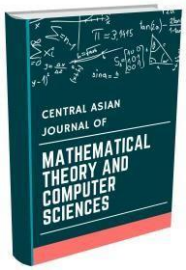

CENTRAL ASIAN JOURNAL OF MATHEMATICAL THEORY AND COMPUTER SCIENCES
https://cajmtcs.centralasianstudies.org/index.php/CAJMTCS
Volume: 05 Issue: 03 | July 2024 ISSN: 2660-5309

Article

Analyzing Factors Affecting Cholesterol Levels: A Data-Driven Study Using Statistical Models and Machine Learning

Zahraa Tariq Mohammed Taher[1]

1. Medicine College- Family and Community Medicine/ Ninevah University- Mosul-Iraq
* Correspondence author email: zahraa.mohammed@uoninevah.edu.iq

Abstract: High cholesterol levels are associated with various health complications, particularly cardiovascular diseases. Predicting an individual's cholesterol levels can be crucial in healthcare analytics to prevent and manage these conditions, leading to long-term health benefits and potential economic savings in healthcare systems. This paper focuses on analyzing the factors influencing blood cholesterol levels and developing predictive models for cholesterol levels using statistical and machine learning techniques. The study conducted an extensive analysis of determinants governing cholesterol levels using a feature importance ratio and leveraged the Framingham Heart Study (FHS) dataset with machine learning techniques to predict cholesterol levels. The study found that age, BMI, and glucose levels consistently influenced cholesterol levels, whether classified into three levels or two levels. Machine learning models exhibited varying performance, with models like Random Forest and Gradient Boosting excelling in precision, recall, and F1-score in specific cholesterol categories. The results emphasize the importance of addressing age, BMI, and glucose levels in healthcare strategies for cholesterol management. They also highlight the need for continuous model refinement and fine-tuning to improve predictive accuracy in different cholesterol classification scenarios.

Keywords: Cholesterol, Predictive Models, Feature Importance, Framingham Heart Study, Classification, Risk Factors, Data Analysis.

Citation: Zahraa Tariq Mohammed Taher. Analyzing Factors Affecting Cholesterol Levels: A Data-Driven Study Using Statistical Models and Machine Learning Central Asian Journal of Mathematical Theory and Computer Sciences 2024, 5(3), 220-230.

Received: 10th Apr 2024
Revised: 11th Mei 2024
Accepted: 24th Jun 2024
Published: 27th Jul 2024

1. Introduction

Cholesterol stands as a fundamental bioactive compound within the human organism, holding pivotal roles in a multitude of physiological processes. Nevertheless, elevated cholesterol concentrations can engender health complexities, primarily in the context of cardiovascular afflictions [1-3]. Therefore, grasping the determinants that modulate cholesterol levels assumes paramount importance in the realms of both public health and medical research [4-5]. Cholesterol manifests in three primary categories within cellular structures: High-density lipoprotein (HDL), acknowledged as "beneficial" cholesterol, facilitates the expulsion of surplus cholesterol from the organism [6-8]. Conversely, low-density lipoprotein (LDL), often dubbed "detrimental" cholesterol, contributes to the accumulation of arterial plaque. Additionally, very low-density lipoprotein (VLDL) plays a role in promoting the accrual of atheromatous deposits [8][9].

Predicting an individual's vulnerability to non-communicable chronic conditions, such as high cholesterol, closely tied to modifiable lifestyle choices and

attitudes, represents a pivotal goal in the realm of healthcare predictive analytics [10]. These predictions carry substantial implications for an individual's enduring health, their potential for active and autonomous aging, and, of no less significance, the potential for significant economic savings within societal healthcare frameworks. Recent research underscores the capability of machine learning tools to forecast an individual's risk of hospitalization, relying exclusively on socioeconomic and behavioral data, while circumventing the necessity for clinical risk factors [11-13]. The evaluation of cholesterol levels typically involves lipid profiles or blood cholesterol assessments [14]. Despite the absence of overt clinical symptoms associated with elevated cholesterol levels, elucidating the interplay among various cholesterol subtypes offers valuable insights into predisposition to cardiovascular maladies [14][15]. Consequently, strategies aimed at averting or mitigating high cholesterol levels translate directly into a reduction of cardiovascular disease susceptibility. Several pivotal risk factors influencing heightened cholesterol levels encompass gender, age, familial proclivity to heart diseases, dietary habits, body mass index (BMI), physical activity levels, alcohol consumption, smoking history, and the presence of diabetes [16-19].

