Diabetes Detection Using Machine Learning

Mohamed Faadil

U87311082

Introduction

This data set has been taken from Kaggle. This dataset has a subset of the features from the health-related telephone survey data that is collected annually by the CDC which are considered as important risk factors for diabetes. This dataset is for the year 2015. The dataset contains answers to the questions asked in the survey. Most of the features in the dataset have Yes or No answers indicated by 1 or 0.

Features

Feature Name	Feature Description
HighBP	Yes or No, if a person has been diagnosed with high blood pressure
HighChol	Yes or No, if a person has ever been diagnosed with high cholesterol levels
CholCheck	Yes or No, if cholesterol check in last 5 years
BMI	Body Mass Index
Smoker	Yes or No, ever smoked at least 100 cigarettes in entire life
Stroke	Yes or No, ever had a stroke
HeartDiseaseorAttack	Yes or No, ever reported having coronary heart disease (CHD) or myocardial infarction (MI)
PhysActivity	Adults who reported doing physical activity or exercise during the past 30 days other than their regular job
Fruits	Yes or No, Consume Fruit 1 or more times per day
Veggies	Yes or No, Consume Vegetables 1 or more times per day
HvyAlcoholConsump	Yes or No, Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week)
AnyHealthcare	Yes or No, if a person has any kind of health care coverage, including health insurance, prepaid plans

Features

Feature Name	Feature Description
NoDocbcCost	Yes or No, Was there a time in the past 12 months when you needed to see a doctor but could not because of cost
GenHlth	1, 2,3 means good or better health. 4 or 5 means fair or poor health.
MentHlth	For how many days during the past 30 days was your mental health not good?
PhysHlth	For how many days during the past 30 days was your physical health not good?
DiffWalk	Do you have serious difficulty walking or climbing stairs?
Sex	Gender of person
Age	Age of the person (different numbers used to show age range)
Education	Level; of education (different numbers used to show if in high school, etc.)
Income	Annual Household Income
Diabetes_binary	Yes or No, Ever been diagnosed with diabetes.

