International Journal of Scientific Research in Science, Engineering and Technology

Print ISSN - 2395-1990
Online ISSN : 2394-4099

OPEN ACCESS

Available Online at : www.ijsrset.com
doi : https://doi.org/10.32628/IJSRSET2310540

Prognosis of Hyper Triglycerides Using Data Science and Machine Learning

S. Alagu Thangam, S. Vilma Veronica, G. Suresh, S. Hemalatha

Department of Computer Science and Engineering, Kings Engineering College, Chennai, Tamilnadu, India

ARTICLEINFO

Article History :

Accepted: 01 Oct 2023
Published: 11 Oct 2023

Publication Issue :
Volume 10, Issue 5
September-October-2023
Page Number :
229-239

ABSTRACT

Triglycerides are a type of fat. They are the most common type of fat in our body. They emanate from foods, especially butter, oils, and other fats we eat and also come from extra calories. These are the calories that we eat, but our body does not need right away. Our body naturalizes these extra calories into triglycerides and stores them in fat cells. When our body needs energy, it disseminates the triglycerides. Our VLDL cholesterol particles carry the triglycerides to our tissues. Hyper triglycerides can increase the risk of heart diseases in particular, CAD, stroke, liver, kidney and other chronic diseases. In recent years Data science is one of the progressing demense due to the profusion of data sources and resulting data. The realm of healthcare is substantially ameliorated from Data science and Machine Learning applications because of these intuitive solutions. Using Data science techniques and Machine learning algorithms with ANN, we can prognoses the disease. The WHF dossier says that every year nearly 4.4 million death occurs due to heart diseases and WHO says that nearly 2.6 million deaths occur due to cholesterol.

Keywords : Triglycerides, Disseminates, Chronic Diseases, Prognoses, Ameliorated, Dossier.

I. INTRODUCTION

Our blood contains a particular form of lipid called triglycerides. When we eat, our body turns any calories into triglycerides that it won't utilize straight away. The triglycerides are kept in our fat cells. These Triglycerides are later released by hormones to provide energy between meals. We may have excessive triglycerides (hyper triglyceridemia) if we consistently consume more calories than we burn, especially from diets heavy in carbohydrates.

High triglyceride levels sometimes occur together with other health issues such as high blood pressure, diabetes, obesity, high "bad" LDL cholesterol, and low "good" HDL cholesterol. As long as it is functioning at a healthy level, it plays an important part in the human body's ability to produce new cells. When cholesterol

levels are too high, it has the opposite effect and seriously compromises heart health.

A high cholesterol level (hypercholesterolemia) causes the blood arteries to become clogged and the fats, making it more difficult for blood to flow through the arteries. The heart doesn't get the oxygen it need, and the chance of having a heart attack rises.

Hypertriglycerides can raise the chance of developing several chronic diseases, including liver, renal, and heart disease, as well as CAD, stroke, and other vascular problems.

Due to the abundance of data sources and consequent data in recent years, data science is one of the fields that are progressing. And Machine learning (ML) has also piqued the interest of the medical professionals because of its essential capacities in health-related situations, including risk assessment, prognosis, and management of a variety of illnesses. This piece features a supervised machine learning methodology whose major goal is to produce extremely reliable risk prediction tools for the development of hypercholesterolemia. A data understanding analysis is specifically carried out to investigate the features' associations with and relevance to hypercholesterolemia. These variables are used to train and evaluate various ML models to determine which is the most effective for our needs.

II. LITERATURE REVIEW

Goran Walldiusa and Ingmar Jungner (2007) demonstrated to compare the potential of high-density lipoprotein (HDL) cholesterol and apolipoprotein (apo) A-I, the major protein in HDL particles, in predicting cardiovascular risk. Pros andcons for using these risk markers are discussed.

Jing Ma et al., (2017) analysed the relationship of triglyceride (TG) and cholesterol (TC) with indexes of liver function and kidney function, and to develop a prediction model of TG, TC in overweight people.

Sajida Perveen et al., (2018)presented this research is to develop machine learning based method in order to identify individuals at an increased risk of developing Non-Alcoholic Fatty Liver Disease using risk factors of Metabolic Syndrome validate the relative ability of quantitative score.

Nahuel García-D'urso et al., (2022) described the machine learning approach to predict cholesterol levels using non-invasive and easy-to-collect data is presented. Specically, it uses clinical and anthropometric data gathered by nutritionists during weight loss intervention (dieting) periods.

III. SYSTEM ARCHITECTURE

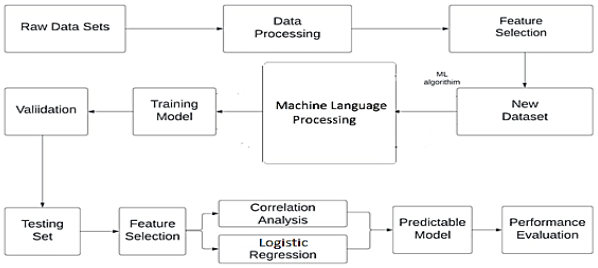

Fig. 3.1 Architecture Diagram of the model

3.1. Raw Data Sets:

Data that have been gathered from sources like Google form, kaggle and other health related data sets from different sources are known as raw data or primary data. To forecast the HTG and HC, this system's data was collected from a variety of sources to compile the Dataset from many sources. The first phase in the Data Science process is data collection for the ML model's training.

