نرم افزار مورد استفاده:

این پروژه با استفاده از زبان پایتون و در محیط برنامه نویسی ژوپیتر انجام شده است.

توصیف Dataset و پیاده سازی:

داده این مسئله CausesOfDeath_France_2001-2008 مربوط به دلایل مرگ و میر افراد بین سال های ۲۰۰۱ تا ۲۰۰۸ در کشور فرانسه می باشد. این دیتاست دارای ۸ ویژگی است شامل ویژگی های "TIME","GEO","UNIT","SEX","AGE","ICD10","Value","Flag and Footnotes" که ۴ ویژگی ","GEO","UNIT","AGE"","Flag and Footnotes" بلا استفاده است زیرا برای همه نمونه ها مقادیر یکسانی دارند و هیچ فایده ای در امر پیشبینی برای ما ندارند که ابتدای کار این ویژگی های زائد را حذف کرده ایم. به صورت زیر

```
# delete columns
data = data.drop("GEO", axis=1)
data = data.drop("UNIT", axis=1)
data = data.drop("AGE", axis=1)
data = data.drop("Flag and Footnotes", axis=1)
```

ویژگی اول "TIME" است که تاریخ سال مربوط به نمونه در آن است. ویژگی بعدی مربوط به جنسیت مرد یا زن بودن است و ویژگی ICD10 نیز مربوط به دلیل مرگ و میر است.

	TIME	SEX	ICD10	Value
0	2001	Males	All causes of death (A00-Y89) excluding S00-T98	277858
1	2001	Males	Certain infectious and parasitic diseases (A00...	5347
2	2001	Males	Tuberculosis	545
3	2001	Males	Meningococcal infection	30
4	2001	Males	Viral hepatitis	471

در اینجا ویژگی نهایی value که تعداد مرگ و میر است را میخواهیم پیشبینی کنیم که با توجه به اینکه کمیت تعداد است و این برچسب در بازه بزرگی است و از صفر تا ۲۹۰۰۰۰ مقدار دارد، بنابراین با استفاده از رگرسیون این مقدار را باید پیشبینی کرد. بهترین و منطقی ترین کار برای این داده پیشبینی مقدار value است: به این صورت که بعنوان مثال داده تست می پرسد با توجه به داده های سال های پیش، در سال آینده چه تعداد مرد (یا زن) بر اثر بیماری سرطان معده فوت خواهند کرد. جهت استحضار تنها راه استفاده از کلاسیفیکیشن، پیشبینی ویژگی جنسیت است که منطقی و معقولانه نیست، به این خاطر که هدف داده تست در این حالت به این صورت است که: در سال آینده می دانیم ۲۲۰۰۰۰ نفر بر اثر بیماری سرطان معده فوت

خواهند کرد، این تعداد مربوط به مردان هست یا زنان. که سوال معقولانه ای نیست و این نکته را صرفا جهت استحضار شما آورده ایم که این روش بهترین و معقولانه ترین کار ممکن برای حل این مسئله است.

با توجه به داده ها ۶۶ دلیل مرگ از جمله بیماری های مختلف وجود دارند (ویژگی ICD10 ۶۶ مقدار یونیک دارد) شامل دلایل و بیماری های زیر:

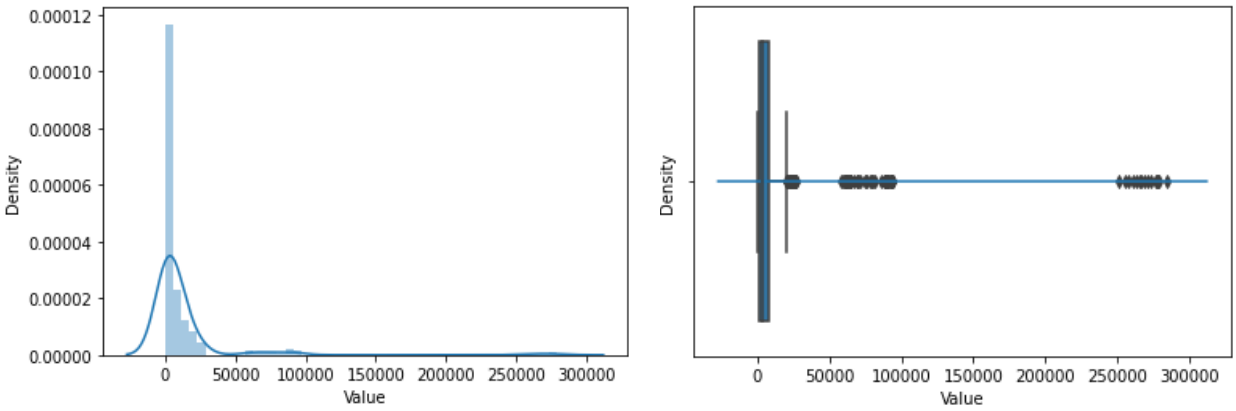
دلایل و بیماری های ذکر شده
' All causes of death (A00-Y89) excluding S00-T98,'
' Certain infectious and parasitic diseases (A00-B99,'(
' Tuberculosis', 'Meningococcal infection', 'Viral hepatitis,'
' Human immunodeficiency virus [HIV] disease', 'Neoplasms,'
' Malignant neoplasms (C00-C97,'(
' Malignant neoplasm of lip, oral cavity, pharynx,'
' Malignant neoplasm of oesophagus,'
' Malignant neoplasm of stomach', 'Malignant neoplasm of colon,'
' Malignant neoplasm of rectosigmoid junction, rectum, anus and anal canal,'
' Malignant neoplasm of liver and intrahepatic bile ducts,'
' Malignant neoplasm of pancreas,'
' Malignant neoplasm of larynx, trachea, bronchus and lung,'
' Malignant melanoma of skin', 'Malignant neoplasm of breast,'
' Malignant neoplasm of prostate,'
' Malignant neoplasm of kidney, except renal pelvis,'
' Malignant neoplasm of bladder,'
' Malignant neoplasms, stated or presumed to be primary, of lymphoid,'
' Diseases of the blood and blood-forming organs and certain disorders involving,'
' Endocrine, nutritional and metabolic diseases (E00-E90,'(
' Diabetes mellitus', 'Mental and behavioural disorders (F00-F99,'(
' Mental and behavioural disorders due to use of alcohol,'
' Drug dependence, toxicomania (F11-F16, F18-F19,'(
' Diseases of the nervous system and the sense organs (G00-H95,'(
' Meningitis', 'Diseases of the circulatory system (I00-I99,'(
' Ischaemic heart diseases,'
' Other heart diseases (I30-I33, I39-I52,'(
' Cerebrovascular diseases,'
' Diseases of the respiratory system (J00-J99)', 'Influenza,'
' Pneumonia', 'Chronic lower respiratory diseases,'
' Asthma and status asthmaticus,'
' Diseases of the digestive system (K00-K93,'(

```
'    Ulcer of stomach, duodenum and jejunum', 'Chronic liver disease,'
'    Diseases of the skin and subcutaneous tissue (L00-L99,'(
'    Diseases of the musculoskeletal system and connective tissue (M00-M99,'(
'    Rheumatoid arthritis and arthrosis (M05-M06,M15-M19,'(
'    Diseases of the genitourinary system  (N00-N99,'(
'    Diseases of kidney and ureter,'
'    Certain conditions originating in the perinatal period (P00-P96,'(
'    Congenital malformations, deformations and chromosomal abnormalities (Q00-Q99,'(
'    Congenital malformations of the nervous system,'
'    Congenital malformations of the circulatory system,'
'    Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere'
'    Sudden infant death syndrome,'
'    Ill-defined and unknown causes of mortality,'
'    External causes of morbidity and mortality (V01-Y89,'(
'    Accidents', 'Transport accidents (V01-V99)', 'Falls,'
'    Accidental poisoning by and exposure to noxious substances,'
'    Intentional self-harm', 'Assault', 'Event of undetermined intent,'
'    Malignant neoplasm of cervix uteri,'
'    Malignant neoplasm of other parts of uterus,'
'    Malignant neoplasm of ovary,'
'    Pregnancy, childbirth and the puerperium (O00-O99'(
```

که موجب مرگ افراد شده اند. ابتدا این ویژگی categorical را به numerical تبدیل کردیم و با توجه به تعداد زیاد حالت های این ویژگی و کم بودن تعداد ویژگی ها (سه ویژگی) نتیجه نهایی مناسبی بدست نمی آمد. در گام بعدی این ویژگی را به ۶۶ ویژگی جدید به صورت صفر و یک تبدیل کردیم که نشان دهنده بودن یا نبودن آن دلیل برای مرگ افراد است و در این حالت نتیجه نهایی بهبود بسیار قابل ملاحظه ای داشت و از این روش بهره بردیم. به صورت زیر:

```
# create features based on feature of ICD10 values
data = pd.get_dummies(data=data, drop_first=True)
value = data.Value
data = data.drop("Value", axis=1)
data["Value"] = value
```

در ویژگی value بعضی از مقادیر miss شده بودند که ابتدا سعی در پر کردن این مقادیر با روش های بهینه یادگیری ماشین داشتیم. با توجه به distribution مقادیر داده ها که غیرتقارنی بودند و بزرگ بودن برخی مقادیر تاثیر زیادی در mean می گذاشت پس از median داده ها برای پر کردن این miss value ها استفاده کردیم. (distribution مقادیر داده ها به شکل زیر است)



