بسمه تعالی

پروژه درس مبانی داده‌کاوی:	شماره ۲
تاریخ تحویل پروژه:	
عنوان پروژه:	پروژه پیش بینی کووید-۱۹
زبان برنامه نویسی:	پایتون (Python)، آر (R)، متلب (MATLAB) یا زبان انتخابی با اطلاع استاد درس

توضیحات پروژه:

در این پروژه دانشجویان باید بر روی فایل جدید مجموعه داده آمار بیماری کووید-۱۹ (آمار جهانی مبتلایان، مرگ و میرها، واکسیناسیون کووید-۱۹) **عملیات پیش پردازش داده‌ها را بر روی آن، همانند بخش اول پروژه پیش پردازش (پروژه قبلی) مجددا انجام دهند**، سپس برنامه ای طراحی و پیاده سازی کنند که موارد زیر را انجام دهد:

۱. ابتدا باید ردیف داده‌هایی را که در آن‌ها صفت‌های زیر مقدار تهی دارند (صفر را تهی در نظر نگیرید) را **برای هر کشور به صورت جداگانه** تبدیل به مقادیر صفر نمایید.

total_vaccinations	people_vaccinated	people_fully_vaccinated	new_cases	new_deaths

نکته: به عنوان مثال نمونه‌ی زیر غیر قابل قبول است و این ردیف داده‌ها باید در مجموعه داده به صفر تبدیل شوند.

نمونه‌ی غیر قابل قبول:

total_vaccinations	people_vaccinated	people_fully_vaccinated	new_cases	new_deaths
54000	54000			

نمونه‌های قابل قبول:

total_vaccinations	people_vaccinated	people_fully_vaccinated	new_cases	new_deaths
0	0	0	5	0
54000	54000	0	10	1

توجه ۱: دو کشور در جدول شماره ۲ را مطمئن شوید که از مجموعه داده حذف کرده اید.

Northern Cyprus	International

جدول شماره ۲. کشورهایی که باید از مجموعه حذف شده باشند.

نکته: همچنین، باید صفت‌های مردمی که به صورت کامل واکسینه‌شده اند (people_fully_vaccinated) و کل واکسیناسیون انجام یافته (total_vaccinations) **برای هر کشور به صورت جداگانه**، پس از اولین مشاهده مقدار عددی در این خانه‌ها، در صورت مشاهده خانه یا خانه‌های خالی بعد از آن‌ها، مقدار مشاهده شده آن‌ها تا مشاهده مقدارهای بعدی باید ثابت باشند. به عبارت دیگر نباید مقدار صفر به داده‌های بعد از اولین مقدار مشاهده شده داد، تا زمانی که مقدار جدیدی در این خانه یا خانه‌ها دیده شده و بروز شود. به عنوان مثال جدول شماره ۱ را می‌توان مشاهده کرد.

در جدول شماره ۱ تغییر مقدار انجام یافته توسط برنامه برای یک تغییر با رنگ قرمز، برای دو تغییر با رنگ زرد و برای سه تغییر با رنگ سبز مشخص شده است.

total_vaccinations	people_vaccinated	people_fully_vaccinated	new_cases	new_deaths
0	0	0	5	0
8200	8200	0	7	0
8200	0	0	19	1
54000	54000	0	10	1
120000	120000	0	94	0
240000	240000	0	98	4
240000	0	0	230	4
504502	448878	55624	340	12
504502	0	55624	315	3
547901	470341	77560	453	10

جدول شماره ۱. صفت هایی که باید برای هر کشور به صورت جداگانه عملیات پیش پردازش روی آن ها انجام یابد.

توجه ۲: برخی از دانشجویان به اشتباه موارد جدول شماره ۱ را در پروژه قبلی مرتب و پردازش کرده بودند، که این اشتباه است. **باید برای هر کشور به صورت جداگانه عملیات بر روی آن‌ها انجام یابد.**

۲. با استفاده از سه متد: جنگل تصادفی (Random Forest)، بیزساده (Naïve Bayes)، و درخت تصمیم (Decision Tree) که آموخته اید، باید برنامه ای در دو حالت طراحی و پیاده سازی کنید که موارد زیر را به ترتیب انجام دهید، و سپس در هر مرحله نتایج حاصل شده را در قالب نمودار های مرتبط خواسته شده و فایل با فرمت اکسل[1] (یا مقادیر جداشده با ویرگول[2]) ذخیره نمایید.