This study makes significant contributions in two primary facets: Firstly, it conducts an exhaustive examination of determinants governing cholesterol levels using a feature importance ratio, delineating the predictors of heightened cholesterol risk by leveraging the Framingham Heart Study (FHS) dataset derived from case records-based machine learning techniques.

This section presents an overview of relevant studies in the field of cholesterol prediction and its associated factors using machine learning techniques. The literature is summarized to identify gaps and unresolved issues that warrant further investigation.

The paper [20] focuses on the estimation of low-density lipoprotein-cholesterol (LDL-C) levels using machine learning techniques. It highlights the limitations of the Friedewald equation in estimating LDL-C, especially in cases of high triglycerides or non-fasting states. The authors propose novel machine learning algorithms, LDL-CX and LDL-CN, which outperform conventional methods. However, the study does not delve into the broader factors affecting cholesterol levels, leaving room for further exploration.

In [21], the authors introduce a non-invasive machine learning approach for predicting total cholesterol levels. This method utilizes clinical and anthropometric data collected during weight loss interventions. The study focuses on improving non-invasive diagnosis quality and disease screening. Through clustering analysis, it identifies patient groups with shared characteristics that may hold valuable diagnostic information. The results demonstrate the potential of machine learning to predict cholesterol levels with low mean absolute percentage error rates, offering a non-invasive tool for clinical applications.[22] Explores the integration of health data-driven machine learning algorithms to assess the risk factors of early-stage hypertension, especially in individuals with dyslipidemia. The study leverages a large dataset and various machine learning techniques to identify the complex relationships between risk factors and early-stage hypertension incidence. Notably, it highlights the importance of variables like age, body mass index, glucose levels, and C-reactive protein in predicting hypertension. This research emphasizes the potential of data-driven machine learning in early disease prediction.

In [23], the authors investigate the application of machine learning in managing lipid disorders, particularly in high-risk patients receiving cholesterol-lowering medications in primary care. They developed machine learning algorithms based on lipid management guidelines and used natural language processing to extract medication information from electronic records. The study showcases how

machine learning can identify suboptimal prescribing patterns, target high-risk patients for more intensive therapy, and suggest evidence-based therapeutic options. However, it mainly focuses on optimizing medication management, leaving unexplored areas regarding the broader factors influencing cholesterol levels.[24] evaluates the applicability of a machine learning-based method for estimating LDL-C levels and assesses the influence of training dataset characteristics. The study identifies the importance of dataset characteristics in achieving accurate estimates. It underscores the need to train machine learning models on datasets with matched characteristics, highlighting the versatility of machine learning methods.[25] addresses the estimation of very low-density lipoprotein cholesterol (VLDL-C) using interpretable machine learning techniques. It aims to predict VLDL-C values based on attributes such as age, sex, and various laboratory measurements. The study finds that the generalized linear model (GLM) yields the best results among the techniques employed.

in paper [26] explores the prediction of blood pressure and cholesterol using machine learning, with a focus on a Multiple Linear Regression Analysis (MRA) approach. The study analyzes primary data collected from various districts in West Bengal, India. It demonstrates that machine learning models can predict the presence of blood pressure and cholesterol-related conditions with a high level of accuracy.In [27], the authors delve into the use of ensemble learning techniques for disease prediction, specifically targeting diabetes and cholesterol diseases. The study demonstrates that ensemble learning algorithms, such as Adaboost, Random Forest, Bagging, Voting, and Stacking, outperform individual algorithms. This research emphasizes the role of ensemble learning in predicting non-communicable diseases and the significance of key parameters in disease prediction.