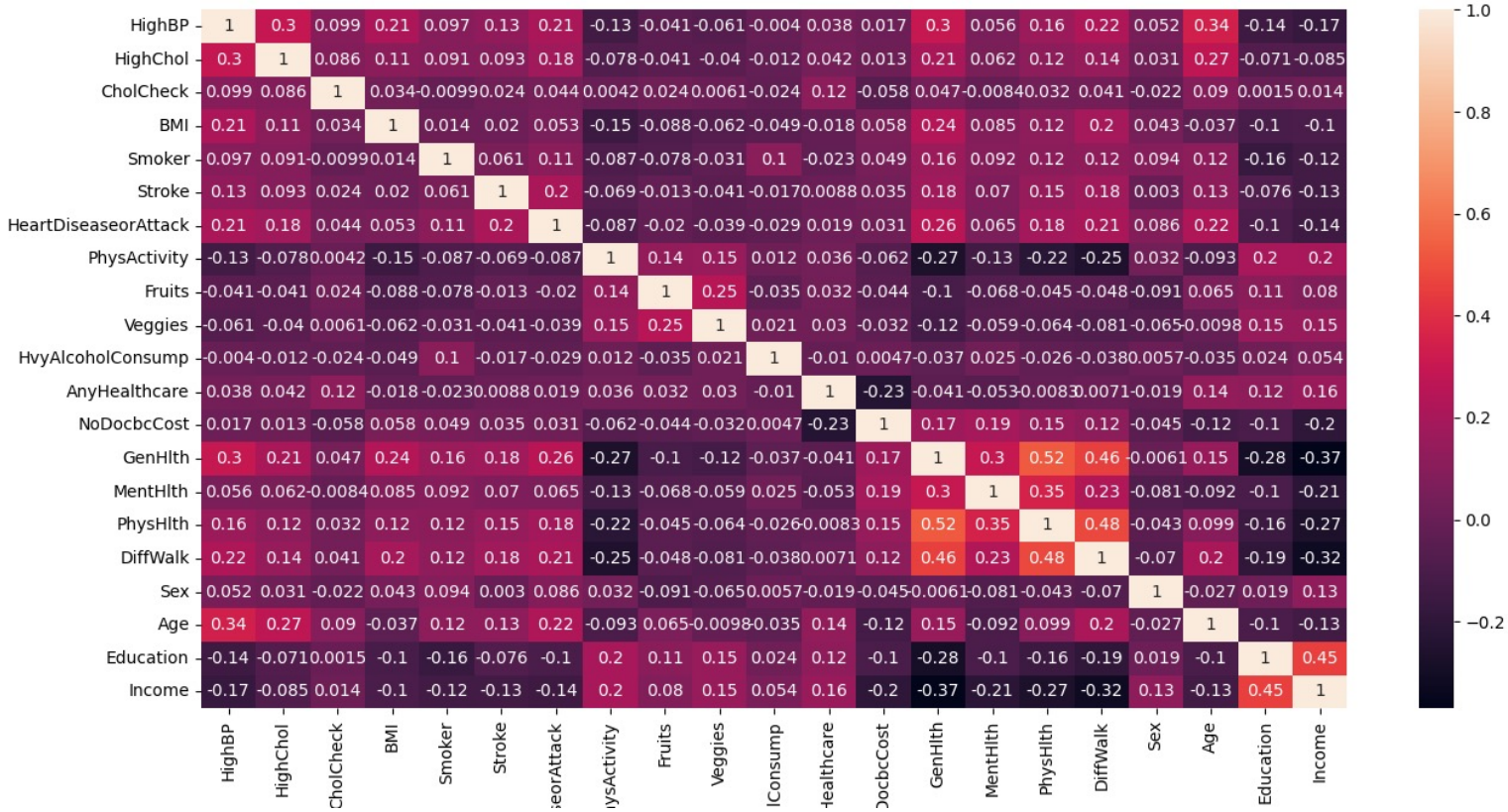
Diabetes CDC BRFSS 2015

Diabetes_binary	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	Veggies	HvyAlcoholConsump	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	PhysHlth	DiffWalk	Sex	Age	E
0.0	1.0	1.0	1.0	40.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	5.0	18.0	15.0	1.0	0.0	9.0	
0.0	0.0	0.0	0.0	25.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	3.0	0.0	0.0	0.0	0.0	7.0	
0.0	1.0	1.0	1.0	28.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	1.0	5.0	30.0	30.0	1.0	0.0	9.0	
0.0	1.0	0.0	1.0	27.0	0.0	0.0	0.0	1.0	1.0	1.0	0.0	1.0	0.0	2.0	0.0	0.0	0.0	0.0	11.0	
0.0	1.0	1.0	1.0	24.0	0.0	0.0	0.0	1.0	1.0	1.0	0.0	1.0	0.0	2.0	3.0	0.0	0.0	0.0	11.0	
0.0	1.0	1.0	1.0	25.0	1.0	0.0	0.0	1.0	1.0	1.0	0.0	1.0	0.0	2.0	0.0	2.0	0.0	1.0	10.0	
0.0	1.0	0.0	1.0	30.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	3.0	0.0	14.0	0.0	0.0	9.0	
0.0	1.0	1.0	1.0	25.0	1.0	0.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	3.0	0.0	0.0	1.0	0.0	11.0	
1.0	1.0	1.0	1.0	30.0	1.0	0.0	1.0	0.0	1.0	1.0	0.0	1.0	0.0	5.0	30.0	30.0	1.0	0.0	9.0	
0.0	0.0	0.0	1.0	24.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	2.0	0.0	0.0	0.0	1.0	8.0	
1.0	0.0	0.0	1.0	25.0	1.0	0.0	0.0	1.0	1.0	1.0	0.0	1.0	0.0	3.0	0.0	0.0	0.0	1.0	13.0	
0.0	1.0	1.0	1.0	34.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	3.0	0.0	30.0	1.0	0.0	10.0	
0.0	0.0	0.0	1.0	26.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	3.0	0.0	15.0	0.0	0.0	7.0	
1.0	1.0	1.0	1.0	28.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	4.0	0.0	0.0	1.0	0.0	11.0	
0.0	0.0	1.0	1.0	33.0	1.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0	1.0	4.0	30.0	28.0	0.0	0.0	4.0	
0.0	1.0	0.0	1.0	33.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	2.0	5.0	0.0	0.0	0.0	6.0	
0.0	1.0	1.0	1.0	21.0	0.0	0.0	0.0	1.0	1.0	1.0	0.0	1.0	0.0	3.0	0.0	0.0	0.0	0.0	10.0	
1.0	0.0	0.0	1.0	23.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	2.0	0.0	0.0	0.0	1.0	7.0	
0.0	0.0	0.0	0.0	23.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	2.0	15.0	0.0	0.0	0.0	2.0	
0.0	0.0	1.0	1.0	28.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	2.0	10.0	0.0	0.0	1.0	4.0	
0.0	1.0	1.0	1.0	22.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0	3.0	30.0	0.0	1.0	0.0	12.0	
0.0	1.0	1.0	1.0	38.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	5.0	15.0	30.0	1.0	0.0	13.0	
0.0	0.0	0.0	1.0	28.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	3.0	0.0	7.0	0.0	1.0	5.0	
1.0	1.0	0.0	1.0	27.0	0.0	0.0	0.0	1.0	1.0	1.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	13.0	
0.0	1.0	1.0	1.0	28.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	3.0	6.0	0.0	1.0	0.0	9.0	
0.0	0.0	0.0	1.0	32.0	0.0	0.0	0.0	1.0	1.0	1.0	0.0	1.0	0.0	2.0	0.0	0.0	0.0	0.0	5.0	
1.0	1.0	1.0	1.0	37.0	1.0	1.0	1.0	0.0	0.0	1.0	0.0	1.0	0.0	5.0	0.0	0.0	1.0	1.0	10.0	
1.0	1.0	1.0	1.0	28.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0	1.0	0.0	4.0	0.0	0.0	0.0	1.0	12.0	
1.0	1.0	1.0	1.0	27.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	4.0	20.0	20.0	1.0	0.0	8.0	
0.0	0.0	1.0	1.0	31.0	1.0	0.0	0.0	1.0	1.0	1.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	12.0	