۲,۱. برای حالت اول، شما باید در بازه زمانی ثبت شده کووید-۱۹ **برای هر کشور به صورت جداگانه و هر قاره به صورت جداگانه**، از اولین تاریخ آن تا تاریخ ۲۰۲۱/۱۰/۳۱ را به عنوان مجموعه داده آموزش (Train Dataset)، و برای تاریخ ۲۰۲۱/۱۱/۰۱ تا ۲۰۲۱/۱۱/۳۰ برای هر کشور و هر قاره را به عنوان مجموعه داده آزمایش (Test Dataset) در نظر بگیرید.

۲,۱,۱. **برای هر کشور به صورت جداگانه**، برای مقادیر خواسته شده در جدول های شماره ۳، ۴، ۵، ۶، ۷ و ۸ را جداگانه با توجه به تاریخ پیش بینی کرده و در حالت های زیر انجام دهید، در پایان هر مرحله نتایج آن‌ها را در نمودار ها نمایش دهید. همچنین برای هر مرحله مقادیر خواسته شده در جدول شماره ۹ را محاسبه کرده و نمودار های مربوط به آن‌ها را نیز نمایش دهید. (**نکته: در هر نمودار و فایل باید مقادیر واقعی که در مجموعه داده آزمایش نیز وجود دارد را در بخش های خواسته شده، جهت مقایسه با نتایج بدست آمده نمایش داده و ذخیره نمایید.**)

۲,۱,۱,۱. برای هر کشور پس از یادگیری مدل‌های خود، برای ۷ روز پس از آخرین روز مورد استفاده قرار گرفته شده در مجموعه داده آموزش، مقادیر خواسته شده را پیش بینی کنید.

۲,۱,۱,۲. برای هر کشور پس از یادگیری مدل‌های خود، برای ۱۴ روز پس از آخرین روز مورد استفاده قرار گرفته شده در مجموعه داده آموزش، مقادیر خواسته شده را پیش بینی کنید.

۲,۱,۱,۳. برای هر کشور پس از یادگیری مدل‌های خود ، برای ۲۱ روز پس از آخرین روز مورد استفاده قرار گرفته شده در مجموعه داده آموزش، مقادیر خواسته شده را پیش بینی کنید.

۲,۱,۱,۴. برای هر کشور پس از یادگیری مدل‌های خود ، برای ۲۸ روز پس از آخرین روز مورد استفاده قرار گرفته شده در مجموعه داده آموزش، مقادیر خواسته شده را پیش بینی کنید.

۲,۱,۱,۵. برای هر کشور پس از یادگیری مدل‌های خود ، برای ۳۰ روز پس از آخرین روز مورد استفاده قرار گرفته شده در مجموعه داده آموزش، مقادیر خواسته شده را پیش بینی کنید.

[1] .xlsx
[2] .csv

۲,۱,۲. **برای هر قاره به صورت جداگانه**، برای مقادیر خواسته شده در جدول های شماره ۳، ۴، ۵، ۶، ۷ و ۸ را جداگانه با توجه به تاریخ پیش بینی کرده و در حالت های زیر انجام دهید، در پایان هر مرحله نتایج آن‌ها را در نمودار ها نمایش دهید. همچنین برای هر مرحله مقادیر خواسته شده در جدول شماره ۹ را محاسبه کرده و نمودار های مربوط به آن‌ها را نیز نمایش دهید. (**نکته: در هر نمودار و فایل باید مقادیر واقعی که در مجموعه داده آزمایش نیز وجود دارد را در بخش های خواسته شده، جهت مقایسه با نتایج بدست آمده نمایش داده و ذخیره نمایید.**)

۲,۱,۲,۱. برای هر قاره پس از یادگیری مدل‌های خود، برای ۷ روز پس از آخرین روز مورد استفاده قرار گرفته شده در مجموعه داده آموزش، مقادیر خواسته شده را پیش بینی کنید.

۲,۱,۲,۲. برای هر قاره پس از یادگیری مدل‌های خود، برای ۱۴ روز پس از آخرین روز مورد استفاده قرار گرفته شده در مجموعه داده آموزش، مقادیر خواسته شده را پیش بینی کنید.