The existing literature highlights the potential of machine learning in predicting cholesterol-related factors and diseases. However, there is a notable gap in comprehensively understanding the multifaceted relationships between cholesterol levels and various influencing factors. Further research is warranted to bridge these gaps and enhance our understanding of cholesterol-related health issues.

The aim of the study is to analyze the factors influencing blood cholesterol levels and develop predictive models for cholesterol levels using statistical and machine learning techniques.

To achieve this aim, the following objectives are accomplished:

– To examine the relationships between Variables and cholesterol levels.

– To apply statistical and machine learning to understand the associations between these factors and cholesterol levels.

– To develop predictive models for cholesterol levels using machine learning techniques.

2. Materials and Methods

The study involved the application of several machine learning models to predict cholesterol levels using the Framingham Heart Study (FHS) dataset. The models were evaluated based on their performance in both three-level and two-level cholesterol classifications. Several metrics were used to evaluate the models for both the three-level and two-level cholesterol classifications. The metrics included precision, recall, F1-score, and accuracy (ACC).

Machine learning model

In this study, various machine learning models were employed to analyze the factors influencing blood cholesterol levels. The determination of a suitable learning algorithm for a given dataset and case study is a crucial. Typically, the empirical approach involving the testing and evaluation of multiple algorithms is the preferred methodology to ensure the selection of an optimal approach. In this regard,

have undertaken an investigation of several algorithms, widely used in the literature, with the aim of identifying the most suitable algorithm for our specific needs. Fig 1 shows the framework.

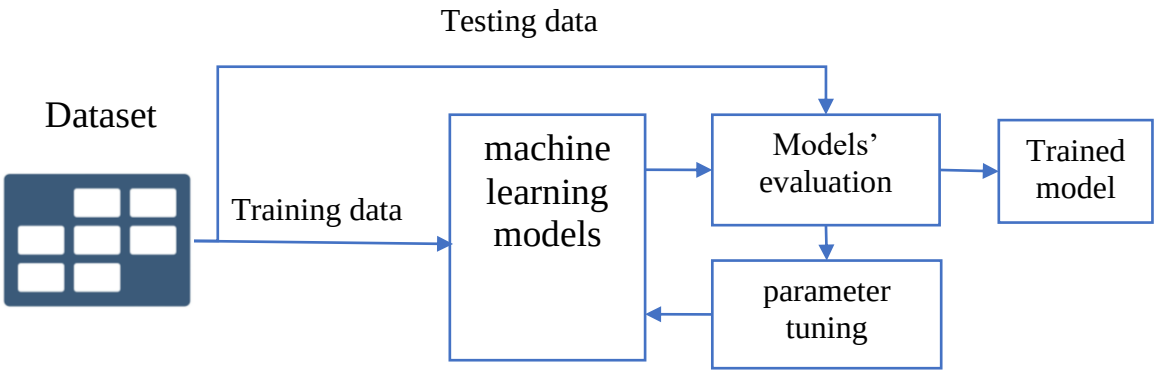

Fig 1: framework of machine learning models

Random Forest (RF) is an ensemble learning method that combines multiple decision trees to improve predictive accuracy and reduce overfitting. It is particularly effective for classification tasks, such as categorizing cholesterol levels [28-29].Gradient Boosting (GB) is another ensemble method that builds decision trees sequentially, with each tree attempting to correct the errors of the previous one. It often provides high predictive performance [30].

Support Vector Classifier (SVC) is a supervised learning algorithm that can be used for classification tasks. It works by finding the optimal hyperplane that best separates different classes, making it suitable for binary and multiclass classification problems [31-32].