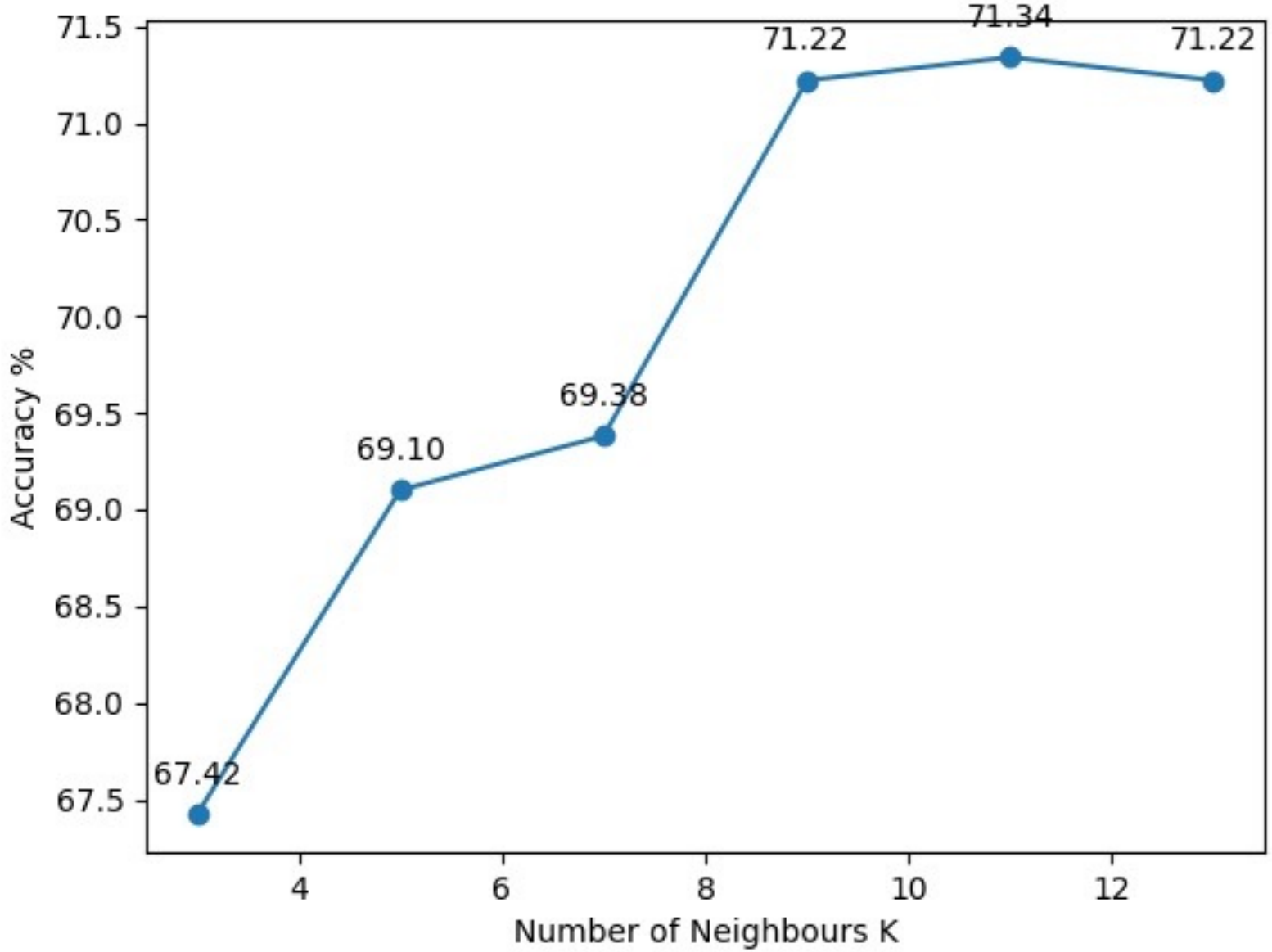
Dataset Description

- No null values found in the data set.
- Dimensions of the dataset: (253680, 22)
- Number of people in Diabetes Class: 35346
- Number of people in Non-Diabetes Class: 218334
- Random sample of 5000 rows each from the two classes, meaning a total of 10000 rows.
- Supervised Learning.

