Predicting the severity of an accident

【Introduction】

In this project, we will try to use other variables to determine the level of severity in an accident.

In these sections, we will first introduce the data, like where it is from, the meaning of each variables.

Then we will use the distribution plot to discuss the distribution of some important variables.

Finally, we will apply three machine learning methods and compare their accuracy.

The problem

The data of this project is obtained from the given data-set of the course,which includes all the variables of the accidents occurred in Seattle in one year recorded by SPD. The aim of this project is to predict those variables that will influence the level of severity in the accident and build a machine learning model to record this.

Why it is interesting

Drivers may be interested in this topic since they want to know the dangerous driving variables so that they can avoid. Urban planners may be interested in this topic since it depicts the detailed information of the road condition that will cause a severe accident so that they can adjust. While the police officers may can use the trained machine learning model to define the severity of an accident in the future.

Data source

This data set is obtained from the link provided by this course, which is published by SPD, recording all variables related to every accident occurred in Seattle in one year. You can download the data in 'https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv'.

SEVERITY	X	Y	OBJECTID	INCKEY	COLDETK	REPORTNO	STATUS	ADDRTYPE	INTKEY	LOCATION	EXCEPTRSN	EXCEPTRSN	SEVERITY	SEVERITY	COLLISIC	PERSONC	PEDCOU	PEDCYLC	VEHCOU	INCDA
2	-122.3231	47.70314	1	1307	1307	3502005	Matched	Intersectio	37475	5TH AVE			2	Injury Coll	Angles	2	0	0	2	2013/0
1	-122.3473	47.647172	2	52200	52200	2607959	Matched	Block		AURORA BR BETWEEN RAYE			1	Property D	Sideswipe	2	0	0	2	2006/1
1	-122.3345	47.607871	3	26700	26700	1482393	Matched	Block		4TH AVE BETWEEN SENECA			1	Property D	Parked Ca	4	0	0	3	2004/1
1	-122.3348	47.604803	4	1144	1144	3503937	Matched	Block		2ND AVE			1	Property D	Other	3	0	0	3	2013/0
2	-122.3064	47.545739	5	17700	17700	1807429	Matched	Intersectio	34387	SWIFT AVE S AND SWIFT AV			2	Injury Coll	Angles	2	0	0	2	2004/0
1	-122.3876	47.690575	6	320840	322340	E919477	Matched	Intersectio	36974	24TH AVE			1	Property D	Angles	2	0	0	2	2019/0
1	-122.3385	47.618534	7	83300	83300	3282542	Matched	Intersectio	29510	DENNY WAY AND WESTLAK			1	Property D	Angles	2	0	0	2	2008/1
2	-122.3208	47.614076	9	330897	332397	EA30304	Matched	Intersectio	29745	BROADW			2	Injury Coll	Cycles	3	0	1	1	2020/0
1	-122.3359	47.611904	10	63400	63400	2071243	Matched	Block		PINE ST BETWEEN 5TH AVE			1	Property D	Parked Ca	2	0	0	2	2006/0
2	-122.3847	47.528475	12	58600	58600	2072105	Matched	Intersectio	34679	41ST AVE SW AND SW THISTI			2	Injury Coll	Angles	2	0	0	2	2006/0
1			14	48900	48900	2024040	Matched	Alley					1	Property D	Other	2	0	0	2	2006/0
1	-122.3338	47.547371	15	38800	38800	C654800	Matched	Intersectio	33194	1ST AV S BR NB AND EAST M			1	Property D	Angles	2	0	0	2	2005/0
1	-122.3563	47.571375	16	2771	2771	1211870	Unmatche	Block		SW SPOK			1	Property D	Rear Ende	0	0	0	2	2006/0
1	-122.324	47.606374	17	32800	32800	2128498	Matched	Block		TERRY AVE BETWEEN JAME			1	Property D	Parked Ca	2	0	0	2	2005/1
2	-122.3174	47.664028	19	1212	1212	3507861	Matched	Block		ROOSEVE			2	Injury Coll	Head On	2	0	0	2	2013/0
1	-122.3377	47.61751	20	330878	332378	3838086	Unmatche	Block		9TH AVE			1	Property	Damage On	1	0	0	0	2020/0
2	-122.3445	47.692012	21	46300	46300	2023080	Matched	Intersectio	37365	AURORA AVE N AND N 87TH			2	Injury Coll	Left Turn	3	0	0	2	2005/0
1			23	23000	23000	537838	Matched	Block		BATTERY ST TUN ON RP BET			1	Property D	Rear Ende	0	0	0	2	2004/0
2	-122.3283	47.57142	24	330833	332333	EA29752	Matched	Block		S SPOKAN			2	Injury Coll	Rear Ende	4	0	0	3	2020/0
1	-122.3838	47.583715	25	97100	97100	2894590	Matched	Block		41ST AVE SW BETWEEN SW W			1	Property D	Parked Ca	2	0	0	2	2009/0
2	-122.2924	47.732847	26	1347	1347	3608880	Matched	Block		LAKE CIT			2	Injury Coll	Rear Ende	3	0	0	2	2013/0
2	-122.3138	47.708535	28	1323	1323	3502831	Matched	Intersectio	36505	14TH AVE			2	Injury Coll	Angles	5	0	0	3	2013/0
1	-122.3182	47.615837	29	80000	80000	2882620	Matched	Block		11TH AVE BETWEEN E PINE			1	Property D	Parked Ca	2	0	0	2	2008/0
1	-122.3375	47.589746	31	28700	28700	1213894	Matched	Block		ALASKAN WY VI SB BETWEE			1	Property D	Other	1	0	0	1	2004/0
2	-122.2797	47.553405	33	1268	1268	3672152	Matched	Intersectio	33499	RAINIER			2	Injury Coll	Rear Ende	3	0	0	2	2013/0

