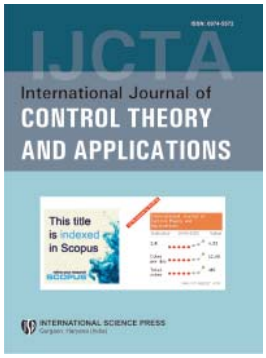

International Journal of Control Theory and Applications

ISSN : 0974–5572

Volume 10 • Number 23 • 2017

Prediction Model on Knee Osteoarthritis

K. Vijay Kumar[a] K. Shyamala[a] and D. Naresh[c]

[a]*Research Scholar, Vels University Chennai, India*

E-mail: vijaydharshan@gmail.com

[b]*Associate Professor, Dr. Ambedkar Government College Chennai, India*

E-mail: shyamalakannan2000@gmail.com

[c] *?*

E-mail: ed_naresh@rocketmail.com

Abstract: Healthcare Analytics, a multidisciplinary field deals with computer science, biology, medicine and data science has vast potential which remains underutilized. Big Data Analytics being the buzz of IT industry enlightens this Healthcare Informatics. In this paper, we attempt to address age relevant diseases, from a real world dataset on Osteoarthritis (OA); the identified attributes were taken from a healthcare database to build a predictive model for Osteoarthritis risk prediction [1][3]. Data cleansing techniques were applied to make the data conducive enough for performing analytics to show the risk of a patient obtaining OA. Evidence to show the modifiable factors that influence in acquiring OA was carried out to showcase the predictive modelling. Machine Learning Algorithms, Logistic regression (LR) and Naive Bayes (NB) were used to predict the estimated risk on a patient's chance of obtaining OA. The cases of both the Incident Knee and the Symptomatic Knee risks [1][3] were applied on Supervised Machine Learning Algorithms. The accuracy of the algorithms was checked by Receiver Operating Characteristics (ROC). The prediction values were checked with an Osteoarthritis data from other source. The performances of the algorithms were compared to show the best fitting model for prediction.

Keyword: *Big data Analytics, Osteoarthritis, Prediction Model, Machine Learning Algorithms.*

1. INTRODUCTION

Big Data Analytics plays a burgeoning role in healthcare domain by handling hurdles in the form of distributed, disjoint and diverse datasets in its roadmap. Apache Hadoop being a distributed framework provides the foundation for managing and processing any data. Applying data mining techniques with enhancements of tools and techniques to handle the characteristics of Big Data is the key answer for the enlightenment of Big Data Analytics. R tool rich in built-in libraries and functions is capable of visualizing and presenting data. In this paper, our initiative is to demonstrate the heuristics applicable on the hospital data requesting anonymity to build a predictive model to estimate the risk of acquiring OA in patients. The entire experiment was carried out on 'R' analytical and visualizing tool. Our next initiative will be to work on such applications by leveraging RHadoop. To run the processing in parallel paradigm, R integrated with Hadoop called RHadoop can allow users to manage, analyze and visualize data. Basically RHadoop comes with 5 packages namely *rhdfs*, *rmr*2, *plyrmr*, *ravro* and *rhbase* [10].

By resourcefully putting the data into use, there can be many solutions that can be unlocked in diagnosing and treating the patients. Osteoarthritis (OA), a prevalent arthritic disorder, is one of a kind on age relevant diseases. Prediction model to estimate the risk of OA was built with guidance of a literature survey and not on the expertise of medical specialist. The predictors that cause this condition in patients were found from a healthcare database on surveying literatures. The algorithm giving the highest accuracy was adapted to find the risk of a patient acquiring OA. Like OA, predictors of many other medical conditions can be found and a consolidated data of those predictors can be prepared that helps to predict a patient's future chance of getting a disease.

2. RELATED LITERATURE

Predictive model is basically built from statistical determination using past data to find what is likely to happen in the future, such that some pre-emptive measures could be taken to maximize the efficiency [6]. A model itself is not complex, it depends on the process and work required to build an efficient model. Each model built is effective only to a specific problem definition.