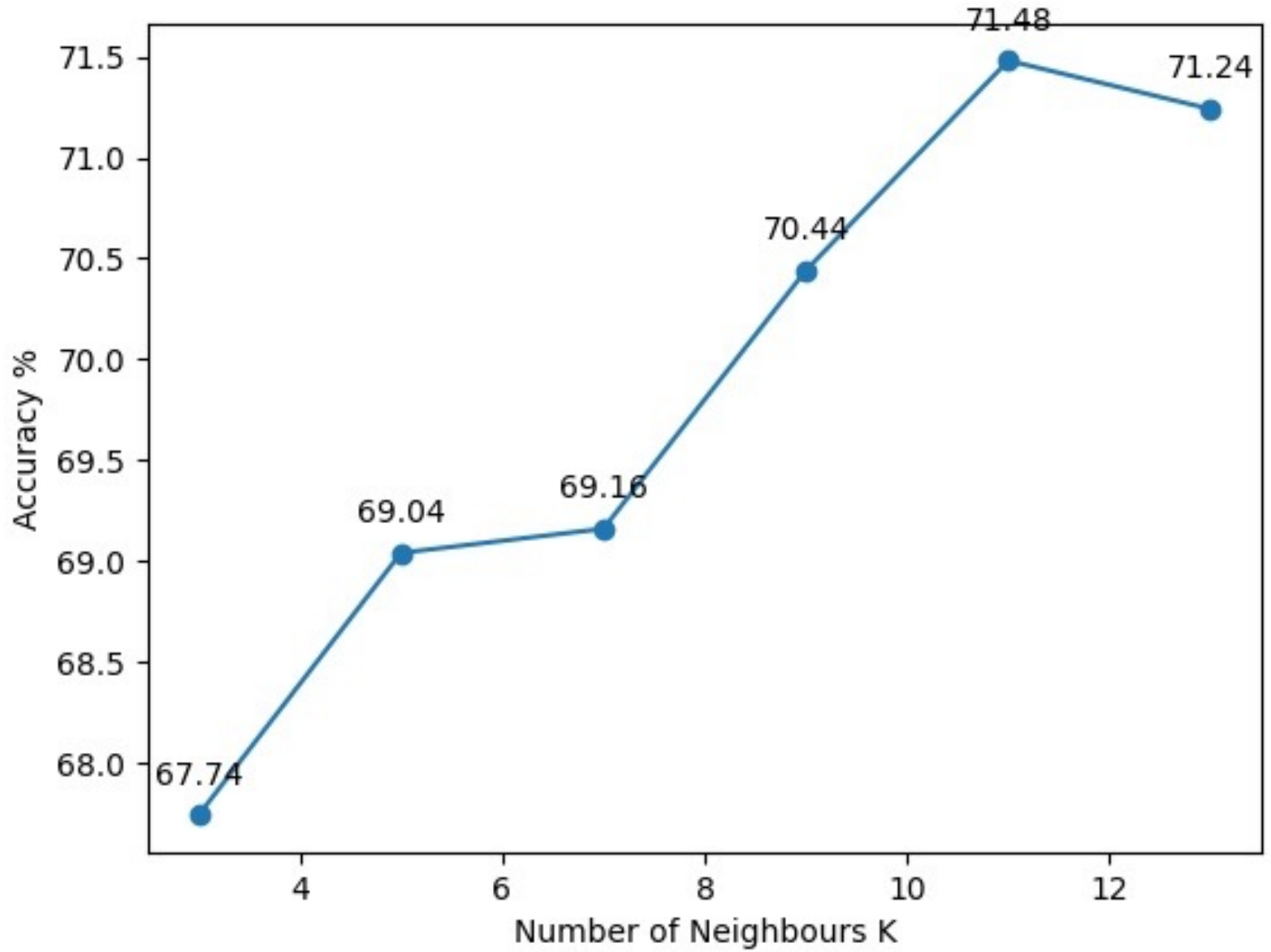
Pearson Correlation Plot



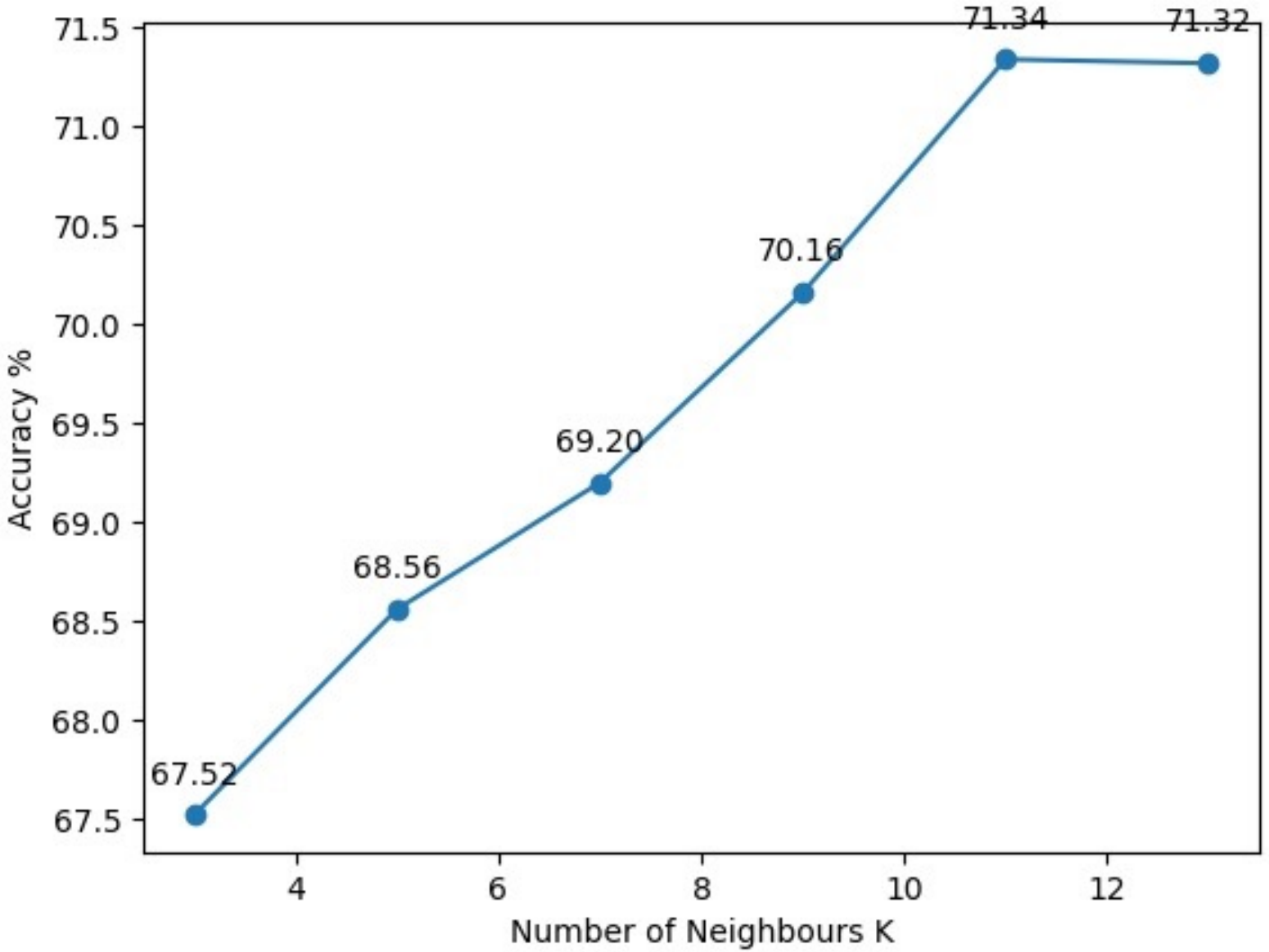
K Nearest Neighbor P = 1



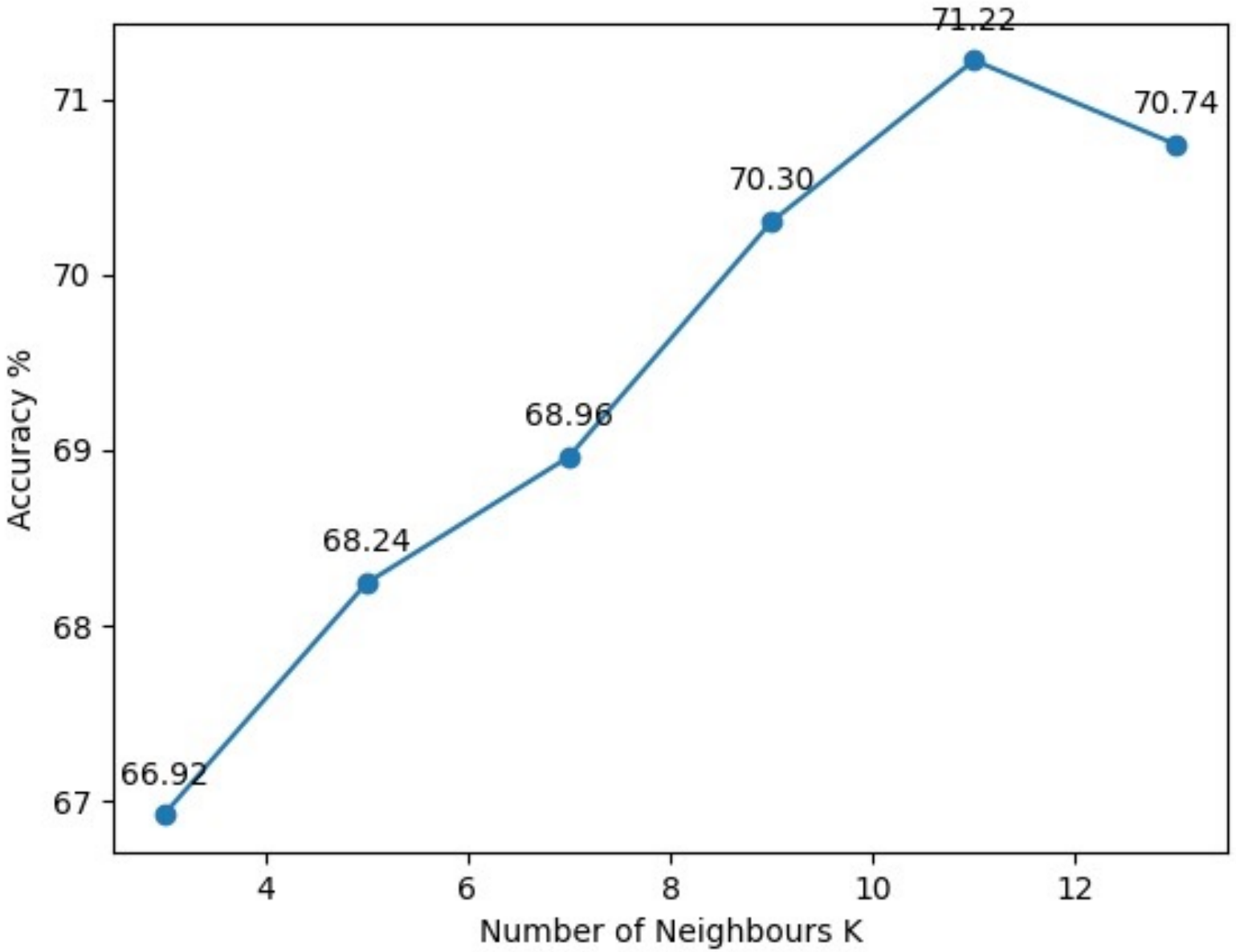
K Nearest Neighbor P = 1.5



K Nearest Neighbor P = 2



K Nearest Neighbor P = 3



K Nearest Neighbor Conclusion

- For all values of P, the highest accuracy is for 11 neighbors.
- The max accuracy for all values of P is almost the same, the highest being 71.48% for P = 1.5
- Would prefer P = 1 as not much computation required and can be easily run for large datasets such as this.