Data explaning

The SEVERITYCODE is the dependent variable that defines the severity of the accidents. It is divided in 5 levels respectively 3 for fatality, 2b for serious injury, 2 for injury, 1 for prop damage and 0 for unknown. The aim of this project is to use other variables to define and predict the SEVERITYCODE.

The ADDRTYPE defines the address of the car accidents, including alley,block and intersection. From this variable we know in what kind of areas the severe accidents are mostly likely to occur.

The LOCATION that describe the general location of the collision. The frequency of the location may be influenced in several ways like the road condition, the light condition and even the character of neighbors. From the LOCATION, we can predict the synthesis situation of those areas.

The SEVERITYDESC is used to define the type of collision, including Injury Collision and Property Damage Only Collision etc, while the COLLISIONTYPE is used to describe the detailed type of collisions including Angles and Sideswipes etc. From those two variables, we can understand the detailed information of this collision.

The PERSONCOUNT, PEDCOUNT, VEHCOUNT, INJURIES,SERIOUSINJURIES, FATALITIES record the total number of people, pedestrians, bicycles, vehicles, injuries, serious injuries and fatalities in the collision respectively. Obviously, those variables like injuries or number of cars are important factors that will influence the severity.

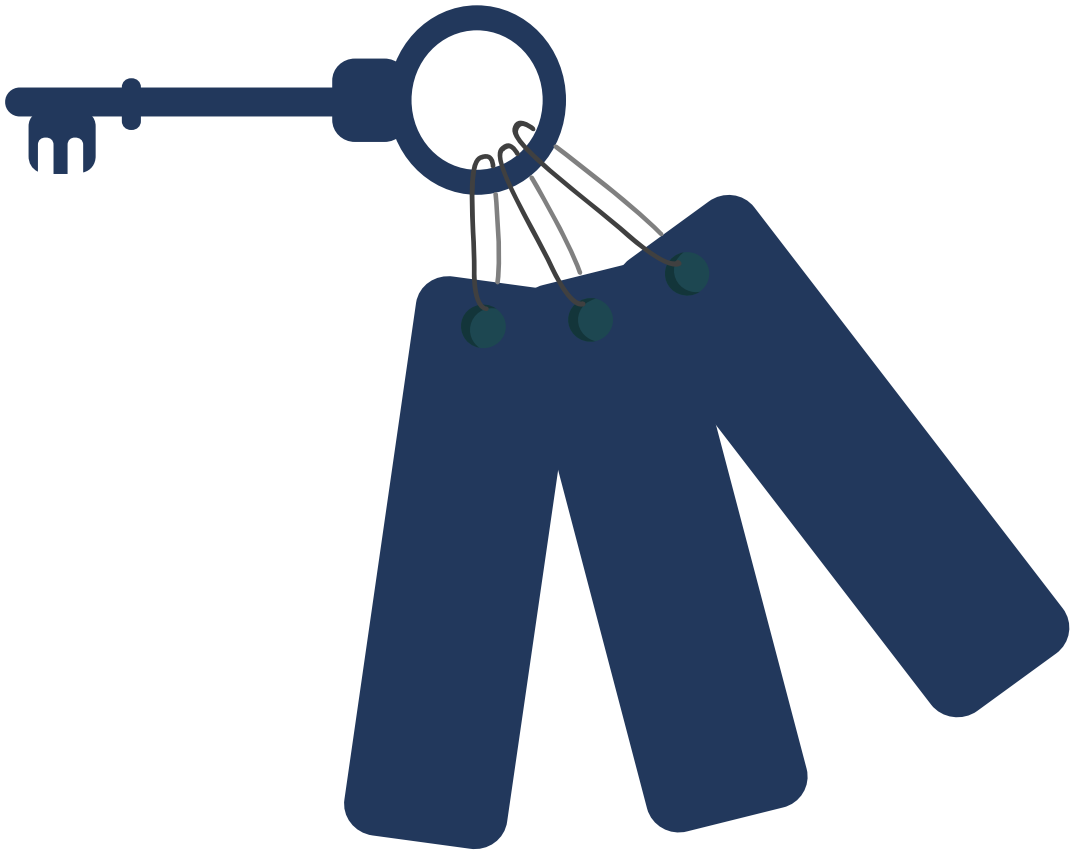
The INCDATE and INCDTTM that define the date and date,time of the incident. It is possible that in some dates the frequency of the accidents will increase like in some holidays festivals that people tend to go out and get drunk. Or even the time of the final exam that students are more likely to stay up till night to study which makes them tired when driving.