3.2. Data Preprocessing:

The classifier purges the data set of unnecessary information. Thus, a number of preprocessing stages are used to assure data validity, and data cleaning techniques are used to ensure data quality. Some

instances of data cleaning techniques involve getting rid of redundant data, preventing typos, handling missing values, data imputation, etc. We decided to eliminate outliers (i.e., occurrences with missing and invalid feature values) from the present data set.

3.3. Feature selection:

In machine learning, the process of feature selection identifies critical elements in a dataset to enhance the model's performance and interpretability. It is mainly for characteristic's significance to the target variable and to compute scores for each feature[12]. Based on their ratings, it chooses a subset of the most significant features and uses them to train the predictive model.

3.4. New Data Set:

The new data is relevant, reliable, and consistent with the existing data. All outliers, duplicates, and redundant data have been removed from the new data set. And thus the required or necessary variables like

3.5. Machine Learning Algorithm:

In this system, the model is trained using Supervised Machine learning algorithms[1]. Supervised learning is a method of machine learning in which the output is predicted by the machines using well-labelled training data that has been used to train the machines. The term "labelled data" refers to input data that has already been assigned the appropriate output.

In supervised learning, the training data that is given to the computers serves as the supervisor, instructing them on how to correctly predict the output. It employs the same approach that a pupil would learn under a teacher's guidance.

The method of supervised learning involves giving the machine learning model appropriate input data as well as the output data. Identifying a mapping function to link the input variable (x) with the output variable (y) is the goal of a supervised learning algorithm. It also encompasses applications in the real world such as risk assessment, image categorization, fraud detection, spam filtering, etc.

3.5.1 Random Forest Algorithm:

The supervised learning method includes the well-known machine learning algorithm Random Forest. It can be applied to ML Classification and Regression issues[2]. Its foundation is the idea of ensemble learning, which is the process of mixing various classifiers to solve a challenging problem and enhance the performance of the model.

Random Forest is a classifier that uses numerous decision trees based on different subsets of the provided dataset and aggregates the results to increase the dataset's predicted accuracy. Rather than depending on a single decision tree, the random forest uses forecasts from each tree and predicts the result based on the votes of the majority of predictions.

Larger the number of trees in the forest, higher the accuracy and overfitting are hence prevented.

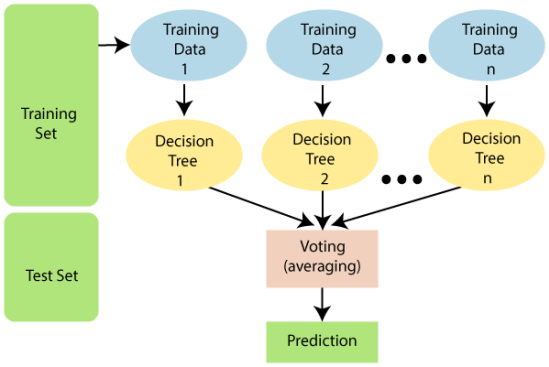

Fig.3.5.1 Random Forest Algorithm process

Considering the fact that the random forest aggregates several trees to estimate the dataset's class, it is possible that some decision trees will predict the correct result while others won't. But when evaluated together, each tree accurately predicts the outcome.

For the feature variable to predict real outcomes as opposed to hypothetical ones there must be some real values in the dataset. The correlations between predictions from each tree must be exceedingly low. Comparatively speaking, the Random Forest technique needs less training time. Additionally, it provides output predictions with high accuracy and runs well even with a huge dataset. Accuracy may still be preserved even though a sizable portion of the

data is missing. N decision trees are joined to create the random forest, and the predictions are then created for each of the first phase's trees individually.

3.5.2. KNN Algorithm:

Select K data points at random from the training set. After that, create the decision trees connected to the selected subsets of data. Next, decide what size N we want our decision trees to have. Repetition of the same action is required. Locate each decision tree's predictions and group them according to the group that receives the most votes when dealing with freshly acquired data points.

If there are two categories, Category A and Category B, and we have a new data point, x1, which category does this data point belong in? We require a K-NN algorithm to address this kind of issue. K-NN makes it simple to determine the category or class of a given dataset.