Logistic Regression

- Accuracy for Logistic Regression with all features: 74.22%
- Weights for the features are as follows:

Feature Name	Weight	Feature Name	Weight	Feature Name	Weight
HighBP	0.85668608	PhysActivity	0.07846623	MentHlth	-0.15072095
HighChol	0.62942776	Fruits	-0.13423889	PhysHlth	-0.25841351
CholCheck	1.26243219	Veggies	-0.13935377	DiffWalk	-0.08787142
BMI	5.40626823	HvyAlcoholConsump	-0.64823837	Sex	0.13613309
Smoker	-0.04415837	AnyHealthcare	0.25729719	Age	1.61390534
Stroke	0.07920962	NoDocbcCost	0.23874808	Education	-0.47340977
HeartDiseaseorAttack	0.24708055	GenHlth	2.54029155	Income	-0.23468117

Logistic Regression

- Removing features with negative weights and checking accuracy. Number of features removed are 9.
- Accuracy for Logistic Regression without features with negative weights: 74.36%
- Accuracy slightly increased and we have also significantly reduced the size of the data set.

Feature Name	Weight	Feature Name	Weight	Feature Name	Weight
HighBP	0.78964572	PhysActivity	0.05887075	MentHlth	
HighChol	0.58008288	Fruits		PhysHlth	
CholCheck	1.0561844	Veggies		DiffWalk	
BMI	5.89740956	HvyAlcoholConsump		Sex	0.15586659
Smoker		AnyHealthcare	0.05962112	Age	1.5132662
Stroke	0.14197646	NoDocbcCost	0.03727246	Education	
HeartDiseaseorAttack	0.39350825	GenHlth	2.44785979	Income	

Linear Discriminant

- Accuracy for Linear Discriminant: 75.0%
- Weights for the features are as follows:

Feature Name	Weight	Feature Name	Weight	Feature Name	Weight
HighBP	0.86345273	PhysActivity	0.05923623	MentHlth	-0.01639663
HighChol	0.59304363	Fruits	-0.05458343	PhysHlth	-0.30914761
CholCheck	0.75594544	Veggies	-0.17954158	DiffWalk	-0.0608103
BMI	6.96269783	HvyAlcoholConsump	-0.60084862	Sex	0.17021504
Smoker	-0.03948506	AnyHealthcare	0.29443414	Age	1.7600519
Stroke	0.17766079	NoDocbcCost	0.13171834	Education	-0.22667397
HeartDiseaseorAttack	0.27907794	GenHlth	2.55747984	Income	-0.18176663

Linear Discriminant

- Removing features with negative weights and checking accuracy. Number of features removed are 9.
- Accuracy for Linear Discriminant without features with negative weights: 74.36%
- Accuracy slightly increased and we have also significantly reduced the size of the data set.

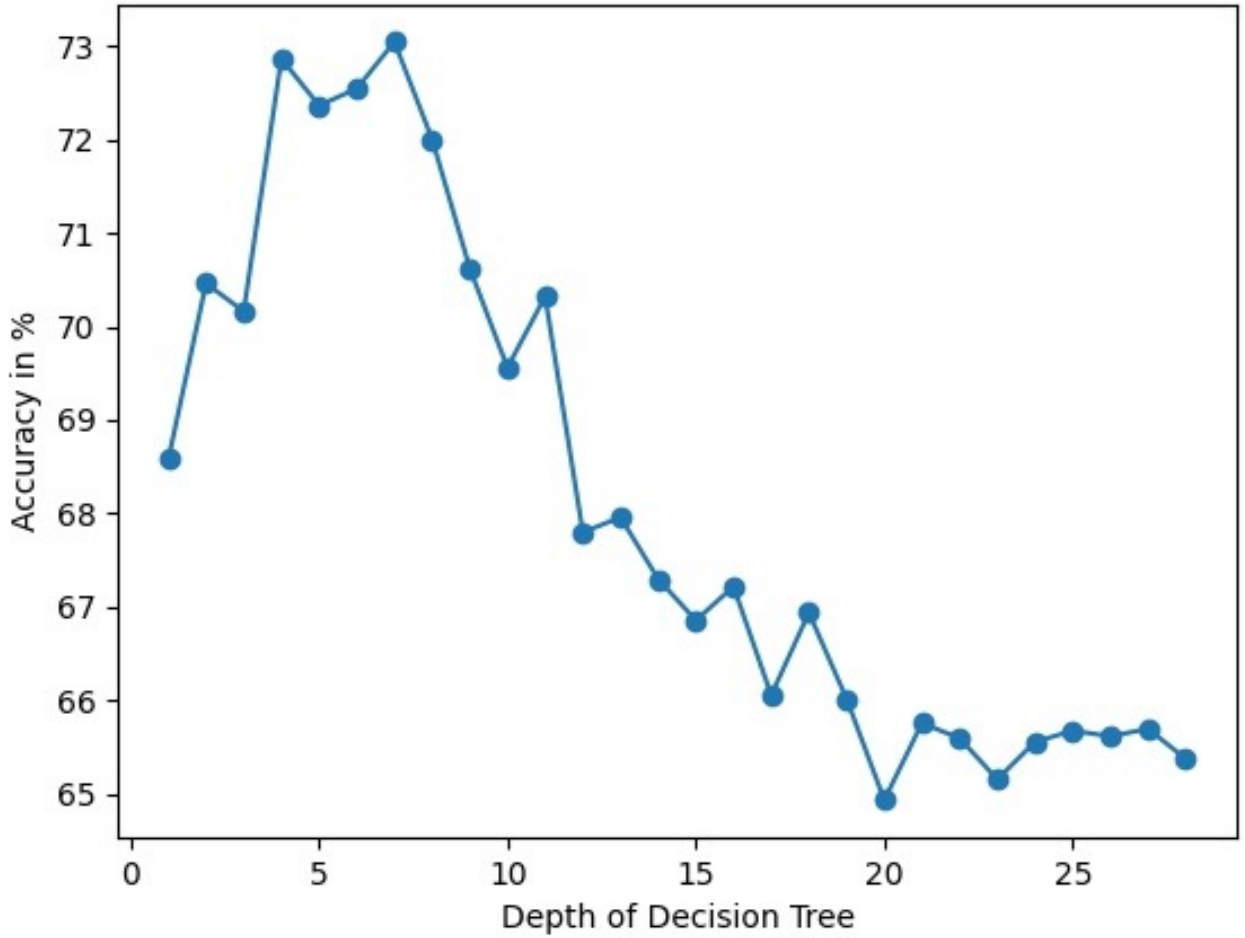
Feature Name	Weight	Feature Name	Weight	Feature Name	Weight
HighBP	0.78964572	PhysActivity	0.05887075	MentHlth	
HighChol	0.58008288	Fruits		PhysHlth	
CholCheck	1.0561844	Veggies		DiffWalk	
BMI	5.89740956	HvyAlcoholConsump		Sex	0.15586659
Smoker		AnyHealthcare	0.05962112	Age	1.5132662
Stroke	0.14197646	NoDocbcCost	0.03727246	Education	
HeartDiseaseorAttack	0.39350825	GenHlth	2.44785979	Income	