Data explaning(2)

The INATTENTIONIND, UNDERINFL,PEDROWNOTGRNT,SPEEDING and HITPARKEDCAR which defines whether whether the accidents was due to inattention, whether the driver involved was under the influence of drugs or alcohol, whether the pedestrian right of way was granted, whether or not speeding was a factor in the collision and whether the collision involved hitting a parked car respectively which could be used to predict the cause of the accident.

The WEATHER, ROADCOND, LIGHTCOND that describe the weather, road and light condition of the accident, which could be used as natural factors of the accident.

Data cleaning



01 **Drop some data**

Drop some columns that are not useful in this project

02 **Fill in the blanks**

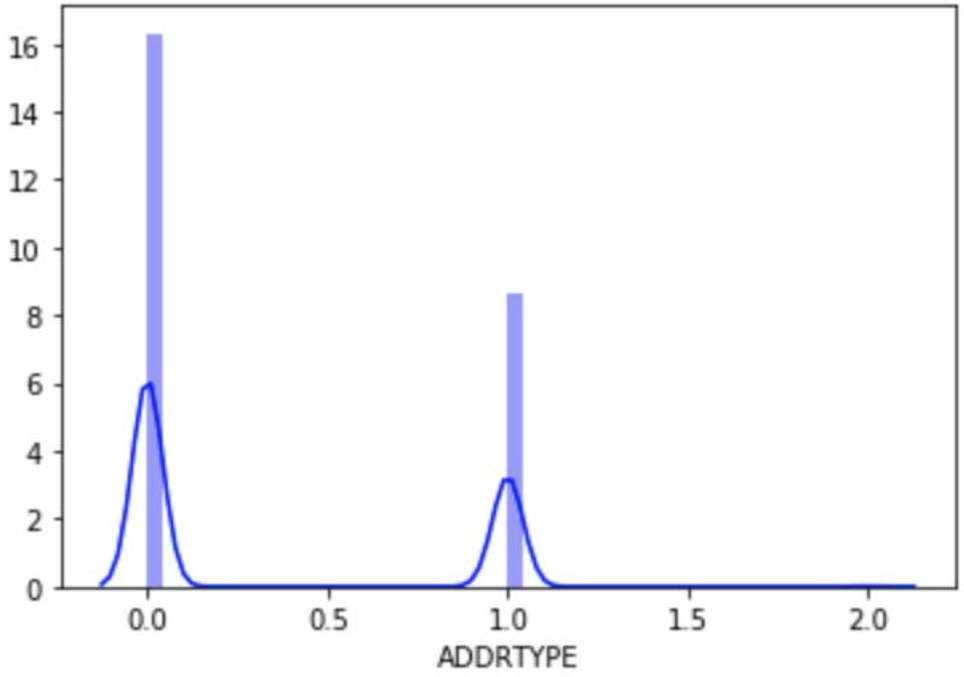
use the dropna of fillna to fill in those blanks.