In Healthcare, Predictive Analytics (PA) benefit the medical community viz., insurance companies, physicians, health providers, medicine providers and patients in all their walks of workflow to attain optimization and efficiency. The decisions of PA in the domain like Health are not to replace the judgments of doctors but to assist them. PA helps not only in predictions, but also reveals associations in data that a superhuman understanding will never suspect. Reinstating as in [9], PA differs from conventional statistics and evidence based medication: first, predictions for patient profiling and not for groups; second PA does not rely on a bell-shaped curve.

Predictive modelling applications are applied when one of the main components in predictive modeling is the application of machine learning algorithms [11]. The complexity of a model depends on the processes and work required to generate a good model [6]. The workflow towards developing an optimized predictive model involves a strategy that can be applied before making the data conducive enough for building prediction model.

The Figure 1 depicts the workflow to develop predictive models. We are required to define the objectives to the strategic plan. Apply pre-processing techniques to create datasets conducive enough to apply machine learning algorithms. In the course of building algorithm, the success of prediction depends on the score obtained eventually leading to implementation of the predictive model on the best algorithm.

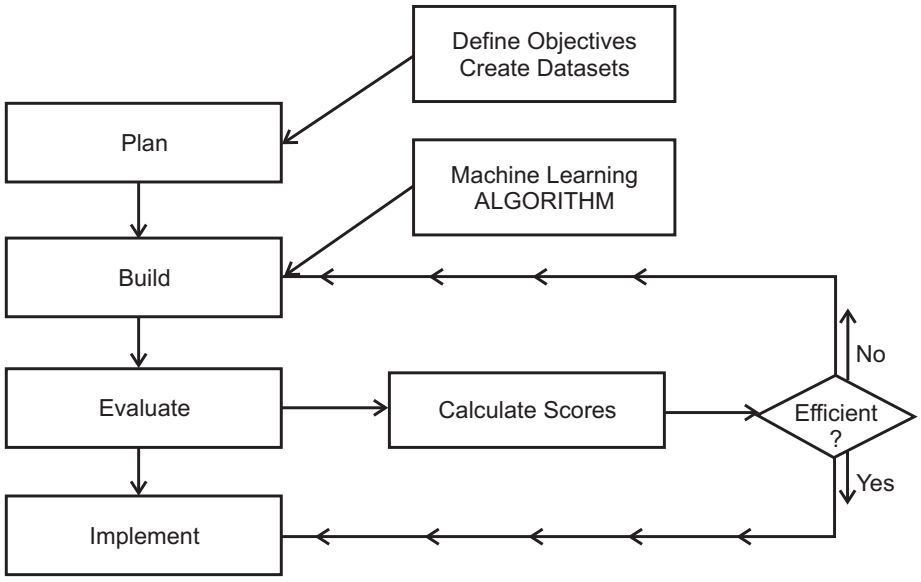

Figure 1: Workflow involved in building Predictive Model

Feature Extraction helps to identify the objectives for customizing the prediction model by selecting the attributes that influences our objectives [5]. After Feature Selection, it is necessary to identify the composite attributes to build the feature rich selection. The importance of attributes is one of the prime steps to build models in NB or SVM. This enhances the efficiency of supervised learning [6].

The risk factors that help in finding the risk of OA in a patient were identified by the literatures [1][3], the attributes which acts as a cause for OA. The usage of Machine Learning Algorithms for data stated in [1][2][3] [12] were very much helpful in doing Analytics practically explaining the way of how algorithms are used. [1] [4] helped to find the accuracy of the models that was used in the risk prediction development like Odd Ratio (OR) and Receiver Operating Characteristics (ROC). On the whole, every reference was very much helpful in the final outcome of the OA risk prediction.

3. MATERIALS AND METHODS

In relation to the above section, number of predicting variables could be found to predict a condition. Since the condition Osteoarthritis is an age relevant disease, all the possible ways of attributes to predict this condition has to be analysed. The coming 9 predictors play an important role in prediction of Osteoarthritis with a person [7][8].

