به نام پروردگار بخشنده و مهربان

دانشگاه کاشان
گروه مهندسی کامپیوتر

نیم‌سال اول ۱۳۹۷-۱۳۹۸

تکلیف سوم

پروژه درس با رپیدماینر

Foundations of Data Mining

مدرس: سید مهدی وحیدی پور
vahidipour@kashanu.ac.ir

توضیحات

هدف از این پروژه آشنایی عملی با برخی از امکانات ابزار داده‌کاوی Rapid miner است. خروجی این پروژه یک گزارش کامل است. به سوالات زیر پاسخ دهید. برای گرفتن نمره کامل باید برای هر سوال و به ازای هر مرحله، گزارش انجام کار را تهیه کنید: روند انجام تمرین در ابزار، تصاویر انجام مراحل مختلف، نحوه محاسبه مقادیر، تنظیمات و پارامترهای هر روش، هر فرضی که دانشجو برای انجام الگوریتم داشته است، هر توضیحی که برای فهم روش نیاز است، ...

متن ساده و روان گزارش نمره بالاتری خواهد داشت. دقت کنید که تمام تصاویر و جداول گزارش ارسالی باید شماره یا برچسب (Caption) داشته باشند؛ مانند جدول ۱ یا شکل ۳. از فونت‌های استاندار استفاده کنید. برای آشنایی با نحوه صحیح نگارش می‌توانید به فایل "ارائه چهارم: ارائه نوشتاری" در درس شیوه پژوهش وارائه اینجانب مراجعه کنید[۱].

سوالات

با استفاده از داده های بیماران قلبی مدلهای زیر را با پیش فرض های مشخص شده ایجاد کنید.

متغیر C متغیر هدف می باشد.

۱. تمام موارد ذیل را در یک فرآیند (process) ایجاد کنید:

a. نویزها را با استفاده از روش Detect Outlier Distances و بامعیار و شرط پیشنهادی خودتان و با فیلتر حذف کنید.

b. داده های از دست رفته(missing value) را باروش replace missing value و با مقدار max جایگزین نمایید.

c. نویزها را با استفاده از روش LOF (حد پایین : ۳ نزدیکترین همسایگی، حد بالا: ۷ نزدیکترین همسایگی)مشخص کنید و بامعیار و شرط پیشنهادی خودتان و با فیلتر حذف کنید.

d. نویزها را با استفاده از روش impute missing value و با دسته بند K-NN و با معیار سنجش ۱۰ نزدیکترین همسایگی پیش بینی نمایید.

e. با استفاده از خروجی مراحل a,b,c,d و با استفاده از روش split validation و با ۸۰ درصد داده آموزشی و روش K-NN بطوریکه هر نمونه با ۸ نزدیکترین همسایگی اش سنجیده شود، مدل را ایجاد کنید ومعیارهای کارایی accuracy, precision, recall را در خروجی نمایش دهید.

f. با استفاده از خروجی مراحل a,b,c,d و با استفاده از روش cross validation و با پارامترهای number of folds:8 و sampling type: stratified sampling و انتخاب دسته بند K-NN با معیار سنجش ۸ نزدیکترین همسایگی مدل جدیدی ایجاد کنید و ومعیارهای کارایی accuracy, precision, recall را در خروجی نمایش دهید.

g. با استفاده از خروجی مراحل a,b,c,d و با استفاده از روش split validation و و با ۸۰ درصد داده آموزشی و انتخاب دسته بند Decision tree با پارامترهای criterion: gini index و maximal depth :8 مدل جدیدی ایجاد کنید و ومعیارهای کارایی accuracy, precision, recall را در خروجی نمایش دهید.

h. با استفاده از خروجی مراحل a,b,c,d و با استفاده از روش cross validation و با پارامترهای number of folds:8 و sampling type: stratified sampling , و انتخاب دسته بند Decision tree با پارامترهای criterion: gini index و maximal depth :8 مدل جدیدی ایجاد کنید و ومعیارهای کارایی accuracy, precision, recall را در خروجی نمایش دهید.

[۱] https://faculty.kashanu.ac.ir/vahidipour/fa/page/14142/ شیوه ارایه و پژوهش

۳- برای هرکدام از حالتهای e و f بهترین پارامترها را به صورت مستقل بدست آوردید .(بهترین تعداد فولدها، بهترین میزان تقسیم داده آموزشی و آزمایشی، بهترین تعداد نزدیکترین همسایگی، بهترین نوع نمونه برداری)(مثلا در حالت e، اگر بخواهیم بالاترین میزان accuracy را داشته باشیم، اگر حالت a و b انتخاب شود، بهترین مقدار number of distances و number of outliers چقدر باشد، بهترین پارامتر جایگزینی مقادیر از دست رفته چه معیاری باشد، بهترین میزان تقسیم داده آموزشی و آزمایشی به چه میزان باشد و بهترین میزان k که نشان دهنده تعداد همسایگی ها می باشد، چقدر باشد).

۴- برای هرکدام از حالتهای g و h بهترین پارامترها را به صورت مستقل بدست آوردید .(بهترین نوع تقسیم بندی، بهترین عمق درخت، بهترین تعداد فولدها، بهترین میزان تقسیم داده آموزشی و آزمایشی، بهترین نوع نمونه برداری)

- **مهلت تحویل تمرین حداکثر تا ساعت ۲۴:۰۰ مورخه ۱۱ بهمن ۱۳۹۷ می‌باشد (غیر قابل تمدید). به ازای هر لحظه تاخیر پس از مهلت مقرر تا ۲۴ ساعت اول از مهلت مقرر ۳۰٪ از نمره، از ۲۴ ساعت تا ۴۸ ساعت تاخیر از مهلت مقرر ۶۰٪ از نمره و پس از ۷۲ ساعت از مهلت مقرر ۱۰۰٪ از نمره تمرین به عنوان جریمه بی‌نظمی کاسته می‌شود.**
- **گزارش کامل به همراه فایلهای رپیدماینتر را به صورت فشرده به آدرس vahidipour@chmail.ir بفرستید. اسم فایل فشرده DM-HW3-ID-Name باشد که ID شماره دانشجویی و Name نام خانوادگی شما است.**