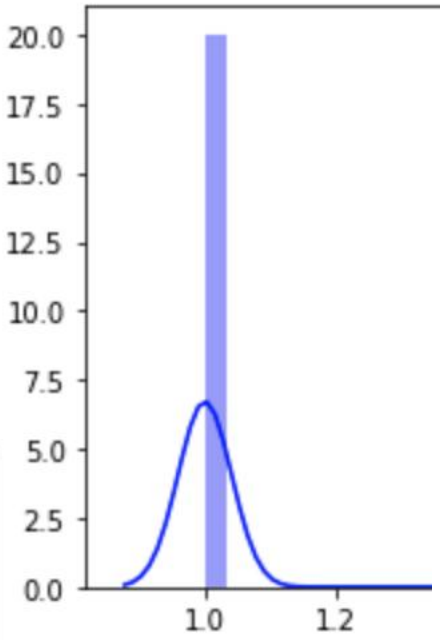
03 **Transfer data type**

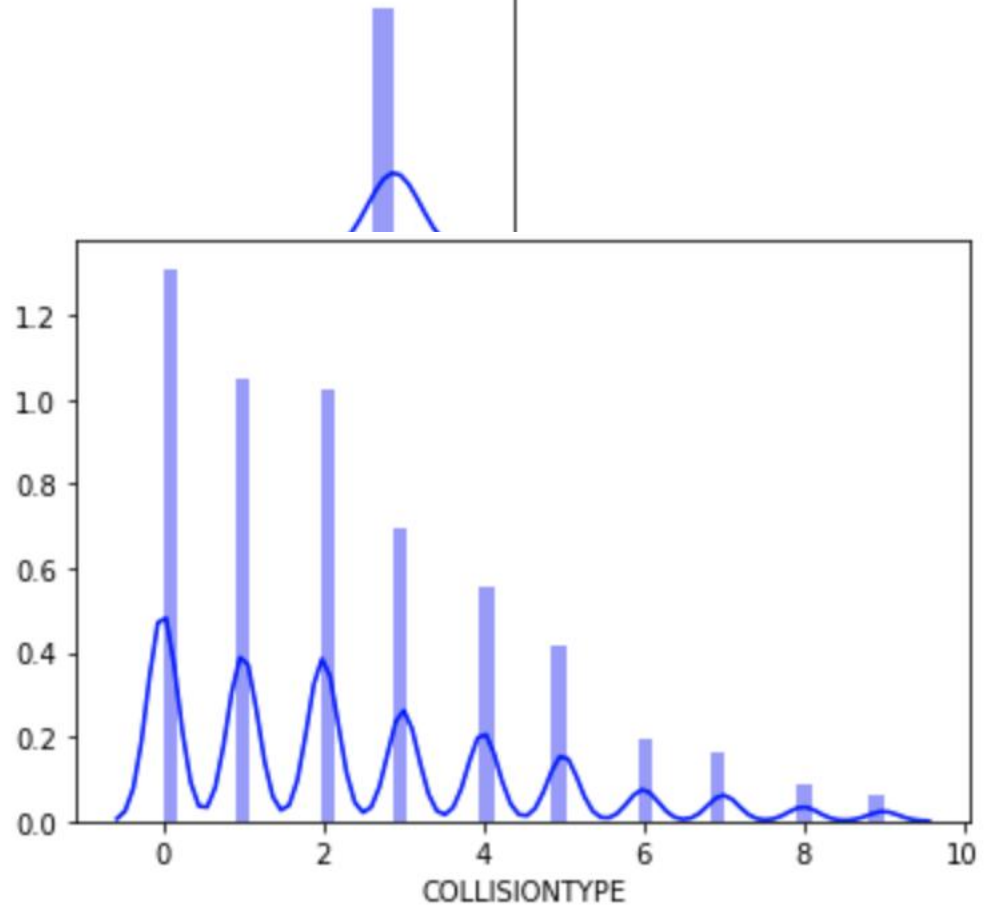
Transfer those categorical string variable to int variable.

Data Visualization

This section we show some distribution plot of importan variables.

