# به نام خدا

# تکلیف چهارم درس مبانی داده کاوی

# ترم بهار ۱۴۰۰

**راهنمایی** :

زبان برنامه نویسی سئوالات پایتون است.

پیشنهاد می شود از محیط Jupyter notebook استفاده کنید.

پکیج‌های اصلی استفاده شده numpy, pandas, sklearn ,seaborn می باشند.

دیتاست های مورد نیاز در ادامه معرفی شده اند.

**روش تحویل**:

الف) فایلهای مربوط به کدهای هر سوال در یک فایل با نام Qx.zip که x شماره سوال است زیپ شوند (برای نمایش خروجی دستورات، هرجا مقدور است نام دیتافریم را بزنید تا خلاصه آن را نشان دهد و در سایر حالات از دستور head استفاده کنید)، سپس کلیه این فایلهای زیپ در یک فایل واحد با نام HW4-Lastname-StudentCode.zip که Lastname نام خانوادگی و StudentCode شماره دانشجویی شما است، زیپ شده و روی سامانه تا **زمان مشخص شده** آپلود شوند.

ب) گزارش نهایی باید شامل پاسخ تمامی سوالات (سوالات تحقیقی و سوالات پیاده سازی) باشد که برای سوالات پیاده سازی شامل کد نوشته‌شده، توضیحی درمورد کد و نتیجه اجرا و تفسیر نتیجه می‌باشد.

ج) زمان و نحوه تحویل تکلیف در فایل راهنمای ترم مشخص شده است.

د) تحویل خارج سامانه و خارج ساعت مشخص شده قابل قبول نیست.

**دیتاست شماره ۱:** مربوط به اطلاعات تعدادی خودرو است. با نام Vehicle در سئوالات به آن اشاره شده است.

**دیتاست شماره ۲:** مربوط به اطلاعات تعدادی بیماران دیابتی است. با نام Diabetes در سئوالات به آن اشاره شده است.

**نکته ۱** : برای پاسخ به سوالات **تحقیقی** و **تفسیری**، پس از مطالعه منابع موردنیاز فقط **برداشت خود** از مسئله را توضیح دهید.

**نکته ۲** : برای نمایش گرافیکی گراف های سوال پنج از دو کتابخانه pydotplus و graphviz استفاده نمایید. برای دانلود graphviz از لینک زیر استفاده نمایید: ( https://graphviz.gitlab.io/download/ )

**نکته ۳** : برای حل سوال ۳ از کلاس metrics از کتابخانه sklearn استفاده نمایید.

**نکته ۴** : سوال ۴ را روی کاغذ بصورت واضح و تمیز بنویسید و با CamScanner عکس بگیرید و در صورت نیاز فیلتر Magic را اجرا کنید تا واضح شود و در فایل پاسخ خود قرار دهید.

## ۱. دسته‌بندی با درخت تصمیم

a) فایل csv دیتاست Vehicle را خوانده و تبدیل به دیتافریم نمایید.

b) دیتافریم موجود را از نظر داده‌های گم شده (Missing Value)، نرمال‌سازی داده‌های عددی و همچنین encode کردن ستون‌های دسته‌ای بررسی کرده و کارهای لازم برای آماده سازی دیتافریم به منظور ایجاد مدل را انجام دهید.

c) همبستگی متغیرها را نسبت به یکدیگر با استفاده از نمودار heatmap رسم کنید. آیا میتوان مهم ترین ستون‌های موثر در فیلد Class را از روی این نمودار پیش بینی کرد؟

d) مقادیر همه ستون ها به جز ستون Class را در متغیر x قرار داده و ستون Class را در متغیر y قرار دهید.

e) مجموعه‌های آموزشی و تست را با نسبت ۰.۸ به ۰.۲ ایجاد کنید و توزیع دسته‌های ستون Class را در دو مجموعه نمایش دهید.

f) با استفاده از کلاس DecisionTreeClassifier مدل دسته‌بندی موردنظر خود را با روش C5.0 ایجاد نمایید. (راهنمایی: مقدار پارامترهای ورودی را به صورت زیر قرار دهید: criterion='entropy')

g) داده های تست را به مدل بدهید و میزان دقت مدل را نمایش دهید.

h) تحقیق کنید پارامترهای max_leaf_nodes و max_features در کلاس DecisionTreeClassifier به چه منظور استفاده می‌شوند.

i) خروجی تابع export_graphviz را بر روی مدل ایجاد شده نمایش دهید.