۲,۱,۲,۳. برای هر قاره پس از یادگیری مدل‌های خود ، برای ۲۱ روز پس از آخرین روز مورد استفاده قرار گرفته شده در مجموعه داده آموزش، مقادیر خواسته شده را پیش بینی کنید.

۲,۱,۲,۴. برای هر قاره پس از یادگیری مدل‌های خود ، برای ۲۸ روز پس از آخرین روز مورد استفاده قرار گرفته شده در مجموعه داده آموزش، مقادیر خواسته شده را پیش بینی کنید.

۲,۱,۲,۵. برای هر قاره پس از یادگیری مدل‌های خود ، برای ۳۰ روز پس از آخرین روز مورد استفاده قرار گرفته شده در مجموعه داده آموزش، مقادیر خواسته شده را پیش بینی کنید.

۲,۲. برای حالت دوم، شما باید در بازه زمانی ثبت شده کووید-۱۹ **برای هر کشور به صورت جداگانه و هر قاره به صورت جداگانه**، از اولین تاریخ آن تا تاریخ آخرین تاریخ ثبت شده برای هر کشور و هر قاره را به عنوان مجموعه داده آموزش (Train Dataset) در نظر بگیرید.

۲,۲,۱. **برای هر کشور به صورت جداگانه**، برای مقادیر خواسته شده در جدول های شماره ۳، ۴، ۵، ۶، ۷ و ۸ را جداگانه با توجه به تاریخ پیش بینی کرده و در حالت های زیر انجام دهید، در پایان هر مرحله نتایج آن‌ها را در نمودار ها نمایش دهید. همچنین برای هر مرحله مقادیر خواسته شده در جدول شماره ۹ را محاسبه کرده و نمودار های مربوط به آن‌ها را نیز نمایش دهید. (**نکته: در هر نمودار و فایل باید مقادیر واقعی که در مجموعه داده آزمایش نیز وجود دارد را در بخش های خواسته شده، جهت مقایسه با نتایج آینده بدست آمده نمایش داده و ذخیره نمایید.**)

۲,۲,۱,۱. برای هر کشور پس از یادگیری مدل‌های خود، برای ۷ روز آینده پس از آخرین روز مورد استفاده قرار گرفته شده در مجموعه داده آموزش، مقادیر خواسته شده را پیش بینی کنید.

۲,۲,۱,۲. برای هر کشور پس از یادگیری مدل‌های خود، برای ۱۴ روز آینده پس از آخرین روز مورد استفاده قرار گرفته شده در مجموعه داده آموزش، مقادیر خواسته شده را پیش بینی کنید.

۲،۲،۱،۳. برای هر کشور پس از یادگیری مدل‌های خود، برای ۲۱ روز آینده پس از آخرین روز مورد استفاده قرار گرفته شده در مجموعه داده آموزش، مقادیر خواسته شده را پیش بینی کنید.

۲،۲،۱،۴. برای هر کشور پس از یادگیری مدل‌های خود، برای ۲۸ روز آینده پس از آخرین روز مورد استفاده قرار گرفته شده در مجموعه داده آموزش، مقادیر خواسته شده را پیش بینی کنید.

۲،۲،۱،۵. برای هر کشور پس از یادگیری مدل‌های خود، برای ۳۰ روز آینده پس از آخرین روز مورد استفاده قرار گرفته شده در مجموعه داده آموزش، مقادیر خواسته شده را پیش بینی کنید.

۲،۲،۲. **برای هر قاره به صورت جداگانه**، برای مقادیر خواسته شده در جدول های شماره ۳، ۴، ۵، ۶، ۷ و ۸ را جداگانه با توجه به تاریخ پیش بینی کرده و در حالت های زیر انجام دهید، در پایان هر مرحله نتایج آن‌ها را در نمودار ها نمایش دهید. همچنین برای هر مرحله مقادیر خواسته شده در جدول شماره ۹ را محاسبه کرده و نمودار های مربوط به آن‌ها را نیز نمایش دهید. (**نکته: در هر نمودار و فایل باید مقادیر واقعی که در مجموعه داده آزمایش نیز وجود دارد را در بخش های خواسته شده، جهت مقایسه با نتایج آینده بدست آمده نمایش داده و ذخیره نمایید.**)

۲،۲،۲،۱. برای هر قاره پس از یادگیری مدل‌های خود، برای ۷ روز آینده پس از آخرین روز مورد استفاده قرار گرفته شده در مجموعه داده آموزش، مقادیر خواسته شده را پیش بینی کنید.

