Computer Vison And Machine Learning Assignment

Abhijay Hazarika
*dept. Electrical and Computer Engineering;*Aarhus University*(AU)*
Aarhus, Denmark

***Abstract*—This paper develops an AI prediction model to be used as a clinical decision model for determining patient specific risk factors for knee arthroplasty survival and complications based on patient demographics, lifestyle factors, radiographic evaluation, bone biomarkers, bone mineral density and radio stereometry measures. To accomplish this task different machine learning algorithms had been implemented for the classification problem and optimal algorithm was chosen. It includes four classification schemes: Logistic Regression, k-Nearest Neighbor, Random Forest Classifier and Support Vector Machine Classifier. These schemes are used on the imputed data from the Training Data set provide.**

Keywords—Machine Learning, Logistic Regression, Nearest Neighbor Classifier, Random Forest, Support Vector Machines

I. INTRODUCTION

Total Knee Arthroplasty is an effective treatment of knee osteoarthritis which is one of the most common degenerative joint diseases. Once the implant has been done in the surgery to meet the necessary requirement there are various factors which can impact the longevity of the surgery. To successfully predict the outcome of a surgery we have been given health data from over 400 patients. The data has been divided into three sets of test data, train data and validation data. The model was trained on the training data to predict the outcome column in the test data set to figure out the accurate prediction for the MiG_Group.

The training data set had some missing values so the missing values have been filled by utilizing an Machine Learning Algorithm knows as knn-imputer. This helped to fill the missing values in the given dataset and pre-process the data to fit into Standard Scalar format. From here on different Machine Learning algorithms have been implemented on the data to generate a predictive model to predict the MIG_Group. In the dataset the MIG_Group returns a true value of 1 in case the migration of the implant is greater than 0.2 mm between 1 and 2 years after surgery. The scope of this paper is correctly predict if the patient would be in need of surgery on the basis of classification of the MIG_Group. The trained model on the test data was then used to predict the MIG_Group of the test dataset.

To solve this classification problem four different classification algorithms have been utilized to predict the MIG_Group. Out of the four classification models, Logistic Regression was chosen and further tuned by finding out the hyperparameters. The resultant model was then used to predict the outcome of the test data set. The scope of this paper have been limited to working with the imbalanced data, and accuracy have been used as a metric to measure the value of the prediction. Imbalanced data refers to those types of datasets where the target class has an uneven distribution of observations, i.e. one class label has a very high number of observations and the other has a very low number of observations. In our dataset there is an uneven distribution of the MIG_Group data where the majority data is classified as 0 and minority class has been classified as 1.

This paper also discusses the different methodologies which can be implemented or used to deal with the imbalanced data such as increasing the count of the minority class for a better judgement. But the paper is limited to the theoretical discussions about the different models or technologies which can be implemented or worked on further.

II. DESCRIPTION OF THE MACHINE LEARNING ALGORITHMS

A. K-Nearest Neighbour

The k-nearest neighbour classifier is a non-parametric algorithm which assigns the class to a testing signal by analysing the number k of the nearest reference signals in the feature space. This classifier requires the evaluation of distances between the testing signal and the reference signal. Given a set of features F = {F1, F2...F_L} where L is the number of features the Euclidean distance between the feature set of signals A and B is calculated by the following equation

$$D(F(A), F(B)) = \sqrt{\sum_{l=1}^{L} [F_l(A) - F_l(B)]^2}$$

KNN Classifier Pseudo Code	
Input	*M* reference signals from every candidate modulation *m*;(i), i = 1, 2, …. I, each with a set of extracted feature sets $F(m)$, an observed unknown signal with extracted feature set F, and a pre-defined k value
Step 1	The distance between F and every reference feature is calculated using the above equation
Step 2	The resulting distances D (F, F^i(m)) are sorted in descending order.
Step 3	The first k distances are selected
Step 4	The modulation label *i* for each distances D (F, F^i(m)) is extracted.
Step 5	The mode of the set of extracted labels is set in *i'* is used to identify the modulation.
Step 6	Modulation *m;(i')* is returned as the classification decision *m'*
Output	*m'*

B. Logistic Regression

Logistic regression is a typical binary classification algorithm in machine learning theory. The idea comes from linear regression. Both logistic regression and linear regression belong to generalized linear model. Under the condition of fixed model parameters, for the given independent variables, the values of the model are subject to exponential cluster distribution, linear regression value is subject to normal distribution, and logistic regression value is subject to Bernoulli distribution. The main idea is to add a layer of nonlinear function between the characteristics and results of linear regression, namely, the sigmoid function. By smoothing and non-linear processing the linear regression value, the probability value of regression classification is obtained, so as to classify the nonlinear discrete data.