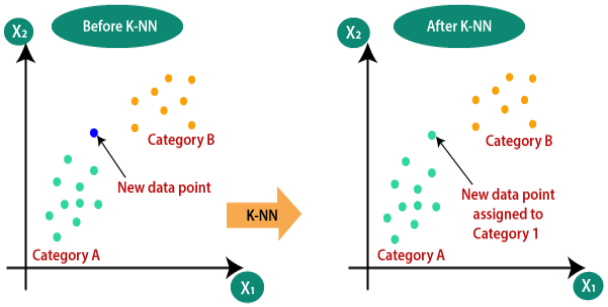

Fig.3.5.2.1 KNN Algorithm for assigning new data points

The K-NN algorithm chooses the Kth neighbor. The Euclidean distance between K neighbors is then calculated. Also, depending on the estimated Euclidean distance, choose the K closest neighbors. By calculating the number of data points in each category among these k neighbors, assign the new data points to the category with the most neighbors.

Consider the scenario where we need to classify a new data point in order to use it.

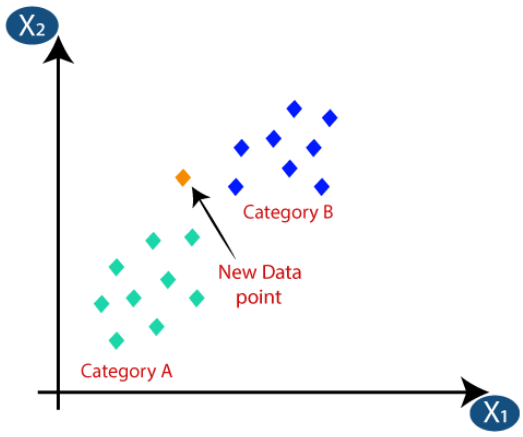

Fig.3.5.2.2 Classifying new data point in KNN algorithm.

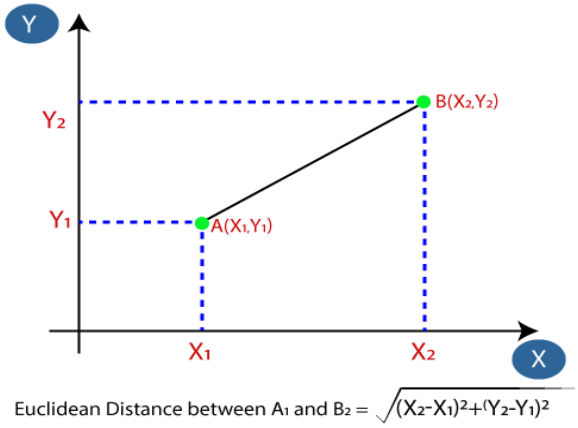

Fig. 3.5.2.3. Euclidean Distance calculation in KNN Algorithm.

By computing the Euclidean distance, we were able to determine the nearest neighbors, with three being in category A and two being in category B.

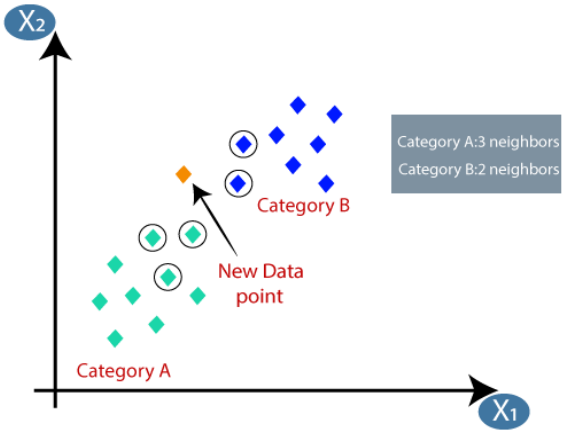

Fig.3.5.2.4 Determining K-value.

We need to try various values to determine which one performs the best since the ideal value for "K" cannot be found in a specific design. The number 5 is the most appropriate representation of the letter K. A small number of K, like 1 or 2, may be noisy and lead to outlier effects in the model. There might be some problems even though K ought to have high values.

3.5.3. Logistic Regression in Machine Learning:

Among the most popular Machine Learning algorithms, logistic regression [15] is a component of the supervised learning approach. With the help of a number of independent variables, the categorical dependent variable is predicted. Logistic regression is used for predicting a categorical dependent variable's outcome. Since the outcome must be discrete or categorical, it must be so. It can be True or False, Yes or No, 0 or 1, etc., but instead of giving the precise values of 0 and 1, it gives the probability values that are in the range between 0 and 1.

The method of application is the primary distinction between logistic regression and linear regression. While logistic regression is used to address classification problems, linear regression is used to solve regression problems. In logistic regression, compared to fitting a regression line, we fit a logistic function with a "S" shape that predicts two maximum values (0 or 1). The logistic function's curve demonstrates the likelihood of a number of different events, like determining if a mouse is obese based on its weight and whether the cells are cancerous, among other possibilities. It is a crucial machine learning technique because it can categorize new data using continuous and discrete datasets and also providing probabilities.