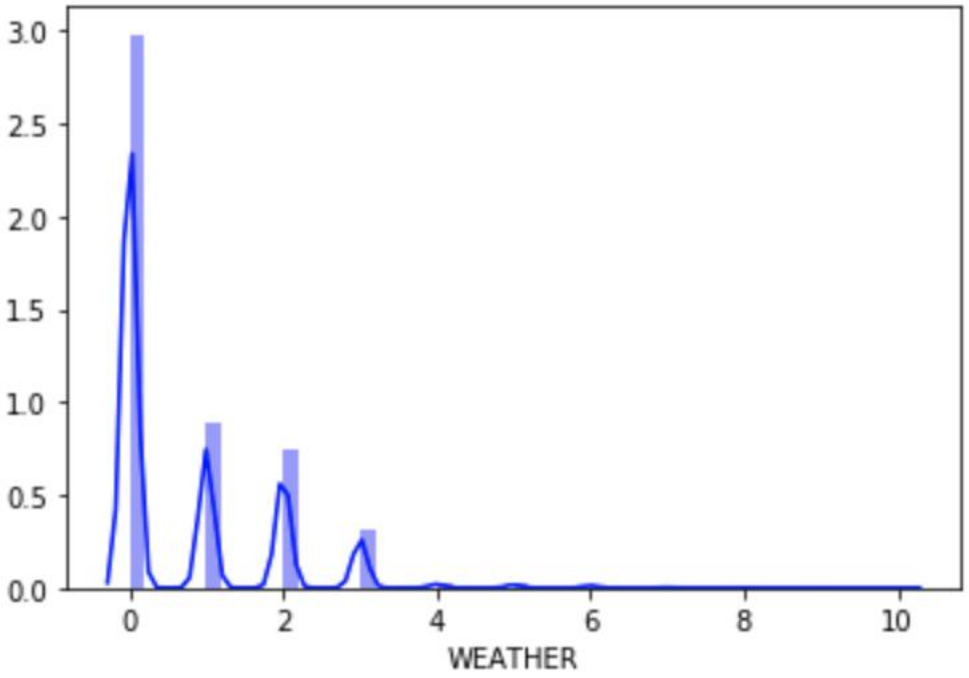


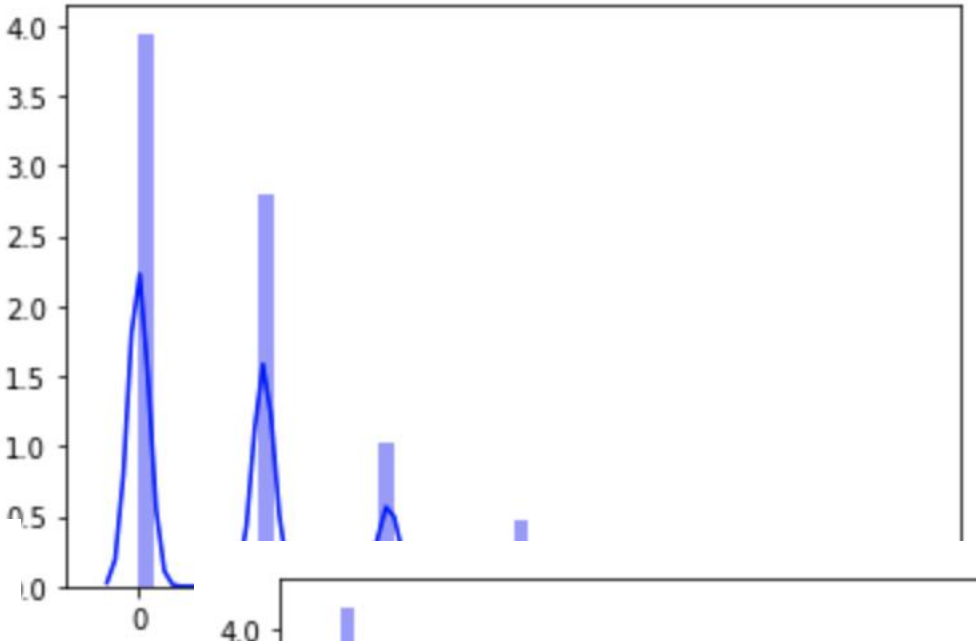


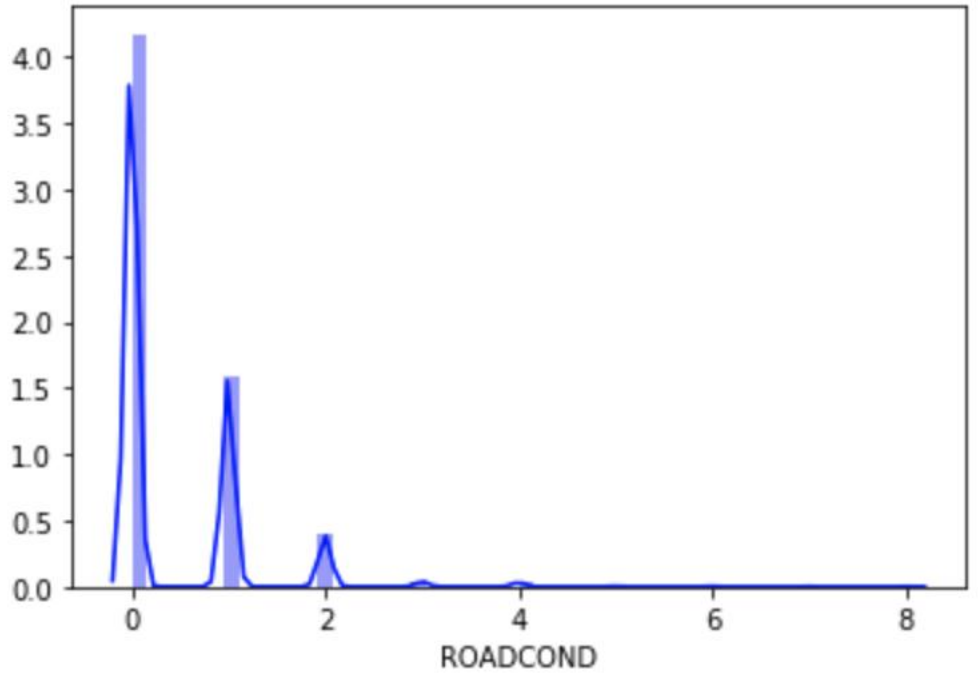


Data Visualization

This section we show some distribution plot of importan variables.



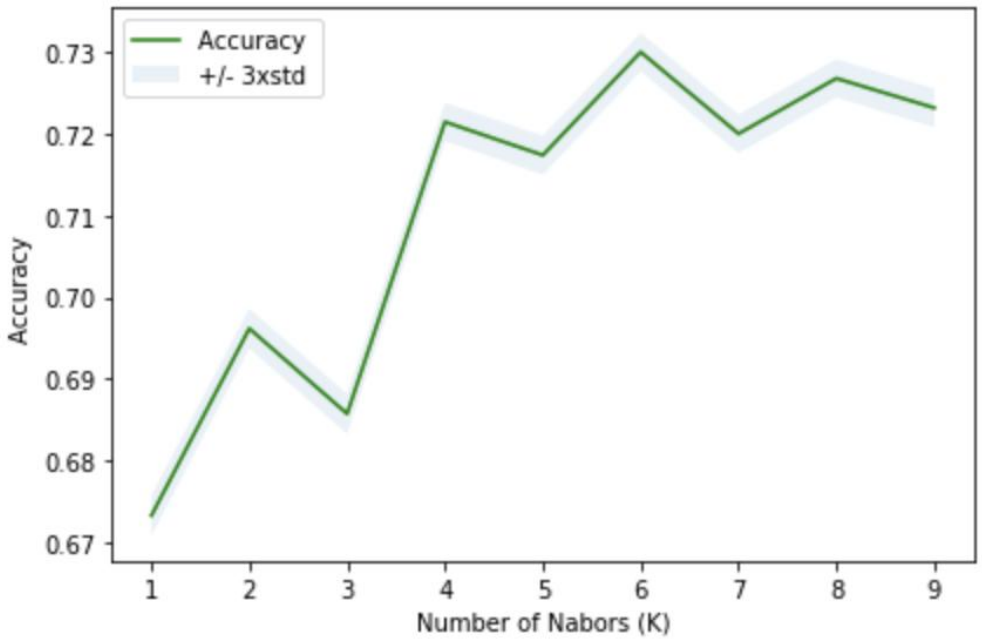




Machine learning

Firstly we apply the k-nearest neighbor to the data. To find the best k, we use the int from 1-10 and test their accuracy and draw the graph:

From this graph, we know that k=6 holds the highest accuracy. Therefore, we apply k=6 to the model and get the result. Then we get the jaccard similarity score to test the accuracy.



Machine learning

After that, we further apply the decision tree model and logistic regression model to the data and get the jaccard similarity score. Then we compare the three accuracy scores and find the highest one.

Method	K-nearest neighbors	Decision tree process	Logistic regression
Score	0.7300636977500752	0.7466305803876541	0.7380190819869324

Thanks