For a given data(X,Y), where X is a matrix with 'm' examples and 'n' features and Y is a vector with 'm' examples. The objective is to train the model to predict which class the future value belongs to. [1]

Logistic Regression	
Step 1	A weight matrix is created with random initializations $a = w_0 + w_1x_1 + \ldots . w_nx_n$
Step 1	The output is then passed on to a link function $\hat{yi} = {}^{1}/_{(1+\ e^{-a})}$
Step 3	Then the cost function is calculated $cost(w) = \left(-\frac{1}{m}\right)\sum_{i=1}^{i=m} y_i \log(\hat{y_i}) + (1 - y_i)\log(1 - \widehat{y_i})$
Step 3	Then the derivative of this cost function is calculated. $dw_j = \sum_{i=1}^{i=n} (\hat{y} - y){x_j}^i$
Step 4	The weights are then updated accordingly $w_i = w_j - (\alpha *) dw_j$
Step 6	Modulation *m;(i')* is returned as the classification decision *m'*

C. Random Forest

Random Forest falls under the category of supervised learning algorithms. It utilizes ensemble learning technique which means multiple algorithms can be used at a time or a single algorithm can be used multiple times to make the model more robust and powerful. Random Forest can be used for both classification and regression problems. Random Forest Classifier is a set of decision trees from randomly selected subset of training set. [2]

Random Forest Classification	
Step 1	The model starts by selecting a random sample from a given dataset
Step 2	Then this algorithm will construct a decision tree for every sample. From there it will then get the prediction result from every decision tree.
Step 3	In this step, voting will be performed from every decision tree.
Step 4	At the end the most voted prediction results as the final prediction result.

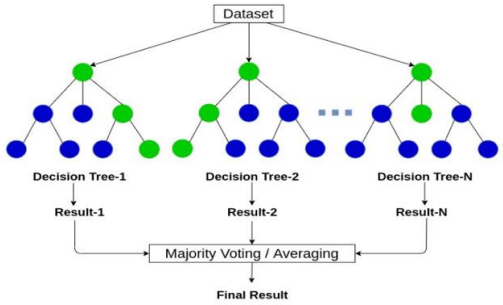

Fig. 1. Strucutre of Random Forest

D. SVM Classifier

A Support Vector Machine (SVM) is a supervised classification technique and the essence of this algorithm simply involves in finding a boundary that separates classes from each other. The SVM helps create a boundary that maximizes the boundary between two classes. The SVM can be considered as an optimization problem where the goal is to minimize the following quantity.

$$MIN_{a_0,\ldots a_m} : \sum_{j=1}^{n} MAX\left\{0, 1 - \left(\sum_{i=1}^{m} a_i x_{ij} + a_0\right) y_j\right\} + \lambda \sum_{i=1}^{m} ({a_i}^2)$$

The first part of this equation focuses on minimizing the error, which is the number of falsely classified points that the SVM makes. The second part of this equation focuses on maximizing the margin. The distance between the two lines from the two classes can be represented by this equation below

$$Distance = \frac{\left|a_{0_{upper\ boundary}} - a_{0_{lower\ boundary}}\right|}{\sqrt{\sum_{i=1}^{m} ({a_i}^2)}}$$

The main objective of the Support Vector Machine is to minimize the total error and maximize the margin by minimizing a_i. [3]

III. Preparation Of Data

From the training data set and the test data set it could be seen that there is a case of missing values. Since this is a real-

Identify applicable funding agency here. If none, delete this text box.

life data set, this problem is bound to arise. Missing values can bias the results of the machine learning models or reduce the accuracy of the prediction from the models. Missing data in the case of this dataset can be defined as the values of data which were represented by NaN. The reasons for these missing values can be from any arbitrary reason such as:

- Data got corrupted due to improper maintenance.
- Observations could not be recoded from the field due to patients refusing to provide information.

Missing data can be classified into three different categories.

A. Missing Completely at Random

In this category the probability of data being missing is same for all the observations. There is no relationship between the missing data and any other observed or unobserved values within our given dataset. The missing values are completely independent of the other data and no pattern could be detected.

B. Missing at Random

In this category the cause of missing values can be explained by variables on which the complete information is missing and there is a relationship which could be identified between the classes and the missing values. The probability of data being missing depends only on the observed data.