Quadratic Discriminant

- Accuracy for Quadratic Discriminant: 70.72%

Decision Tree

- Max accuracy for Decision Tree is: 73.04% for depth: 7



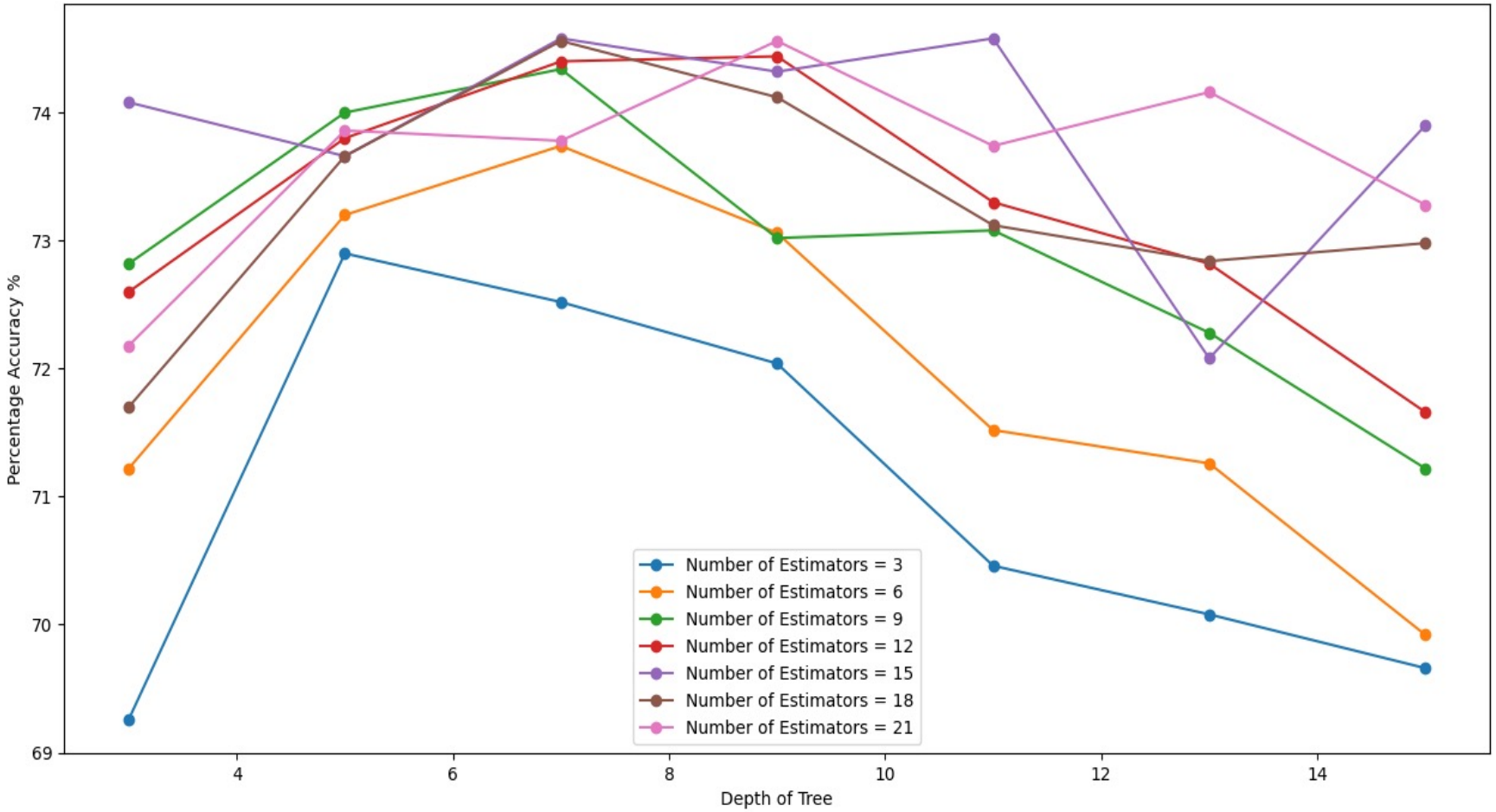
Decision Tree

- The feature importance are as follows:

Feature Name	Importance	Feature Name	Importance	Feature Name	Importance
HighBP	18.67897192	PhysActivity	0.49800583	MentHlth	2.54409196
HighChol	5.42536772	Fruits	0.77067695	PhysHlth	2.39353647
CholCheck	0.53544433	Veggies	0.65808931	DiffWalk	0.71277698
BMI	13.68247227	HvyAlcoholConsump	0.38523137	Sex	0.22795587
Smoker	0.35521333	AnyHealthcare	0.22753388	Age	8.34774112
Stroke	0.2630675	NoDocbcCost	0.38255329	Education	2.56276888
HeartDiseaseorAttack	0.97767836	GenHlth	38.44944428	Income	1.92137837