سپس به این نکته توجه کردیم که بعنوان مثال برای پر کردن تعداد افراد بر اثر بیماری سرطان پستان median تعداد مرگ و میر افراد بر اثر این سرطان که دارای تعداد ۲۰۰۰۰۰ نفر است با مرگ و میر بر اثر تصادف و یا افتادن بسیار می تواند متفاوت باشد. پس برای هر بیماری خاص median همان بیماری یا دلیل را آمدیم جای گذاری کنیم. ولی این مقدار صفر شد. که در اینجا متوجه شدیم که این miss value ها درست هستند و مقدار صفر دارند زیرا به عنوان مثال تعداد مرگ و میر مردان بر اثر سرطان پستان صفر است و برای زنان نیز همچنین مقادیری صفر است و کلا همچنین ردیف هایی نیازی نیست زیرا اساسا این سوال غلط است که تعداد مردان فوت شده بر اثر بیماری سرطان پستان چند نفر است و این ردیف ها را برای بهبود کار حذف نمودیم. به صورت زیر:

```
# Preprocessing
# replace ":" characters in Value column with 0 and remove these rows in 4 next line
data.Value[data.Value == ":"] = "0"
# remove space characters in Value column and convert str to integer
data["Value"] = data["Value"].apply(lambda x: int(x.replace(" ", "")))
data = data.loc[data["Value"]!=0]
```

نمایش تعدادی از ویژگی ها و نمونه ها: ویژگی های بدست آمده نهایی به شکل زیر است.

	TIME	SEX_Males	ICD10_Accidents	ICD10_All causes of death (A00-Y89) excluding S00-T98	ICD10_Assault	ICD10_Asthma and status asthmaticus	ICD10_Cerebrovascular diseases	ICD10_Certain conditions originating in the perinatal period (P00-P96)	ICD10_Certain infectious and parasitic diseases (A00-B99)	ICD10_Chronic liver disease	...	ICD
0	2001	1	0	1	0	0	0	0	0	0	...	
1	2001	1	0	0	0	0	0	0	1	0	...	
2	2001	1	0	0	0	0	0	0	0	0	...	
3	2001	1	0	0	0	0	0	0	0	0	...	
4	2001	1	0	0	0	0	0	0	0	0	...	

5 rows × 68 columns

در اینجا با استفاده از داده های سال ۲۰۰۱ تا ۲۰۰۷ میخواهیم پیشبینی کنیم که در سال ۲۰۰۸ چه تعداد افراد به ازای بیماری ها و دلایل مختلف فوت کرده اند. پس داده های سال ۲۰۰۸ را بعنوان داده آزمایش میگیریم (یک هشتم داده ها را داده آزمایش گرفتیم) و بقیه داده ها بعنوان داده آموزش. به صورت زیر:

```
# train and test splits
x_train = data.iloc[0:889,0:-1] # 2001 to 2007
y_train = data.iloc[0:889,-1]
x_test = data.iloc[889::, 0:-1]# 2008
y_test = data.iloc[889::, -1]
```

با استفاده از روش های رگرسیون خطی و همچنین رگرسیون درخت تصمیم مدل را ساخته شد و برروی داده های آزمایش اعمال شد. به صورت زیر:

```
# create linear regresion model and predict test labels
Model = LinearRegression()
Model.fit(x_train, y_train)
y_pred = Model.predict(x_test)
rmse = np.sqrt(mean_squared_error(y_test,y_pred))
print("Normalized RMSE = " ,rmse/(data.Value.max()-data.Value.min()))
Model = DecisionTreeRegressor(max_depth=5)
Model.fit(x_train, y_train)
y_pred = Model.predict(x_test)
rmse = np.sqrt(mean_squared_error(y_test,y_pred))
print("Normalized RMSE2 = " ,rmse/(data.Value.max()-data.Value.min()))
```

در نهایت با استفاده از خطای جذر میانگین مربعات rmse خطای این پیشبینی را اندازه گرفتیم و باتوجه به زیاد بودن مقادیر value که از صفر تا ۲۹۰۰۰۰ هستند این خطا را بصورت نرمالایز شده محاسبه کردیم که مقدار ۰,۰۱۱ بدست آمد. این خطا با استفاده از فرمول زیر بدست می آید.

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \hat{x}_i)^2}{N}}$$

RMSD = root-mean-square deviation
i = variable i
N = number of non-missing data points
x_i = actual observations time series
$\hat{x}_i$ = estimated time series

نتایج بدست آمده را در جدول زیر نمایش داده ایم که در روش Linear Regression خطای کمتری مشاهده شده است.

نام روش	Normalized RMSE
Linear Regression	۰,۰۱۱۱۱
DecisionTreeRegressor	۰,۰۱۹۹

در ادامه با استفاده از روش های کلاسیفیکیشن روش های KNeighborsClassifier و DecisionTreeClassifier و RandomForestClassifie را برای این داده با درنظر گرفتن ویژگی جنسیت بعنوان برچسب نهایی اعمال کرده ایم که نتایج را در جدول زیر می توان مشاهده کرد. همانطور که میبینیم در روش RandomForestClassifier دقت بهتری بدست آمده است.

نام روش	Accuracy
KNeighborsClassifier	۵۹,۸۴
DecisionTreeClassifier	۸۱,۸۸
RandomForestClassifier	۹۰,۵۵