1. Age (in years)
2. Gender (Male-0, Female-1)
3. BMI (Kg/m2)
4. Work Risks (never-0, very rarely-1, sometimes-2, often-3, always-4)
5. Family OA History(yes[first relative with OA]-1, no[first relative with no OA]-0)
6. Knee Injury History (yes[Injured]-1, no[not injured]-0)
7. Knee Pain for a Month (yes[pain for a month]-1, no[no pain]-0)
8. Sports Activity History (yes[played]-1, no[no sports activity]-0)
9. Tibiofemoral and Platellofemoral Radiographic Grade (definite-5, multiple-4, large-3)

The analytics portion involves the prediction of Osteoarthritis Risk in a patient coming for diagnosis. Two of the Machine Learning Algorithms ‘Logistic Regression’ and ‘Naive Bayes’ were used on the data. Validation of the algorithm run was done using Receiver Operating Characteristic (ROC). The performance of each of the fitted models should be noted and the algorithm with highest accuracy can be implemented for Osteoarthritis Risk Prediction. The descriptions of the algorithms used in this project are as follows,

3.1. Logistic Regression

Logistic Regression is a supervised classification algorithm used for predicting an event by estimating the relationship between the dependent variable and the independent variable. There can be one or more independent variables and the dependent variable should be categorical to use this algorithm. It is prevalently used in qualitative response models. We have used this algorithm to leverage the capability in providing the functionality of forecasting group associations, *i.e.*, the ability to predict the class of individuals.

The numerical values are changed to nominal values for Logistic Regression. This is done because when the summary of the model is printed the result gives a clean understanding of the labels in data and how it shows probabilities.

Table 1
Snippet of data built for LR

Age	*Gender*	*bmi*	*Occupational Risk*	*Radiographic Grade*	*Family History*	*Knee Injury*	*Knee Pain*	*Sports activity*
73	(Male)	32.6	(Never)	3	(First relative with no OA)	(Not Injured)	(No Pain)	(Played)
45	(Female)	29.7	(Sometimes)	3	(First relative with no OA)	(Not Injured)	(No Pain)	(Played)
60	(Female)	26.7	(Sometimes)	3	(First relative with no OA)	(Not Injured)	(No Pain)	(No Sports Activity)
77	(Female)	27.5	(Sometimes)	2	(First relative with no OA)	(Injured)	(No Pain)	(No Sports Activity)
48	(Female)	36.9	(Sometimes)	3	(First relative with no OA)	(Injured)	(Pain for a month)	(Played)
48	(Female)	26.7	(Very Rarely)	3	(First relative with no OA)	(Not Injured)	(No Pain)	(Played)
69	(Male)	36.6	(Very Rarely)	3	(First relative with no OA)	(Injured)	(No Pain)	(Played)
55	(Female)	26.7	(Often)	2	(First relative with no OA)	(Not Injured)	(No Pain)	(Played)

Snippet of the code for LR

```
OA$gender<-factor(OA$gender)
OA$radiographic.grade<-factor(OA$radiographic.grade)
OA$occupational.risk<-factor(OA$occupational.risk)
OA$sports.activity<-factor(OA$sports.activity)
OA$family.history<-factor(OA$family.history)
OA$knee.injury<-factor(OA$knee.injury)
OA$knee.pain<-factor(OA$knee.pain)
logistc<- glm(radiographic.grade ~ age + gender + bmi + occupational.risk + family.history
+ knee.injury + sports.activity + knee.pain, data = OA, family =”binomial”)
summary(logistc)
```

3.2. Naïve Bayes

Naive Bayes is another supervised classification algorithm that judges documents as belonging to one category or the other. The classification is done by word frequency which acts as a feature. This is applicable when dependent variable has classes, where the classification becomes an inference in the probabilistic model.

For processing in NB, the independent variable has to be in classes and the programming does not take processing in numerical values. So independent variable radiographic grade is converted to nominal for proper results and all other values are made numerical.

