای 0.20 و کمتر از آن در این مسابقه بسیار عالی است.)
- در تمامی بخشهاي پروژه، استفاده از کتابخانه SciKit-Learn مجاز است. دقت کنید که هدف پروژه تحلیل نتایج است بنابراین از ابزارهاي تحلیل داده بطور مثال نمودارها استفاده کنید و توضیحات مربوط به هر بخش از پروژه را بطور خلاصه و در عین حال مفید در گزارش خود ذکر کنید.
- **توجه کنید که باید از Jupyter Notebook استفاده کنید** و نتایج و گزارش خود را در یک فایل فشرده با عنوان **AI_CA4_<#SID>.zip** تحویل دهید. محتویات پوشه باید شامل تمامی پیاده سازي ها در یک فایل Notebook به همراه خروجی html آن ارائه دهید.
- در صورتی که سوالی در مورد پروژه داشتید بهتر است در فروم درس یا گروه تلگرام مطرح کنید تا بقیه از آن استفاده کنند؛ در غیر این صورت توسط ایمیل با طراحان در ارتباط باشید.
- هدف از تمرین، یادگیري شماست، مشورت ایرادی ندارد اما تمرین را باید خودتان انجام دهید.