C. Missing Not at Random

In this category the missing values depend on the unobserved data. If there is a case of some pattern or structure of the data being missing and the observed data cannot be used to explain. This case generally happens due to the reluctance of different people providing information like answering the questions in the survey which led to this creation of this data set.

Even with the presence of different machine learning algorithms like K-nearest neighbor and Naïve Bayes classifier which support the missing values it is better to impute the missing data in the dataset. There are different ways to impute missing values but before imputing the values it's better to figure out how many classes have missing values and the count in the respective classes. The following figure shows the count of the missing values in the training and test dataset.

Number	0	Number	0
Sex	0	Sex	0
Age	0	Age	0
Tscore	1	Tscore	0
Height	3	Height	0
Weight	3	Weight	0
BMI	3	BMI	0
Cem_ucem	0	Cem_ucem	0
TKA	0	TKA	0
side	0	side	0
BASP	7	BASP	0
PTHpmoll	1	PTHpmoll	0
Vitdnmoll	1	Vitdnmoll	0
CTXmygl	7	CTXmygl	1
P1NP	7	P1NP	1
calciumion	1	calciumion	0
creatinin	0	creatinin	0
eGFR	0	eGFR	0
oks_total	9	oks_total	0
VAS_aktiv	9	VAS_aktiv	0
exercise	9	exercise	0
former_alcoholabuse	9	former_alcoholabuse	0
smoker	9	smoker	0
former_smoker	9	former_smoker	0
MIG_group	0		

Fig. 2. Number of missing values belonging to each class from the train dataset and the test data set repectively.

There are two primary ways for tackling the missing values problem, which is to either delete the missing values or impute the missing values. Even though the classic way is to impute the missing values with either mean, median or mode, but in the scope of this paper the missing values have been imputed with the algorithm knn-imputer from Sklearn library.

The knn-imputer is one of the most widely used and common method to impute missing values. The reason behind this is that the best way to impute missing data is with an estimated value based on the missing data. Univariate methods used for missing value imputation are simple ways of estimating the values but may not be able to provide an accurate model. The knn-imputer works on the principle of k-Nearest Neighbor that identifies the neighboring points through a measure of distance and the missing values can be estimated using completed values of neighboring observations.

The main idea in the kNN method is to identify 'k' samples in the dataset which are similar or are in close space then utilize the 'k' samples to estimate the value of the missing data points. Each sample's missing values are imputed using the mean value of the 'k'-neighbors found in the dataset. The Euclidian distance can be calculated as follows:

$$d_{xy} = \sqrt{weight * squared\ distance\ from\ coordinates}$$

where,

$$weight = \frac{Total\ number\ of\ coordinates}{Number\ of\ present\ coordinates}$$

After utilizing this algorithm, it could be seen that the missing values problem is resolved from the following figure below. [4]

Number	0	Number	0
Sex	0	Sex	0
Age	0	Age	0
Tscore	0	Weight	0
Height	0	Cem_ucem	0
Weight	0	TKA	0
BMI	0	side	0
Cem_ucem	0	creatinin	0
TKA	0	eGFR	0
side	0	MIG_group	0
BASP	0	Tscore	0
PTHpmoll	0	Height	0
Vitdnmoll	0	BMI	0
calciumion	0	BASP	0
creatinin	0	PTHpmoll	0
eGFR	0	Vitdnmoll	0
oks_total	0	CTXmygl	0
VAS_aktiv	0	P1NP	0
exercise	0	calciumion	0
former_alcoholabuse	0	oks_total	0
smoker	0	VAS_aktiv	0
former_smoker	0	exercise	0
CTXmygl	0	former_alcoholabuse	0
P1NP	0	smoker	0
		former_smoker	0

Fig. 3. Number of missing values belonging to each class from the test dataset and the train data set repectively.

IV. Data Visualization

It can be seen from the plot of the training data that this is an imbalanced data set. As the Outcome Column Distribution shows that the count of number of people not requiring surgery very high than that of people requiring surgery.

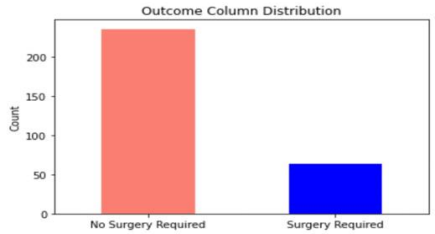

Fig. 4. Distribution of MIG_Group data in the training dataset.