With the use of logistic regression, it is possible to categorize observations using a variety of data sources, and it is also straightforward to decide which variables to use. The logistic feature is depicted in the graphic below:

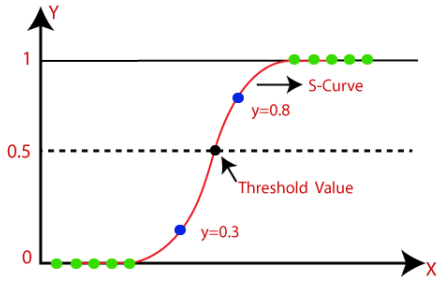

Fig.3.5.3. Logistic Feature

3.5.3.1.Logistic Function (Sigmoid Function):

Using the sigmoid function, a mathematical technique, the anticipated values are transformed into probabilities. It changes any real value from 0 to 1 to another value between those two ranges. The logistic regression has the form of a "S" curve because its value, which must be between 0 and 1, can never go over this range. The S-form [15] curve may also be referred to as the logistic function or sigmoid function. The threshold value concept is used in logistic regression to determine if there is a chance of 0 or 1. When values go below the threshold value, they frequently become zero and values like 1, for instance. A categorical variable must be the dependent one and, there should not be multi-colinearity in the independent variable.

3.5.3.2.Logistic Regression Equation:

$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \cdots + b_n x_n$ is the equation for a straight line. In Logistic Regression [12] y can be between 0 and 1, so for this we divide the above equation by (1-y), $\frac{y}{1-y}$; 0 for y=0, and infinity for y=1. But we need range between -[infinity] to +[infinity], by taking logarithm of the equation : $log \frac{y}{1-y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \cdots + b_n x_n$ is the Logistic Regression equation used.

To implement the Logistic Regression using Python, we will use the same steps as we have done in previous algorithms such as the Data Pre-processing, Fitting Logistic Regression to the Training set, Predicting the test result, Test accuracy of the result

(Creation of Confusion matrix) and Visualizing the test set result.

3.6. Training Model:

The training dataset, which is used to train or fit the machine learning model, is the largest (in terms of size) subset of the original dataset. The machine learning model is trained after it has been created in order to produce the desired outcomesA significant amount of pre-processed data is required to train a machine learning model. Pre-processed data in this context refers to structured data with fewer null values, etc. . To prepare for the ML algorithms to learn how to make predictions for the given task, training data is first fed into them. There are many chances that our model will perform horribly if we do not supply pre-processed data.

3.7. Validation:

By experimenting with various combinations of hyperparameters, we must train several models. The performance of each model is then assessed using the validation set. The validation test is useful for hyperparameter tuning or choosing the best model from a set of models, in this way.

3.8.Testing Data Set:

Once the model has been trained using the training dataset, it is necessary to test the model using the test dataset. This dataset assesses the model's performance and ensures that the model extrapolates well to new or untested datasets. The test dataset is a separate subset of the original data from the training dataset. But once the model training is over, it utilizes it as a benchmark for model evaluation because it has some comparable attributes and a similar class probability distribution. A well-organized dataset called test data contains data for each type of scenario the model might encounter in the actual world. The test dataset for an ML project typically makes up 20–25% of the entire original data.

At this point, we may also examine and contrast the testing accuracy with the training accuracy, or, more specifically, the accuracy of our model when applied to the test dataset in comparison to the training dataset. The model is considered to have overfitted if its accuracy on training data is higher than its accuracy on testing data.

3.9. Correlation:

A correlation heatmap is a visual tool that shows the correlation between many variables as a matrix with different colors. Similar to a color chart, it demonstrates how closely related various variables are.

Each variable is represented by a row and a column in a correlation heatmap, and the cells display the correlation between them. Each cell's color indicates the degree and direction of the link, with stronger correlations represented by hues that are lighter. Analyzing correlation heatmaps could assist us spot patterns and connections between various variables.

3.10. Prediction:

When predicting the likelihood of a specific outcome, "prediction" refers to the output of an algorithm that has been trained on historical data and applied to new data. The algorithm will produce probable values for an unknown variable for each record in the new data, allowing the model builder to determine what that value will most likely be.

3.11. Performance Evaluation:

Another crucial phase in creating a successful machine learning model is evaluating its performance. Different measurements, known as performance metrics or evaluation metrics, are employed to assess the effectiveness or fineness of the model. We may evaluate the effectiveness of our model [5] for the provided data using these performance indicators. By modifying the hyper-parameters, we can enhance the performance of the model. Performance measures are used to assess how successfully an ML model generalizes on new data, with the goal of each model being to do so.

IV. OBSERVATIONS AND CONCLUSIONS

By analyzing non-null values and data types, data preparation and EDA are carried out .The distribution

of all the features can be observed in histogram plots fig …. , along with the null values and duplicate values for each feature.