## ۲. جنگل تصادفی

a) فایل csv دیتاست Diabetes را به صورت دیتافریم بخوانید و اسامی را زیر به عنوان اسامی ستون ها جایگزین کنید.

1. Number of times pregnant.
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test.
3. Diastolic blood pressure (mm Hg).
4. Triceps skinfold thickness (mm).
5. 2-Hour serum insulin (mu U/ml).
6. Body mass index (weight in kg/(height in m)^2).
7. Diabetes pedigree function.
8. Age (years).
9. Class variable (0 or 1).

b) مقادیر دیتافریم موجود را از نظر داده‌های گم شده (Missing Value)، نرمالایز داده‌های عددی و همچنین اینکد کردن ستون های دسته‌ای بررسی کرده و کارهای لازم برای آماده سازی دیتافریم به منظور ایجاد مدل را انجام دهید.

c) مقادیر همه ستون ها به جز ستون Class را در متغیر x قرار داده و ستون Class را در متغیر y قرار دهید.

d) مجموعه‌های آموزشی و تست را با نسبت ۰.۸ به ۰.۲ ایجاد کنید و توزیع دسته‌های ستون Class را در دو مجموعه نمایش دهید.

e) با استفاده از کلاس RandomForestClassifier مدل پیش‌بینی‌کننده را ایجاد کنید. (راهنمایی : مقدار پارامترهای ورودی را به صورت زیر قرار دهید: max_depth = 3, criterion='entropy')

f) داده های تست را به مدل بدهید و میزان دقت مدل را نمایش دهید.

g) افزایش یا کاهش مقدار max_depth چه تاثیری روی دقت مدل خواهد داشت. بهترین مقدار برای عمق درخت این مسئله چه عددی میباشد.

## ۳. ارزیابی مدل

a) در این سوال از مدل ایجاد شده در سوال ۲ استفاده میکنیم بدین منظور تمامی مراحل سوال قبل را انجام داده و مدل موردنظر خود را با بهترین پارامترها ایجاد نمایید.

b) داده های تست را به مدل fit کنید و سپس تابع predict را برای آن فراخوانی کنید و نتیجه را در y_pred ذخیره کنید.

c) متد confusion_matrix را با داده های y_test و y_pred و مقدار برچسب ها مقداردهی کنید و نتیجه را تفسیر کنید. هر کدام از ۴ عدد نشان داده شده در خروجی نشان دهنده چیست؟

d) متد classification_report را با داده های y_test و y_pred مقداردهی کنید و نتیجه را تفسیر کنید. هر کدام از ستون‌های این گزارش نشان دهنده چیست؟

## ۴. مدل Naïve Bayes

a) با استفاده از رکوردهای جدول زیر و قانون بیز محاسبه کنید درصورتی که کسی دارای تب، عدم سرفه و دارای سردرد باشد، آیا آن فرد سرماخوردگی دارد یا خیر.

| شماره رکورد | سردرد | سرفه | تب | سرماخوردگی؟ |
|---|---|---|---|---|
| ۱ | دارد | دارد | دارد | آری |
| ۲ | دارد | دارد | دارد | خیر |
| ۳ | دارد | دارد | دارد | آری |
| ۴ | دارد | دارد | ندارد | خیر |
| ۵ | ندارد | دارد | ندارد | آری |
| ۶ | ندارد | ندارد | ندارد | خیر |
| ۷ | ندارد | ندارد | ندارد | آری |
| ۸ | دارد | ندارد | ندارد | خیر |
| ۹ | ندارد | ندارد | دارد | خیر |
| ۱۰ | ندارد | دارد | دارد | آری |