The visualization of different calsses and the distubution of the different values in the calsses had been done as well using a correlation matrix. A correlation matrix is a tabular data representing the 'correlations' between pairs of variables in a given data. The rows and columns represents a variable, and every value in the given matrix is the correlation coefficient btween the variables represented by the corresponding row and column. The correlation matrix is an unique data analysis metric which is computed to summarize data to understand relationship between various varibales.

The correlation matrix responds to each row and variables by using the correlation coefficient. A correlation coefficient denotes the strength of the relationship between two variables. It is defined as the covariance between two variables divided by the product of the standard deviations of the two variables.

$$\rho(X,Y) = \frac{COV(X,Y)}{\sigma_X \sigma_Y}$$

Where the covariance between X and Y *COV(X,Y)* can be defined as the 'expected value of the product of the deviations of X and Y from their respective means. [5]

$$COV(X,Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

To represent and visualize the data even more clearly Scatter Matrix had been used as a plot. A scatter matrix compactly plots all the numeric variables present in a dataset against one another. In this case the pair-plot shows relationships of each feature with each other in the training dataset. Seaborn plot have been used to represent the data from the training dataset.

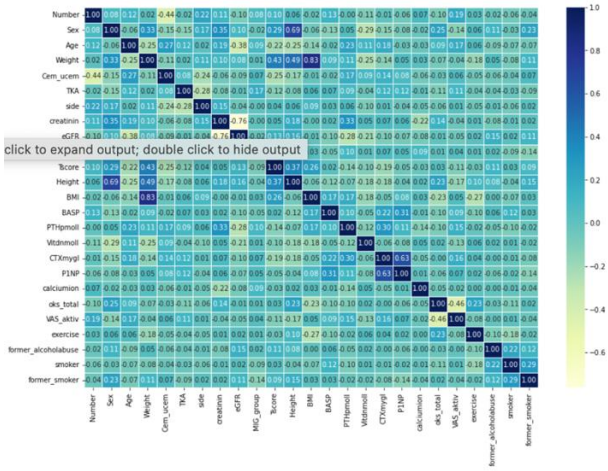

Fig. 5. This figure represents the covariance matrix of the training dataset.

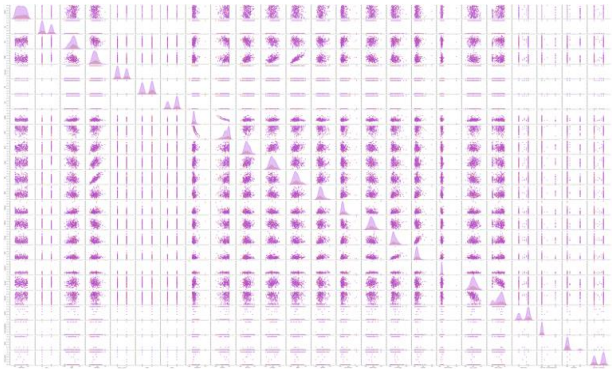

Fig. 6. This figure represents the scatter plot of the data in the training set.

V. Model Selection and Hyperparameter Tuning

The data had been trained on four different machine learning models to give out the prediction results. Since the most common method for binary classification is Logistic Regression, here this algorithm has been implemented to train the model. The initial accuracy results from applying the machine learning algorithm via utilizing the machine learning

model were 78 percent but the results have been improved by utilizing the grid search method to hyper tune the parameters.

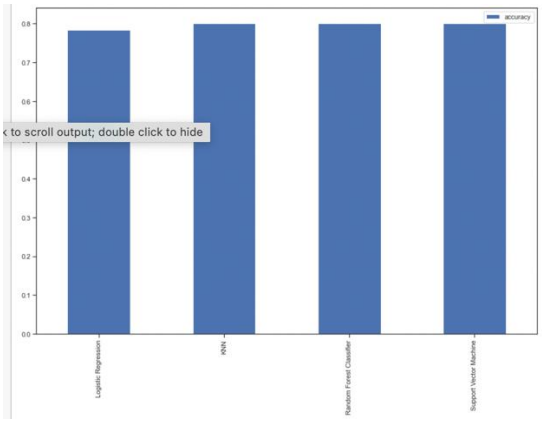

Fig. 7. This figure represents the accuracy from four different machine learning models.

Hyperparameters are the variables that the user specifies while building the model. To find the best hyperparameter values which would fit the goal of out model Grid Search Method had been used. Grid Search uses a different combination of all the specified hyperparameters and their values and calculates the performance for each combination and selects the best value for the hyperparameters. This makes the processing time-consuming and expensive based on the number of hyperparameters involved.