Table 2
Snippet of Data built for NB

Age	*Gender*	*Bmi*	*Occupational Risk*	*Radiographic Grade*	*Family History*	*Knee Injury*	*Knee Pain*	*Sports activity*
73	0	32.6	0	multiple	0	0	0	1
45	1	29.7	2	multiple	0	0	0	1
60	1	26.7	2	multiple	0	0	0	0
77	1	27.5	2	Definite	0	1	0	0
48	1	36.9	2	multiple	0	1	1	1
48	1	26.7	1	multiple	0	0	0	1
69	0	36.6	1	multiple	0	1	0	1
55	1	26.7	3	Definite	0	0	0	1

Snippet of the code for NB

```
nbmodel<- NaiveBayes(radiographic.grade~., data = OA)
score<- nbprediction$posterior[, c("large")]
actual_class<- TEST$radiographic.grade == 'large'
pred<- prediction(score, actual_class)
nbperf<- performance(pred, "tpr", "fpr")
nbauc<- performance(pred, "auc")
nbperf
nbauc<- unlist(slot(nbauc, "y.values"))
plot(nbperf, colorize=TRUE)
abline(0,1,lty=8,col="grey")
nbauc
```

Though there are many theoretical applications of analytics onhealthcare are available, there isn't any proper practical approach to it. Considering large increase in data from time to time makes it a point to deal with this situation practically. Once such initiative and the process adopted for OA risk prediction is explained below.

With the identified problem definition, we followed setting up stages for processing one source of output becoming input for another source. In the first stage, identification of predictor variables that influences OA condition as described in the overview section was finalized. In the second stage, the challenge was to build data marts such that we can do Confirmatory Data Analysis (CDA) on the longitudinal records. Thus we are providing an access layer to get the healthcare data which are of interest to the problem definition in the second stage. The output data marts require cleansing to apply analytical algorithms. In stage three, the objective was to help fixing data inconsistencies by transformation of data, de-duplication, standardization and filtering the data in a final format conducive enough to apply analytical algorithms. In stage four, R-Hadoop programming was used to do analytics on the data. Integrating R with Hadoop called RHadoop leverages parallel processing and distributed file system for storage of voluminous data. Since the data we have received was limited in size, we ran the experiments in R in a single thread processing.

The records containing 'Radio-graphic Grade' less than 2 were removed as it states the condition of a patient with 'No Osteoarthritis'. With Osteoarthritis records, the prediction of a patient's risk of obtaining Osteoarthritis was found. There were two cases in which the prediction was found, one was 'Symptomatic Knee OA' and another was 'Incident Knee OA'. Incident Knee OA is a prediction without the 'Knee Pain for a

Month' attribute. Two of the Machine Learning Algorithms 'Logistic Regression' and 'Naive Bayes' were used on the data. Validation of both the algorithm run was also performed using Receiver Operating Characteristic. This is also tested on other Osteoarthritis data from a different source to see how it performs. The comparisons of performance of each of the fitted models were made and the highest accurate algorithm can be implemented in hospitals to find the risk of a patient acquiring Osteoarthritis.

4. RESULTS AND DISCUSSIONS

Coefficients :

	Estimate	Std. Error	z value
(Intercept)	1.075991	0.442728	2.430
age	0.003183	0.005282	0.603
gender (M)	–0.054996	0.100104	–0.549
bm1	0.002698	0.008331	0.324
occupational. risk(Never)	–0.481273	0.220246	–2.185
occupational. risk(Often)	–0.188361	0.126572	–1.488
occupational. risk(Sometimes)	–0.254996	0.148739	–1.714
occupational. risk(Very Rarely)	–0.275207	0.168716	–1.631
family.history(First relative with OA)	0.013431	0.109780	0.122
knee. injury(Not Injured)	–0.148345	0.097929	–1.515
sports. activity (played)	0.047429	0.120204	0.395

– – –

Signif. codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 ' • ' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1

Null deviance : 2568.8 on 2287 degrees of freedom
Residual deviance : 2558.9 on 2277 degrees of freedom
AIC : 2580 : 2580.9

Number of fisher scoring iterations : 4

Coefficients :