K-Nearest Neighbors (KNN) is a simple yet effective algorithm for classification. It classifies data points based on the majority class among their k-nearest neighbors. It is particularly useful when dealing with local patterns in the data [33-34]. Decision Trees (CT) are a straightforward and interpretable machine learning model. They split data into branches based on feature values, creating a tree-like structure for decision-making [35][29]. Naive Bayes (NB) is a probabilistic classifier that makes predictions based on Bayes' theorem. It is often used for text classification but can also be applied to other classification tasks [36].

Multi-Layer Perceptron (MLP) is a type of artificial neural network with multiple layers of nodes (neurons). It can be used for various machine learning tasks, including classification [37-38]. These models were used to predict cholesterol levels based on the provided dataset, which included various features such as gender, age, education, smoking habits, blood pressure, and more.

Data and preprocessing

In this study, data from the Framingham Heart Study (FHS) were employed [39]. FHS is a renowned and extensive cardiovascular cohort study. The dataset contains various variables relevant to our exploration of factors impacting blood cholesterol levels. Table 1 is a summary of the dataset, along with descriptions of its variables.

Table 1: Dataset Overview and Variables Description

Variable	Description	Type

male	Male or Female	Categorical (0 = Female; 1 = Male)
age	Age at exam time in years	Continuous
education	Education of the patient	Categorical (1 = Some High School; 2 = High School or GED; 3 = Some College or Vocational School; 4 = College)
currentSmoker	At present smoker or not smoking	Categorical (0 = No Smoking; 1 = Smoking)
cigsPerDay	Smoking habits - Average no. of cigarettes/day	Continuous
BPMeds	Blood Pressure medications	Categorical (0 = Not taking any Blood Pressure medications; 1 = Already on Blood Pressure medications)
prevalentStroke	Fasting blood sugar > 120 mg/dl	Categorical (0 = False; 1 = True)
prevalentHyp		
diabetes	Diabetes present or not	Categorical (0 = No; 1 = Yes)
totChol	Total amount of cholesterol in blood	Continuous (mg/dL)
sysBP	Systolic blood pressure	Continuous (mmHg)
diaBP	Diastolic blood pressure	Continuous (mmHg)
BMI	Body Mass Index	Continuous (kg/m2)
heartRate	Beats/Min (Ventricular)	Continuous
glucose	Glucose level in blood	Continuous

In preparation for data analysis, several preprocessing steps were carried out. Firstly, the 'totChol' variable was categorized into three classes for triple classification: 'Desirable', 'Borderline high', and 'High', with specific thresholds applied (see Table 2). Values less than or equal to 200 mg/dL were labeled 'Desirable', those between 201 mg/dL and 239 mg/dL were classified as 'Borderline high', and values equal to or exceeding 240 mg/dL were designated as 'High' [40]. The threshold limit for the three cholesterol levels was chosen based on the National Library of Medicine (NIH) [41]. In the case of binary classification, 'normal' was assigned to values less than or equal to 200 mg/dL, while 'high' covered values between 201 mg/dL and 239 mg/dL.

Table 2: Total Cholesterol Level Classification

Total Cholesterol Level	**Class**
Less than 200mg/dL	Desirable
200-239 mg/dL	Borderline high
240mg/dL and above	High

Feature selection involved the inclusion of variables such as 'male', 'age', 'education', 'currentSmoker', 'cigsPerDay', 'BPMeds', 'prevalentStroke', 'prevalentHyp', 'diabetes', 'sysBP', 'diaBP', 'BMI', 'heartRate', and 'glucose' as predictors for machine learning models. Subsequently, the dataset was split into training (80%) and testing (20%) sets using the `train_test_split` function, and feature scaling was applied through the `StandardScaler` to standardize all features with a mean of (0) and a standard deviation of (1).

3. Results

In the following section, the study presents the outcomes of the analysis, encompassing feature importance and the performance metrics of diverse machine learning models employed for the prediction of cholesterol levels.