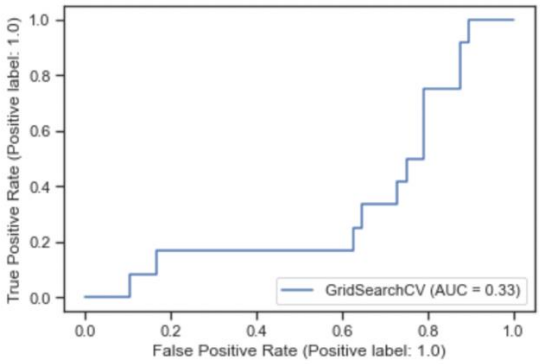

Fig. 8. This figure displays the ROC curve

In GridSearchCV, along with Grid Search, cross-validation is also performed. Cross-Validation is used while training the model. Before training the model with data, the data was divided into two parts – train data and test data. In cross-validation, the process divides the train data further into two parts – the train data and the validation data. Then utilizes the best parameters to further tune the model. [6]

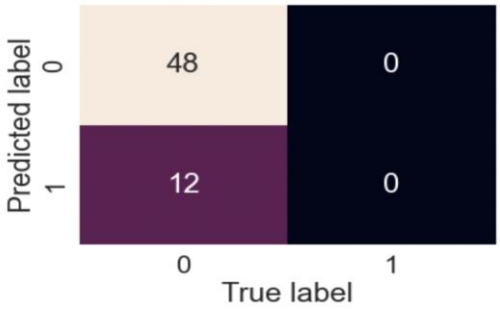

Fig. 9. This figure displays the classification results in a confusion matrix

VI. Discussion

The choice to using kNN imputer to solve the missing value problem can be considered as a point of discussion as the value of 'k' can be changed accordingly. It would be suggested to test the model using cross-validation after performing imputation with different values of 'k'. After cleaning up the data it could be seen that the dataset contains imbalanced data. There is different matrix to measure the prediction such as F1 score instead of accuracy as the performance metric. Various resampling techniques could also be used such as oversampling more copies of the minority class. In this case the technique of under sampling the majority class could not be considered as the quantity of data was very low. [7] This problem could also be tackled by generating synthetic samples to have more of the minority class. One such way to do that would be to utilize imblearn's Synthetic Minority Oversampling Technique. These were some of the discussion points about the various techniques which could have been utilized to tackle the problem of imbalanced data.

VII. Conclusion

This paper discussed about the different machine learning models which could be applied in the context of the given problem of classifying data as a predictive model for Total Knee Arthroplasty surgery sustainability. In the given dataset kNN can be considered as an effective and standard methodology to impute missing data. Logistic regression and Support Vector Machine can be considered as good machine learning models to implement in binary classification problems.

VIII. Works Cited

[1] S. Sekhar, "Math Behind Logistic Regression," Medium, 2019 August 2019. [Online]. Available: https://medium.com/analytics-vidhya/logistic-regression-b35d2801a29c. [Accessed 27 April 2022].

[2] R. RASTOGI, "Random Forest Classification and it's Mathematical Implementation," Medium, 16 June 2020. [Online]. Available: https://medium.com/analytics-vidhya/random-forest-classification-and-its-mathematical-implementation-1895a7bb743e. [Accessed 27 April 2022].

[3] MLMath.io, "Math behind SVM (Support Vector Machine)," Medium, 10 February 2019. [Online]. Available: https://ankitnitjsr13.medium.com/math-behind-support-vector-machine-svm-5e7376d0ee4d. [Accessed 3 May 2022].

[4] K. S. Htoon, "A Guide To KNN Imputation," Medium, 3 July 2020. [Online]. Available: https://medium.com/@kyawsawhtoon/a-guide-to-knn-imputation-95e2dc496e. [Accessed 15 April 2022].

[5] M. Lanhenke, "Understanding the Covariance Matrix," Towards Data Science , 29 December 2021. [Online]. Available: https://towardsdatascience.com/understanding-the-covariance-matrix-92076554ea44. [Accessed 7 April 2022].

[6] R. Shah, "Tune Hyperparameters with GridSearchCV," Analytics Vidhya, 23 June 2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/06/tune-hyperparameters-with-gridsearchcv/. [Accessed 21 April 2022].

[7] T. Boyle, "Dealing with Imbalanced Data," Towards Data Science, 3 February 2019. [Online]. Available: https://towardsdatascience.com/methods-for-dealing-with-imbalanced-data-5b761be45a18. [Accessed 4 May 2022].