	Estimate	Std. Error	z value
(Intercept)	1.067386	0.442843	2.410
age	0.003142	0.005282	0.595
gender (M)	–0.053544	0.100136	–0.535
bm1	0.002603	0.008335	0.312
occupational. risk(Never)	–0.487521	0.220455	–2.211
occupational. risk(Often)	–0.192442	0.126699	–1.519
occupational. risk(Sometimes)	–0.261631	0.149037	–1.755
occupational. risk(Very Rarely)	–0.285037	0.169281	–1.684
family.history(First relative with OA)	0.012343	0.109799	0.112
knee. injury(Not Injured)	–0.147402	0.097940	–1.505
sports. activity (played)	0.046485	0.120221	0.387
knee. pain(pain for a month)	0.084528	0.118799	0.712

– – –

Signif. codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 ' • ' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1

Null deviance : 2568.8 on 2287 degrees of freedom
Residual deviance : 2558.4 on 2276 degrees of freedom
AIC : 2582.4

Number of fisher scoring iterations : 4

Figure 2: Summary Values for Symptomatic Knee OA using LR

The predicting capability of LR and NB can be seen where LR says about the probability of risk in Figure 2 & 3 and NB says about the class of radiographic grade and the probability for each of the classes to acquire this condition in Figure 4. The results for Symptomatic Knee OA data are tested.

By using 'glm' package in R, the estimated value of a patient prone to getting Osteoarthritis was found for both condition symptomatic and asymptomatic (includes knee pain for a month). The snippet of the sample code is in section c. The variables are converted to factors to make it fit for the model after which LR is applied. Once a summary of this is commanded, it returns with the following values of the snippet. It explains as follows for symptomatic and asymptomatic conditions.

1. A person has a chance of increase in symptomatic OA as the age increases by 0.003183 and 0.003142 for asymptomatic conditions.
2. Next when a person's gender is male, he has 0.054996 and 0.053544 less probability for acquiring symptomatic and asymptomatic conditions respectively.
3. The occupational risk of working 'Always' is compared with never, often, sometimes and very rarely. So for a person who 'Never' works when compared with a person who 'Always' works, that person has 0.481273 less probability of acquiring symptomatic OA. It follows for other factors too.

Then if the diagnosing patient has family history of OA, then that person has 0.013431 probability of acquiring symptomatic OA, it decreases by 0.148345 when there was no injury in knee to that person and finally if that patient has history of sports activity then there is 0.047429 probability of acquiring symptomatic OA. The same applies to asymptomatic OA as below the summary of symptomatic OA. The LR applied model is tested on sample records and each records show the probability values of symptomatic OA.

1	1	3	4	5	6	7	8	9
0.7578326	0.7653084	0.7487994	0.7803385	0.7801818	0.7918721	0.7791627	0.7816197	0.7619206
10	11	12	13	14	15	16	17	18
0.7881650	0.7597925	0.7678225	0.7795726	0.7818169	0.7795912	0.7878372	0.7636607	0.7589418
19	20	21	22	23	24	25	26	27
0.7455486	0.8016123	0.7786255	0.7818293	0.7528626	0.7788797	0.8087071	0.7650346	0.7814859
28	29	30	31	32	33	34	35	36
0.7906591	0.7934905	0.7929455	0.7671683	0.7477274	0.7929726	0.7827218	0.7462038	0.7437460
37	38	39	40	41	42	43	44	45
0.7783628	0.7287463	0.7561314	0.7929676	0.7455022	0.7491944	0.7606505	0.7544953	0.7839124
46	47	48	49	50	51	52	53	54
0.7028831	0.7578158	0.7028778	0.7834712	0.7465509	0.7214631	0.7037210	0.7109865	0.7986154
55	56	57	58	59	60	61	62	63
0.7790448	0.7328556	0.7428524	0.7284118	0.7551437	0.7484362	0.7629836	0.6971472	0.8016658
64	65	66	67	68	69	70	71	72
0.7306583	0.7328679	0.6947908	0.7208874	0.7515202	0.7151057	0.7636899	0.7494424	0.7168544
73	74	75	76	77	78	79	80	81
0.7519295	0.7052253	0.7642239	0.7726012	0.7446139	0.7940518	0.7165318	0.7152738	0.7391438
82	83	84	85	86	87	88	89	90
0.7029631	0.7487540	0.7850371	0.7143024	0.8102153	0.7782458	0.7875952	0.7904520	0.7808983
91	92	93	94	95	96	97	98	99
0.7711138	0.6924197	0.7102128	0.7578592	0.7818901	0.7535912	0.7075176	0.7521486	0.7232474
100	101	102	103	104	105	106	107	108
0.7809809	0.7404244	0.7403378	0.6983444	0.7276169	0.7666716	0.7526281	0.6471755	0.6908028
109	110	111	112	113	114	115	116	117
0.7181391	0.7614021	0.7000759	0.6770366	0.7380490	0.6788658	0.7038504	0.6973042	0.6683457
118	119	120	121	122	123	124	125	126
0.6722272	0.6955192	0.6580794	0.7214164	0.7120118	0.6902949	0.6950083	0.6985476	0.7273144