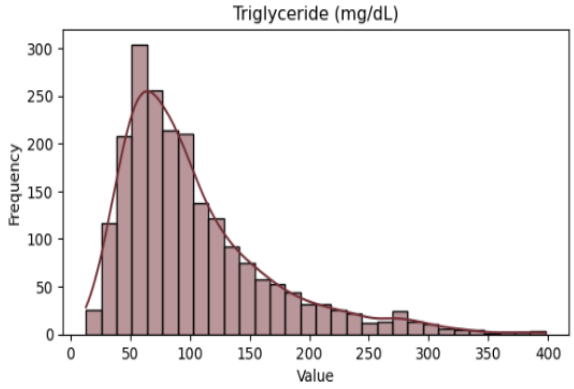

Fig. 4.1An Example for the distribution of given data.

Outliers are handled for all the features and reduced using the IQR (Inter Quartile Range) [2]approach by calculating the IQR=Q3-Q1, where Q1- is the lower limit and Q3 is the upper limit.

The heat map for the correlation matrix[10] is created using the new data set, and the higher the correlation, the lighter the color. According to this heat map, the LDL-Chol(mg/dL) and Tot.Cholestrol(mg/dL) have a strong correlation of 0.92.

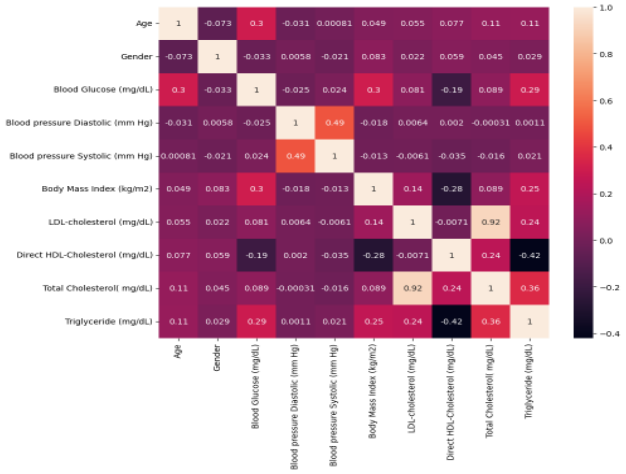

Fig 4.2 Correlation of the Variables

The Total Cholesterol increases when Triglycerides rise and the scatter plots created for Triglyceride levels with regard to age. We would be able to predict the triglycerides in relation to cholesterol by using body mass [12]as well.

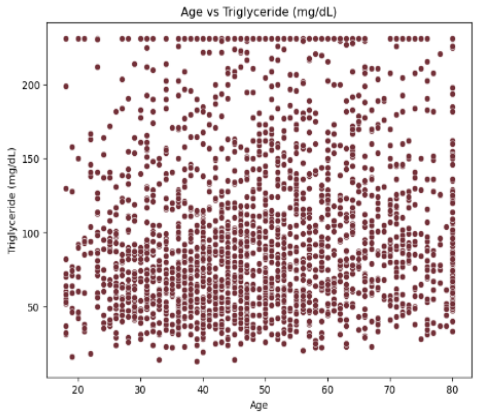

Fig.4.3 Scatter plot for Age Vs Triglycerides

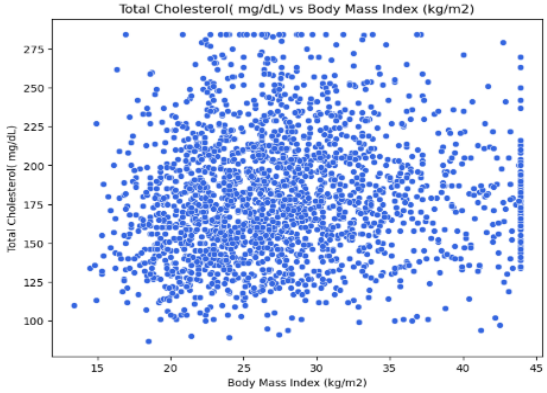

Fig 4.4 Scatter plot for Total Cholesterol vs Body mass Index

Consequently, prediction is carried out with the aid of Training data sets and Testing data sets. The model prediction is made using Supervised Leaning methods[14] like Random Forest Classifier Algorithm, KNN Algorithm, and Logistic Regression Algorithm.

A Random Forest Algorithm overfit model may appear impressive on the training set, but it will be useless in a practical setting. As a result, the typical method for hyperparameter optimization incorporates Grid Search CV for cross validation and builds a new forest for each iteration to account for overfitting.

The OOB score is a measure of the model's performance on unseen data, which means that the data are not used in the training of the model and are used to provide an unbiased estimate of the model's performance. The OOB score is calculated for ensemble models and the error rate varies from 0.036

to 0.042 and above. The OOB error rate for 70 trees is 0.03774 with score of 0.964033.

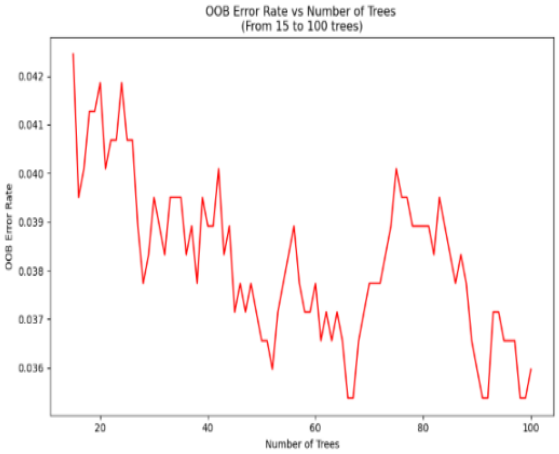

Fig.4.5 OOB Error rate with the trees.