Feature Importance

In this section, the feature importance analysis is explored, a critical aspect of our study designed to comprehend the factors that exert significant influence on blood

cholesterol levels. Feature importance serves to identify the most influential variables for predicting cholesterol levels, offering valuable insights for healthcare professionals and researchers alike. In the analysis, the categorization of cholesterol levels into three classes - 'Desirable,' 'Borderline high,' and 'High' - was initially undertaken using specific thresholds. Fig 1 displays the average importance scores of the chosen features based on this three-level classification.

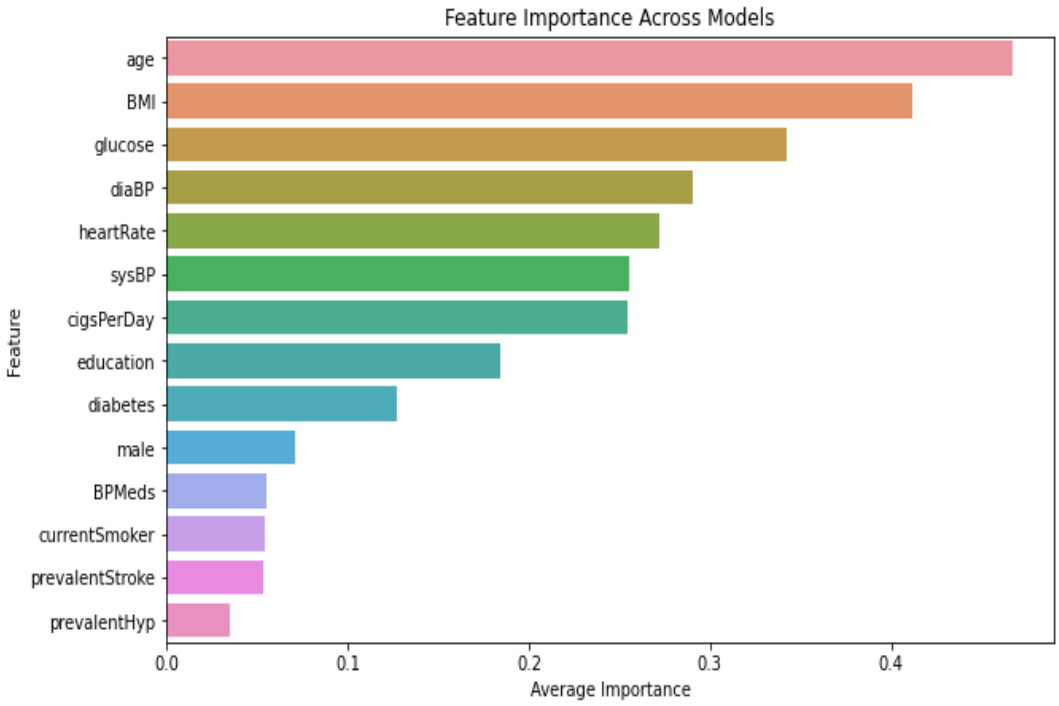

Fig 1: Average Importance based on Three-level Classification

From the information presented in the figure, it becomes apparent that age, BMI (Body Mass Index), and glucose levels hold substantial roles in predicting cholesterol levels. These variables rank among the most influential factors in determining whether cholesterol levels fall within the categories of 'Desirable,' 'Borderline high,' or 'High.' Additionally, factors such as diastolic blood pressure (diaBP), heart rate, systolic blood pressure (sysBP), and the average number of cigarettes smoked per day (cigsPerDay) also demonstrate notable importance in this classification.To simplify the analysis, a two-level classification was also conducted, discerning between 'normal' and 'high' cholesterol levels. Fig 2 showcases the average importance scores of the chosen features based on this binary classification.