Random Forest

- Max Accuracy for Random Forest is: 74.58% for 15 estimators and depth: 7



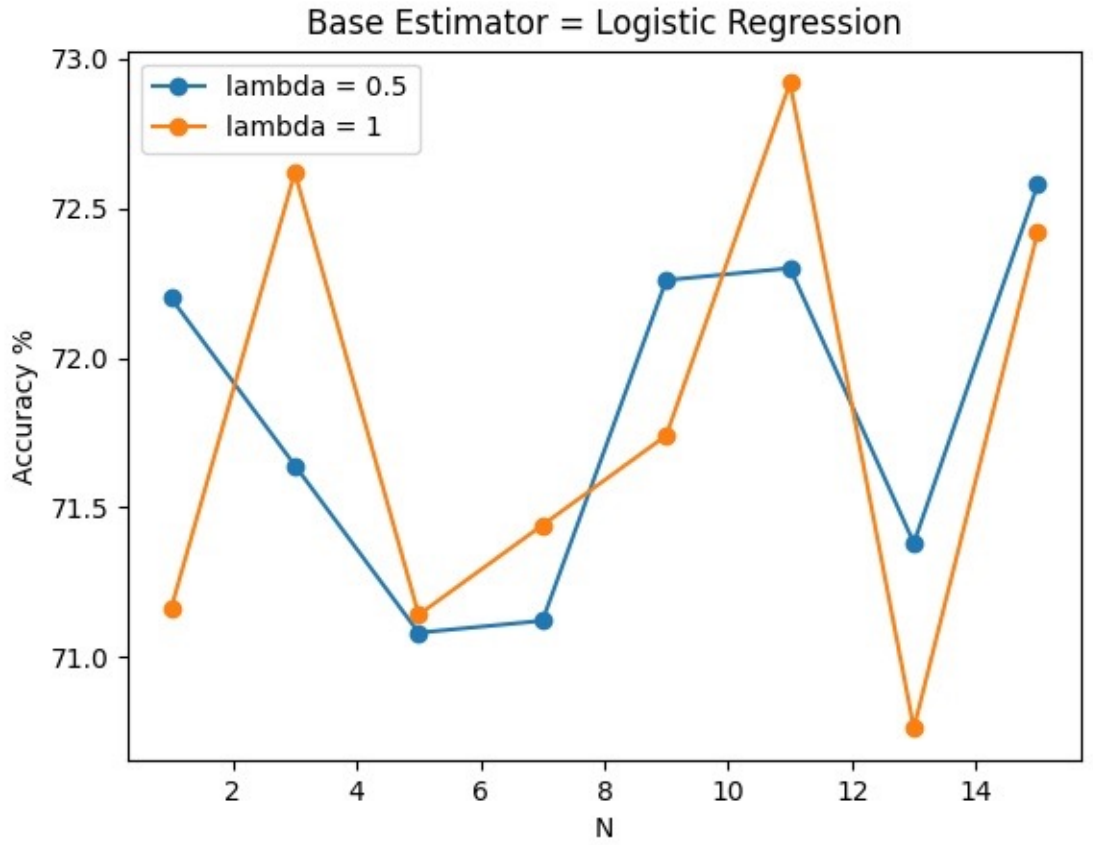
Random Forest

- The feature importance are as follows:

Feature Name	Importance	Feature Name	Importance	Feature Name	Importance
HighBP	20.08	PhysActivity	1.92	MentHlth	1.98
HighChol	7.62	Fruits	0.83	PhysHlth	3.85
CholCheck	0.52	Veggies	0.70	DiffWalk	4.11
BMI	16.14	HvyAlcoholConsump	0.92	Sex	0.70
Smoker	0.80	AnyHealthcare	0.20	Age	11.28
Stroke	0.93	NoDocbcCost	0.54	Education	2.33
HeartDiseaseorAttack	2.60	GenHlth	16.67	Income	5.27

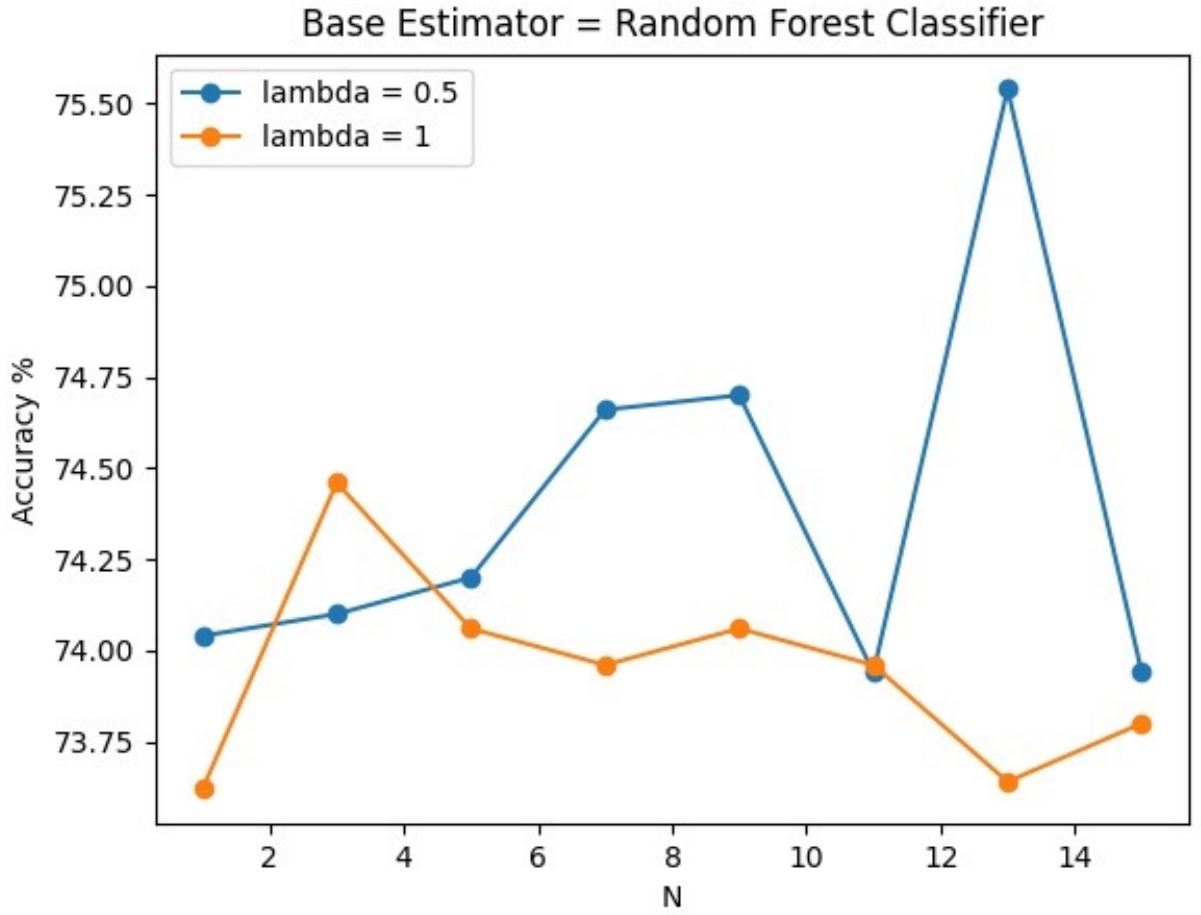
Adaboost – Logistic Regression

- Highest Accuracy for Adaboost with Logistic Regression is: 72.92% with Learning rate: 1 & 11 base estimators.