In order to determine the accurate and incorrect values for the total count, a confusion matrix is created using the actual and predicted data. It aids in the creation of an effective data visualization and provides information on various types of errors that are being made by a classifier in addition to their frequency. The confusion matrix value for the RF algorithm includes 409 correct predictions, and 16 incorrect predictions, resulting in an accuracy level of 96%.

Classification Report:

	precision	recall	f1-score	support
High risk	0.94	0.94	0.94	125
Low risk	0.97	0.97	0.97	300
accuracy			0.96	425
macro avg	0.95	0.95	0.95	425
weighted avg	0.96	0.96	0.96	425

Fig 4.6 Classification report of Random Forest Algorithm

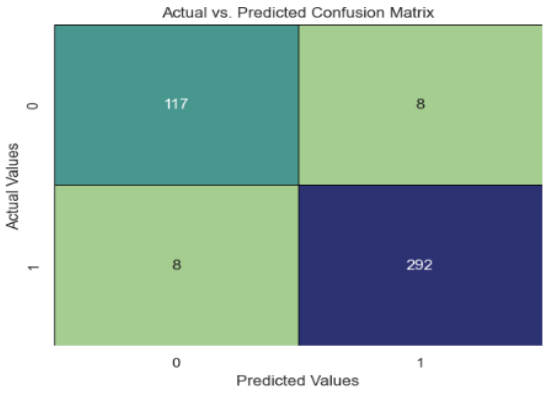

Fig.4.7 The confusion Matrix for Random Forest Algorithm.

We will employ feature scaling with logistic regression since we want accurate predictions. The classifier object is created to fit the model and its accuracy level of 98.12%.The confusion matrix will then be created so that the classification's accuracy may be verified. The function requires two parameters, true (the actual values) and predicted (the targeted value returned by the classifier) value and gives appropriately 417 correct predictions and 8 wrong predictions.

Classification Report:

	precision	recall	f1-score	support
High risk	0.98	0.98	0.98	505
Low risk	0.99	0.99	0.99	1191
accuracy			0.99	1696
macro avg	0.99	0.98	0.98	1696
weighted avg	0.99	0.99	0.99	1696

Fig 4.8 Classification report of Logistic Regression Algorithm

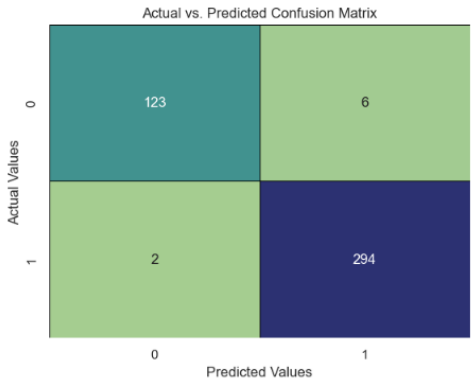

Fig.4.9 The confusion Matrix for Random Forest Algorithm

In KNN Algorithm, the default parameter is K, and N-neighbor determines the distance between the points with N- value of 10 and it will be equivalent to the common Euclidean metric.

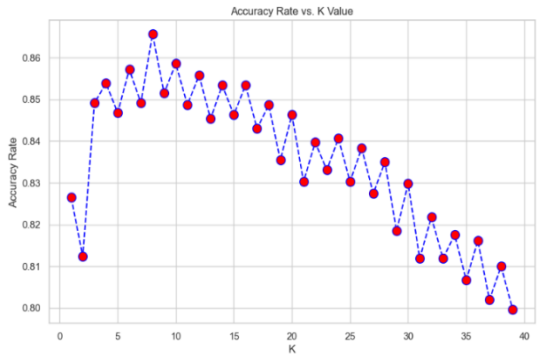

Fig 4.10 Accuracy Rate with respect to the K value

Hence the classifier has been fitted to the training set of data by producing 85.4% of accuracy. We have also imported the confusion matrix function and called it using the variable and generates 363 accurate predictions and 62 inaccurate predictions,.