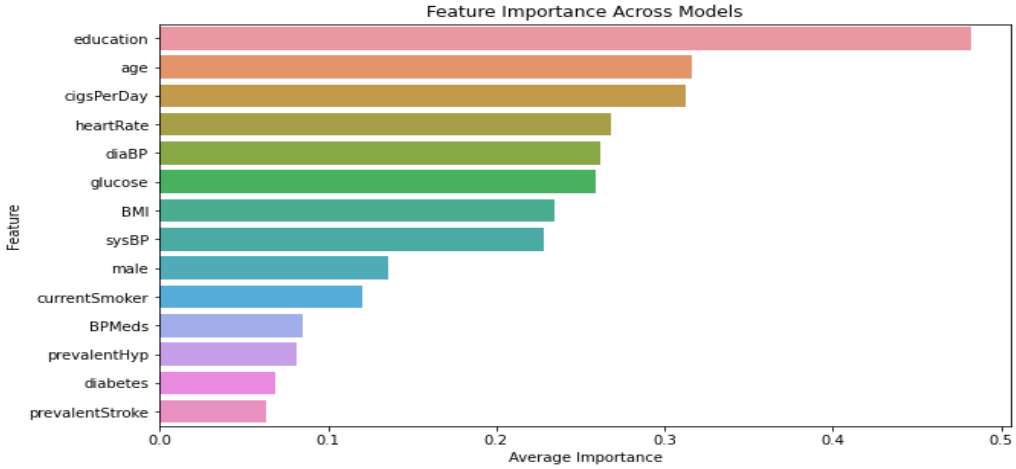

Fig 2: Average Importance based on Two-level Classification

Within the framework of the two-level classification, education, age, and the average number of cigarettes smoked per day (cigsPerDay) emerge as the most influential features for distinguishing between 'normal' and 'high' cholesterol levels. These findings underline the significant role played by education level in predicting whether an individual is predisposed to 'high' cholesterol levels.

Model performance

In this section, the performance of various machine learning models is assessed concerning their ability to predict cholesterol levels. The evaluation is based on metrics such as precision, recall, F1-score, and accuracy (ACC) for both three-level and two-level classifications. The three-level classification comprises cholesterol levels categorized as 'High,' 'Borderline,' and 'Desirable,' while the two-level classification differentiates between 'High' and 'Desirable' cholesterol levels.

Table 3: Three-level classification model performance

	High			Borderline			Desirable			
Model	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	ACC
RF	0.82	0.95	0.88	0.44	0.1	0.16	0.26	0.12	0.16	0.78
GB	0.85	0.65	0.74	0.14	0.33	0.19	0.18	0.31	0.22	0.58
SVC	0.88	0.59	0.7	0.12	0.67	0.2	0.17	0.17	0.17	0.53
KNN	0.82	0.58	0.68	0.12	0.24	0.16	0.16	0.38	0.23	0.53
CT	0.82	0.77	0.8	0.04	0.05	0.05	0.15	0.2	0.17	0.65
NB	0.84	0.06	0.12	0.09	0.74	0.16	0.1	0.35	0.16	0.14
MLP	0.86	0.54	0.66	0.1	0.17	0.12	0.16	0.46	0.24	0.5

From the results, it is evident that different machine learning models exhibit varying performance in the three-level cholesterol classification. Random Forest (RF) achieves the highest precision and recall for the 'High' cholesterol category, indicating that it effectively identifies individuals with high cholesterol levels. However, Gradient Boosting (GB) demonstrates the highest F1-score for 'High' cholesterol, suggesting a balanced trade-off between precision and recall. Support Vector Classifier (SVC) achieves the highest precision for 'Desirable' cholesterol, highlighting its ability to accurately identify individuals with desirable cholesterol levels. Nevertheless, model performance varies across different cholesterol categories.