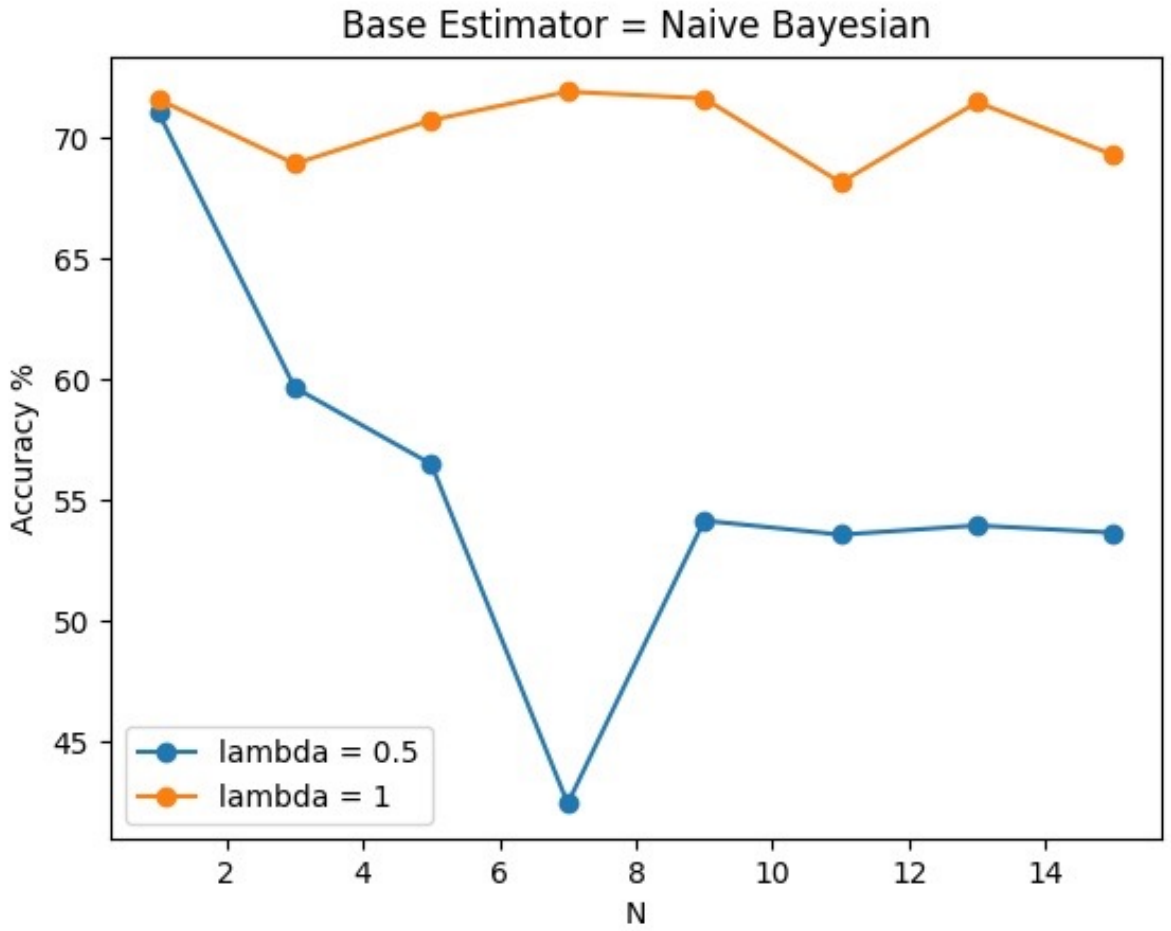
Adaboost – Random Forest

- Highest Accuracy for Adaboost with Random Forest is: 75.54% with Learning rate: 0.5 & 13 base estimators.



Adaboost – Naïve Bayesian

- Highest Accuracy for Adaboost with Naive Bayesian is: 71.90% with Learning rate: 1 & 7 base estimators.



Support Vector Machine

- Accuracy for Linear SVM is: 74.8399999999999%
- Accuracy for Gaussian SVM is: 73.88%
- Accuracy for Polynomial SVM of degree 2 is: 73.88%
- Accuracy for Polynomial SVM of degree 3 is: 73.88%
- Accuracy for Polynomial SVM of degree 4 is: 73.88%
- Accuracy for Polynomial SVM of degree 5 is: 73.88%

Results

Machine Learning Classifier	Accuracy %	Machine Learning Classifier	Accuracy %
KNN with P = 1	71.34%	Adaboost with Naïve Bayesian	71.90%
KNN with P = 1.5	71.48%	Linear SVM	74.84%
KNN with P = 2	71.34%	Gaussian SVM	73.88%
KNN with P = 3	71.22%	Polynomial SVM Degree 2	73.88%
Logistic Regression	74.36%	Polynomial SVM Degree 3	73.88%
Naïve Bayesian	71.90%	Polynomial SVM Degree 4	73.88%
Linear Discriminant	75.00%	Polynomial SVM Degree 5	73.88%
Quadratic Discriminant	70.72%		
Decision Tree	73.04%		
Random Forest	74.58%		
Adaboost with Logistic Regression	72.92%		
Adaboost with Random Forest	75.54%		

Conclusion

- Adaboost with Random Forest has the highest accuracy of 75.54%
- Quadratic Discriminant has the lowest accuracy of 70.72%
- Prefer using Logistic Regression as very easy to compute and low sensitivity to changes in the data set. Accuracy is 74.36% which is very close to the maximum accuracy and, this accuracy has been calculated using the dataset with reduced features. Only 12 features of the dataset were used.

K Means Clustering

- Not suitable for implementation. Entropy too high & requires too many clusters to reduce it.