Classification Report:

	precision	recall	f1-score	support
High risk	0.85	0.61	0.71	125
Low risk	0.85	0.96	0.90	300
accuracy			0.85	425
macro avg	0.85	0.78	0.81	425
weighted avg	0.85	0.85	0.85	425

Fig 4.11 Classification report of KNN Algorithm

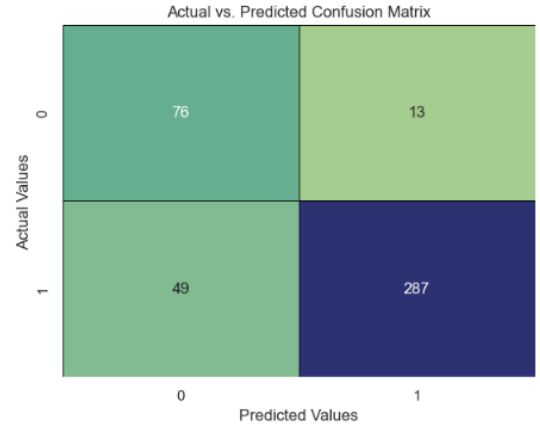

Fig 4.12 Confusion matrix of KNN Algorithm

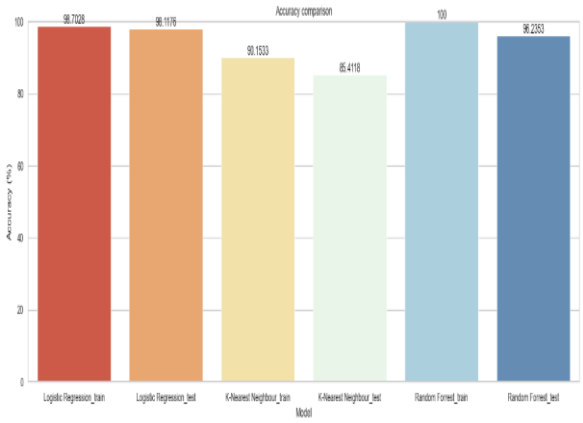

Fig. 4.13The Training and the test dataset accuracy level for various algorithms.

The training data set for the Random Forest Classifier method in the graph above contains no potential outcomes, therefore we chose the Logistic Regression approach instead, which offers a 99% accuracy level.

V. CONCLUSION

In this research, we propose a model for the prediction of hypertriglycerides using supervised machine learning algorithms based on several risk factors, such as heart illnesses, particularly CAD, stroke, liver, renal, and other chronic diseases. The model's significant feature is that it enables medical professionals to reassess the related risk and provide appropriate guidelines and medications to manage or for prevention of its occurance.

According to the performance research, data preparation is vital in order to produce a model that is both accurate and efficient. Thus, the results of the test proved that logistic regression was efficient, with recall of 0.98 and F1-measure of 0.99 with 99% accuracy.

The future model can be evolved through the deep learning techniques to predict triglyceride levels even in infants and newborns. This will enable the development of more precise prognoses and diagnoses with regard to remedies by analyzing large datasets of patient information including genomic data and other medical records.

VI. REFERENCES

[1]. Banda, J.M.; Sarraju, A.; Abbasi, F.; Parizo, J.; Pariani, M.; Ison, H.; Briskin, E.; Wand, H.; Dubois, S.; Jung, K.; et al. Finding missed cases of familial hypercholesterolemia in health systems using machine learning. npj Digit. Med. 2019, 2, 1–8.

[2]. Brautbar, A.; Leary, E.; Rasmussen, K.; Wilson, D.P.; Steiner, R.D.; Virani, S. Genetics of familial hypercholesterolemia. Curr. Atheroscler. Rep. 2015, 17, 1–17.

[3]. Claus Weihs1 · Katja Ickstadt2 “Data Science: the impact of statistics” January 2018.

[4]. Elias Dritsas * and Maria Trigka (2022) “Machine Learning Methods for Hypercholesterolemia Long-Term Risk Prediction” Department of Computer Engineering and Informatics, University of Patras, 26504 Patras, Greece;

MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

[5]. Emilie Westerlin Kjeldsen a,b, Jesper Qvist Thomassen a, Ruth Frikke-Schmidt "HDL cholesterol concentrations and risk of atherosclerotic cardiovascular disease – Insights from randomized clinical trials and human genetics" BBA - Molecular and Cell Biology of Lipids 1867 (2022).

[6]. Flint, A.C.; Conell, C.; Ren, X.; Banki, N.M.; Chan, S.L.; Rao, V.A.; Melles, R.B.; Bhatt, D.L. Effect of systolic and diastolic blood pressure on cardiovascular outcomes. N. Engl. J. Med. 2019, 381, 243–251.

[7]. García-d'Urso, N.; Climent-Pérez, P.; Sánchez-SanSegundo, M.; Zaragoza-Martí, A.; Fuster-Guillo, A.; Azorin-Lopez, J. "A noninvasive approach for total cholesterol level prediction using machine learning". IEEE Access 2022, 10, 58566–58577.

[8]. Goran Walldiusa,b and Ingmar Jungnerc,d "Apolipoprotein A-I versus HDL cholesterol in the prediction of risk for myocardial infarction and stroke" Lippincott Williams & Wilkins 2007.