Figure 3: Probability Values for Symptomatic Knee OA using LR

Estimation of the patient risk of Osteoarthritis was made by the use of 'klaR' library and with 'MASS' package. NB deals prediction based on classes and so the independent variable has to be classified when a prediction is made, the summary returns the class level of symptomatic OA to a person as large, multiple or definite. The snippet in Figure 4 shows the probable class in which the patient is going to come under as large, medium or definite for each patient to which they are likely to come under for a certain class of symptomatic OA. NB trained model was created and was tested on test data and the results show the prediction values. Thus on prediction, records in the test data show the classes in which they are likely to come under in '$class' part whereas in '$posterior' part it says the probability value for each record in each class. The probability values obtained for each record in '$posterior' part amounts to 1 when the class of probability values are summed up. This prediction done is for symptomatic OA as the trained model and does not include 'pain for a month' attribute.

$class

[1]	large	large	large	large	large	large	large
[18]	large	large	large	large	large	multiple	large
[35]	multiple	multiple	large	multiple	multiple	multiple	multiple
[52]	multiple	definite	large	large	multiple	multiple	multiple
[69]	multiple	large	large	multiple	large	multiple	large
[86]	multiple	large	large	large	large	large	multiple
[103]	multiple	multiple	multiple	large	multiple	multiple	multiple
[120]	definite	large	multiple	multiple	multiple	multiple	multiple
[137]	multiple	multiple	multiple	multiple	multiple	multiple	multiple
[154]	large	multiple	multiple	multiple	definite	multiple	multiple
[171]	multiple	multiple	definite	multiple	multiple	multiple	multiple
[188]	large	large	multiple	multiple	multiple	large	multiple
[205]	large	large	multiple	multiple	definite	definite	large
[222]	multiple	multiple	multiple	large	multiple	multiple	multiple
[239]	large	large	definite	large	large	multiple	large
[256]	multiple	multiple	multiple	multiple	multiple	multiple	multiple
[273]	multiple	multiple	multiple	multiple	multiple	multiple	multiple
[290]	multiple	multiple	multiple	large	large	multiple	multiple
[307]	definite	definite	multiple	multiple	definite	multiple	definite
[324]	multiple	multiple	large	multiple	multiple	multiple	multiple
[341]	multiple	large	multiple	multiple	large	definite	definite

Levels : definite large multiple

$posterior

	definite	large	multiple
[1,]	0.2402532	0.710186741	0.04956005
[2,]	0.3055239	0.592828770	0.10164735
[3,]	0.3237344	0.474264042	0.20200156
[4,]	0.2884739	0.474747830	0.23677831
[5,]	0.2893529	0.476184861	0.23446220
[6,]	0.2166647	0.636698697	0.14663662
[7,]	0.2402632	0.698392122	0.06134470
[8,]	0.3000529	0.413351482	0.28659566
[9,]	0.2978331	0.506328427	0.19583844
[10,]	0.1354424	0.840016898	0.02454075
[11,]	0.2701620	0.664285686	0.06555230
[12,]	0.3107679	0.441685284	0.24754679
[13,]	0.2894490	0.473951054	0.23659991

