در ابتدا باید مشخصات و ویژگی های دیتاست را بررسی کنیم مثلا معنی ستون ها هرکدوم چیه؟

یک بخش توضیح به جدول اضافه کنیم و شاخص ها را توضیح بدهیم.

آماده سازی دیتا : دیتا را تمیز کنیم و مثلا میسینگ ولیو ها را پیدا کنیم و برای آنها یک فکر بکنیم و استراتژی جهت حل مشکل میسینگ ولیو را بیان کنیم و از آن استراتژی استفاده کنیم.

گاهی وقت ها بهترین روش این است که همه سطر هایی که میسینگ ولیو هستند را نادیده بگیریم مثلا وقتی که هزینه و وقت پیدا کردن میسینگ ولیو ها خیلی زیاد است و یا تاثیر آن ستونی که میسینگ ولیو دارد در کل دیتاست خیلی کم است.

مسئله داده های پرت: باید ببینیم در هر ستون چقدر داده های پرت وجود دارد و میزان پراکندگی ستون ها چقدر است. بهترین روش استفاده از نمودار جعبه ای است تا ببینیم چه مقدار داده پرت داریم و با آن داده های پرت چه کاری بکنیم؟ حذف کنیم یا نگهداریم و چرا؟

پیش بینی ← امید به زندگی را پیش بینی کنیم (life expectancy)

آیا می شود مرگ نوزادان را پیش بینی کرد؟

مسئله پیش بینی امید به زندگی Classification یا regression؟ رگرشن. چون داده های امید به زندگی بصورت عددی و پیوسته هستند.

باید دیتا ست را به دو بخش Train و test تقسیم کنیم و روی Train مدل را بسازیم و روی Test آزمایش کنیم. معمولا باید ۷۰ درصد دیتا را برای آموزش گذاشته و ۳۰ درصد را برای تست بگذاریم.

کتابخانه سایکیت لرن ?traintestsved این کار را انجام می دهد فقط باید درصد بهش بدهیم. سپس accuracy مدل را می سنجیم با همین کتابخانه سایکیت لرن.

مصور سازی داده ها : دو یا سه تا از این بیماری ها را انتخاب کنیم و تاثیر و ارتباط این بیماری ها با امید به زندگی را مشخص کنیم و مصور کنیم.

در قسمت بعد نسبت این بیماری ها با امید به زندگی در کشور های توسعه یافته و در کشور های در حال توسعه چطوری است (در یک نمودار)

در مصور سازی باید دو یا سه عامل را انتخاب کنیم و تاثیرش را در یک عامل مشخص کنیم با در نظر گرفتن کشور توسعه یافته یا در حال توسعه.

نمودار hitmap را هم حتما بکشیم تا کورلیشن را بتوانیم روی ان به راحتی نمایش دهیم.

میانگین مدت زمانی که فرد در مدرسه می ماند در مقایسه با امید به زندگی با در نظر گیری کشور های درحال توسعه یا توسعه یافته در یک نمودار.