[9]. Jing Ma1, Jiong Yu2, Guangshu Hao3, Dan Wang3, Yanni Sun3, Jianxin Lu3, Hongcui Cao2,3* and Feiyan Lin4* "Assessment of triglyceride and cholesterolin overweight people based on multiple linear regression and artificial intelligence model" Ma et al. Lipids in Health and Disease (2017)

[10]. Karimollah Hajian-Tilaki (PhD) 1* Behzad Heidari (MD) 2 Afsaneh Bakhtiari (PhD) 3 Triglyceride to high-density lipoprotein cholesterol and low-density lipoprotein cholestrol to high-density lipoprotein cholesterol ratios are predictors of cardiovascular risk in Iranian adults: Evidence from a population-based cross-sectional study. Social Determinants of Health Research Center, Health Research Institute, Babol University of Medical Sciences, Babol, Iran. Caspian J Intern Med 2020. 33.

[11]. Khirfan, G.; Tejwani, V.; Wang, X.; Li, M.; DiDonato, J.; Dweik, R.A.; Smedira, N.; Heresi, G.A. Plasma levels of high density lipoprotein cholesterol and outcomes in chronic thromboembolic pulmonary hypertension. PLoS ONE 2018, 13, e0197700.

[12]. Konstantoulas, I.; Kocsis, O.; Dritsas, E.; Fakotakis, N.; Moustakas, K. Sleep Quality Monitoring with Human Assisted Corrections. In Proceedings of the International Joint Conference on Computational Intelligence (IJCCI), SCIPTRESS, Valletta, Malta, 25–27 October 2021; pp. 435–444. 24.

[13]. Krishnan, S.; Geetha, S. Prediction of Heart Disease Using Machine Learning Algorithms. In Proceedings of the 2019 IEEE 1st International Conference on Innovations in Information and Communication Technology (ICIICT), Chennai, India, 25–26 April 2019; pp. 1–5.

[14]. Lee, B.J. Prediction model of hypercholesterolemia using body fat mass based on machine learning. J. Converg. Cult. Technol. 2019.

[15]. Mahendra Kumar Gourisaria1 Gaurav Jee1 G. M. Harshvardhan1 Vijander Singh2 Pradeep Kumar Singh3 Tewabe Chekole Workneh4 "Data science appositeness in diabetes mellitus diagnosis for healthcare systems of developing nations" December 2021.

[16]. Nahuel García-D'urso1, Pau Climent-Pérez 1, Miriam Sánchez-Sansegundo2, Ana Zaragoza-Martí3, Andrés Fuster-Guilló 1, And Jorge Azorín-López 1 1Department of Computer Technology, University of Alicante, 03690 Alicante, Spain "A Non-Invasive Approach for Total Cholesterol Level Prediction Using Machine Learning" May 2022.

[17]. Nusinovici, S.; Tham, Y.C.; Yan, M.Y.C.; Ting, D.S.W.; Li, J.; Sabanayagam, C.; Wong, T.Y.; Cheng, C.Y. Logistic regression was as good as

machine learning for predicting major chronic diseases. J. Clin. Epidemiol. 2020, 122, 56–69.

[18]. Park, H.; Kim, K. Comparisons among machine learning models for the prediction of hypercholestrolemia associated with exposure to lead, mercury, and cadmium. Int. J. Environ. Res. Public Health 2019, 16, 2666.

[19]. Pavithra, V.; Jayalakshmi, V. Hybrid feature selection technique for prediction of cardiovascular diseases. Mater. Today Proc. 2021, in press.

[20]. Pina, A.; Helgadottir, S.; Mancina, R.M.; Pavanello, C.; Pirazzi, C.; Montalcini, T.; Henriques, R.; Calabresi, L.; Wiklund, O.; Macedo, M.P.; et al. Virtual genetic diagnosis for familial hypercholesterolemia powered by machine learning. Eur. J. Prev. Cardiol. 2020, 27, 1639–1646.

[21]. Saba, T. Recent advancement in cancer detection using machine learning: Systematic survey of decades, comparisons and challenges. J. Infect. Public Health 2020, 13, 1274–1289.

[22]. Sajida Perveen1, Muhammad Shahbaz1,2, Karim Keshavjee 2,3 & Aziz Guergachi2,4,5 “A Systematic Machine Learning based Approach for the Diagnosis of Non-Alcoholic Fatty Liver Disease Risk and Progression” February 2018.

[23]. Tangirala, S. Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm. Int. J. Adv. Comput. Sci. Appl. 2020, 11, 612–619.

[24]. Zaibunnisa L. H. Malik 1, Momin Fatema 2, Nikam Pooja 3, Gawandar Ankita 4, “Heart Disease Prediction using Artificial Intelligence” International Journal of Engineering Research & Technology (IJERT) 2021.

Cite this article as :

S. Alagu Thangam, S. Vilma Veronica, G. Suresh, S. Hemalatha, "Prognosis of Hyper Triglycerides Using Data Science and Machine Learning", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 10 Issue 5, pp. 229-239, September-October 2023. Available at doi : https://doi.org/10.32628/IJSRSET2310540
Journal URL : https://ijsrset.com/IJSRSET2310540