Figure 4: Class of radiographic grade and Probability for each of the Classes

To know the accuracy of a model, Receiver Operating Characteristic (ROC) and the Area Under the Curve (AUC) are found [7]. This follows such for symptomatic OA to both LR model and NB model. The curve has to be in the true positive side of a line of 45 degrees. This is because something is true and the prediction is true, only on this case (both true) a system can be trusted. Both LR and NB yielded curves in the true positive side with AUC values of 0.5908963 and 0.5541114 respectively. A system is decently good when it produces such values; however the system with the LR model for OA prediction performs better than a system with NB. An analogy of the predicting performance can be seen for LR in Figure 5 and NB in Figure 6 where the predicting performance is tested through ROC and by finding the AUC. This is a kind of test done on the models that are fitted to see, how well it performs.

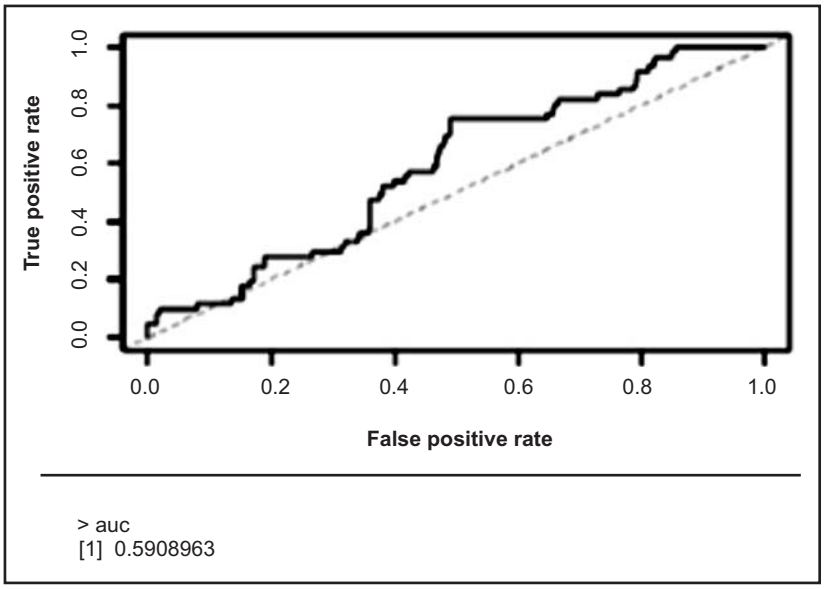

Figure 5: Risk Prediction for Symptomatic Knee OA using LR

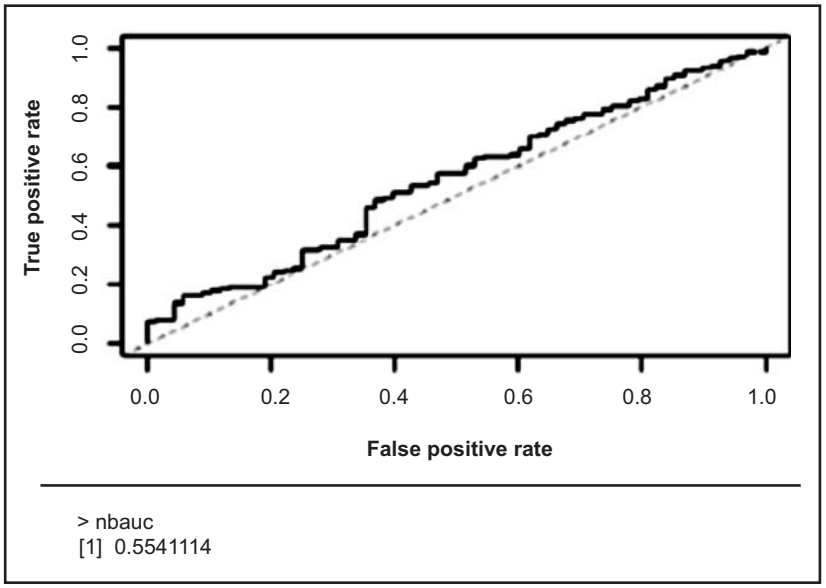

Figure 6: Risk Prediction for Symptomatic Knee OA using NB

5. CONCLUSION

The algorithm has shown the capability to predict the risk of Osteoarthritis to a patient using both LR and NB. An analogy made for LR and NB shows that the Logistic Regression model fitted gives better predicting capability than Naive Bayes. This is found by getting to know the values of the Area under the Curve, which are 0.5908963 and 0.5541114 for LR and NB respectively. Different Algorithms can be fitted and its model fitting can be tested as a future work. As for the results obtained here, Logistic Regression model is suggested for Osteoarthritis Risk Prediction.

REFERENCES

[1] Zhang, W., Mcwilliams, D. F., Ingham, S. L., Doherty, S. A., Muthuri, S., Muir, K. R., & Doherty, M. "Nottingham knee osteoarthritis risk prediction models," *Annals of the Rheumatic Diseases*, vol 70(9), pp 1599-1604.

[2] Hlaudi Daniel Masethe, Mosima Anna Masethe. "Prediction of Heart Disease using Classification Algorithms," *World Congress on Engineering & Computer Science (WCECS).* vol II. pp. 22-24, Oct. 2014.