۲،۲،۲،۲. برای هر قاره پس از یادگیری مدل‌های خود، برای ۱۴ روز آینده پس از آخرین روز مورد استفاده قرار گرفته شده در مجموعه داده آموزش، مقادیر خواسته شده را پیش بینی کنید.

۲،۲،۲،۳. برای هر قاره پس از یادگیری مدل‌های خود، برای ۲۱ روز آینده پس از آخرین روز مورد استفاده قرار گرفته شده در مجموعه داده آموزش، مقادیر خواسته شده را پیش بینی کنید.

۲،۲،۲،۴. برای هر قاره پس از یادگیری مدل‌های خود، برای ۲۸ روز آینده پس از آخرین روز مورد استفاده قرار گرفته شده در مجموعه داده آموزش، مقادیر خواسته شده را پیش بینی کنید.

۲،۲،۲،۵. برای هر قاره پس از یادگیری مدل‌های خود، برای ۳۰ روز آینده پس از آخرین روز مورد استفاده قرار گرفته شده در مجموعه داده آموزش، مقادیر خواسته شده را پیش بینی کنید.

Target
new_cases

جدول شماره ۳. صفتی که باید برای هر کشور (یا قاره) به صورت جداگانه مقادیر آن پس از آموزش مدل، پیش پینی شود.

Target
new_deaths

جدول شماره ۴. صفتی که باید برای هر کشور (یا قاره) به صورت جداگانه مقادیر آن پس از آموزش مدل، پیش پینی شود.

Target
total_cases

جدول شماره ۵. صفتی که باید برای هر کشور (یا قاره) به صورت جداگانه مقادیر آن پس از آموزش مدل، پیش پینی شود.

Target
total_deaths

جدول شماره ۶. صفتی که باید برای هر کشور (یا قاره) به صورت جداگانه مقادیر آن پس از آموزش مدل، پیش پینی شود.

Target
people_fully_vaccinated

جدول شماره ۷. صفتی که باید برای هر کشور (یا قاره) به صورت جداگانه مقادیر آن پس از آموزش مدل، پیش پینی شود.

Target
total_vaccinations

جدول شماره ۸. صفتی که باید برای هر کشور (یا قاره) به صورت جداگانه مقادیر آن پس از آموزش مدل، پیش پینی شود.

Evaluation Metrics					
Mean squared error (MSE)	Root Mean Square Error (RMSE)	R^2 (R-Squared)	Mean absolute error (MAE)	Relative Absolute Error (RAE)	Root Relative Squared Error (RRSE)

جدول شماره ۹. مقادیری که باید برای هر کشور (یا قاره) به صورت جداگانه مقادیر آن ها پس از آموزش مدل پیش بینی، محاسبه شوند.

۳. بهترین هایپر پارامتر ها (hyperparameter) را برای هر سه متد **برای تمامی تک به تک حالت‌ها** گفته شده با کمترین خطا بدست آورید، آن ها را نمایش داده و در یک فایل ذخیره نمایید. (**این قسمت از پروژه اختیاری است**، و دانشجویانی که بتوانند آن را انجام دهند برای آنان نمره اضافه در نظر گرفته خواهد شد.)

توجه: مجموعه داده پردازش شده به همراه تمامی مراحل را به صورت کامل باید در قالب یک فایل با فرمت اکسل[3] (یا مقادیر جداشده با ویرگول[4]) ذخیره نمایید. همچنین تمامی اشکال را باید ذخیره کنید. در زمان تحویل پروژه نیز، برنامه‌ای که نوشته‌اید را باید اجرا کرده تا کارهای ذکر شده را انجام دهد:

۱. پردازش هر مرحله را به ترتیب انجام داده و ذخیره کند. ۲. باید شکل‌ها را در هر مرحله به ترتیب ایجاد، نمایش و به همراه فایل های مربوطه ذخیره کند.

موفق باشید.

[3] .xlsx
[4] .csv