[3] Zhang, W. "Risk factors of knee osteoarthritis – excellent evidence but little has been done," *Osteoarthritis and Cartilage*. vol 18(1). pp. 1-2. 2009.

[4] Cui, J. "Overview of Risk Prediction Models in Cardiovascular Disease Research," *Annals of Epidemiology*. vol 19(10). 2009 .

[5] Sutlive, T. G., Lopez, H. P., Schnitker, D. E., Yawn, S. E., Halle, R. J., Mansfield, L. T., Childs, J. D. "Development of a Clinical Prediction Rule for Diagnosing Hip Osteoarthritis in Individuals With Unilateral Hip Pain," *The Journal of Orthopaedic & Sports Physical Therapy*, vol 38(9). pp. 542-550.

[6] (2016) Predictive Modeling, [Online], Available: www.datamine.com/Resources/whitepapers/Predictive+Modeling.html

[7] Kerkhof, H. J., Bierma-Zeinstra, S. M., Arden, N. K., Metrustry, S., Castano-Betancourt, M., Hart, D. J., Meurs, J. B. "Prediction model for knee osteoarthritis incidence, including clinical, genetic and biochemical risk factors," *Annals of the Rheumatic Diseases Ann Rheum Dis*. vol 73(12). pp. 2116-2121, 2013.

[8] Reijman, M., Hazes, J. M., Bierma-Zeinstra, S. M., Koes, B. W., Christgau, S., Christiansen, C., Pols, H. A. "A new marker for osteoarthritis: Cross-sectional and longitudinal approach," *Arthritis & Rheumatism*. vol 50(8). pp. 2471-2478. 2004.

[9] (2016) Seven ways predictive analytics can improve healthcare, [Online], Available: http://www.elsevier.com/connect/seven-ways-predictive-analytics-can-improve-healthcare.

[10] Shyamala K., Vijay Kumar K., "Distributed Text Mining – An Approach To Sentiment Analysis,". *International Journal of Applied Engineering Research*. vol 10. pp. 14319-14335. 2015.

[11] Dong, Guozhu, and Vahid Taslimitehrani. "Pattern-aided Regression Modeling and Prediction Model Analysis." *IEEE 32nd International Conference on Data Engineering (ICDE)*. 2016.

[12] Ardern, Christopher I., Peter T. Katzmarzyk, Ian Janssen, and Robert Ross. "Discrimination of Health Risk by Combined Body Mass Index and Waist Circumference." *Obesity Research*. vol 11(1). pp. 135-42. 2003.