Table 4: Two-level classification model performance

	High			Desirable			
Model	Precision	Recall	F1	Precision	Recall	F1	ACC
RF	0.91	0.95	0.93	0.29	0.18	0.22	0.88
GB	0.92	0.95	0.93	0.33	0.21	0.25	0.88
SVC	0.93	0.67	0.78	0.15	0.54	0.24	0.66
KNN	0.93	0.69	0.79	0.15	0.51	0.24	0.68
CT	0.91	0.84	0.87	0.15	0.26	0.19	0.78
NB	0.93	0.06	0.11	0.1	0.96	0.18	0.15
MLP	0.92	0.83	0.87	0.16	0.29	0.21	0.78

In the two-level cholesterol classification, where the focus is on distinguishing between 'High' and 'Desirable' cholesterol levels, the models exhibit improved

performance compared to the three-level classification. Support Vector Classifier (SVC) and K-Nearest Neighbors (KNN) achieve high precision for 'High' cholesterol, indicating their accuracy in identifying individuals with high cholesterol levels. Gradient Boosting (GB) and Random Forest (RF) demonstrate strong F1-scores for both 'High' and 'Desirable' categories, reflecting a balance between precision and recall. However, it's important to note that Naive Bayes (NB) has a particularly low recall for 'High' cholesterol, indicating a higher rate of false negatives.

4. Discussion

The feature importance analysis in this study revealed key insights into the factors that significantly influence cholesterol levels. Age, BMI, and glucose levels consistently emerged as some of the most influential variables across both three-level and two-level cholesterol classifications. Healthcare professionals can utilize these insights to tailor interventions for individuals with high cholesterol. For instance, focusing on lifestyle modifications for individuals with high BMI or elevated glucose levels may be particularly effective in managing cholesterol. These findings emphasize the importance of personalized medicine. Different individuals may have varying risk factors for high cholesterol, and understanding these individualized factors can lead to more effective healthcare strategies. The evaluation of machine learning models for cholesterol prediction highlighted their varying performance based on different classification scenarios. Models like Random Forest (RF) and Gradient Boosting (GB) performed well in terms of precision, recall, and F1-score, depending on the classification task. However, Support Vector Classifier (SVC) and K-Nearest Neighbors (KNN) exhibited strengths in precision for certain cholesterol categories. The variability in model performance emphasizes the importance of continual model refinement and fine-tuning. Researchers and data scientists can work on improving models to enhance their predictive accuracy across all cholesterol categories.

5. Conclusion

In Conclusion, this study explored the intricacies of cholesterol levels, emphasizing its pivotal role in human physiology and its direct link to cardiovascular health. Notably, the research uncovered age, BMI, and glucose levels as consistently influential factors in determining cholesterol levels, irrespective of whether the classification was three-level or two-level. These findings underscore the imperative need to address these factors in healthcare strategies and personalized medicine to mitigate the risk of cardiovascular diseases. Furthermore, the assessment of various machine learning models highlighted their diverse performances in cholesterol prediction. Models such as Random Forest (RF) and Gradient Boosting (GB) demonstrated proficiency in different aspects of precision, recall, and F1-score, contingent upon the specific cholesterol classification task. Support Vector Classifier (SVC) and K-Nearest Neighbors (KNN) showcased notable precision in specific cholesterol categories. This variance in model performance underscores the critical importance of carefully selecting an appropriate model based on the intended application and continuously refining these models to enhance their predictive accuracy.

REFERENCES

[1] Huff, Trevor & Brandon Boyd & Ishwarlal Jialal. (2023). Physiology, Cholesterol. In StatPearls. StatPearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK470561/

[2] Duan Y, Gong K, Xu S, Zhang F, Meng X, Han J. Regulation of cholesterol homeostasis in health and diseases: from mechanisms to targeted therapeutics. Signal Transduct Target Ther. 2022 Aug 2;7(1):265. doi: 10.1038/s41392-022-01125-5. PMID: 35918332; PMCID: PMC9344793.

[3] Dybiec, Jill & Baran, Wiktoria & Dąbek, Bartłomiej & Fularski, Piotr & Młynarska, Ewelina & Radzioch, Ewa & Rysz, Jacek & Franczyk, Beata. (2023). Advances in Treatment of Dyslipidemia. International Journal of Molecular Sciences. 24. 13288. 10.3390/ijms241713288.

[4] Sureshbabu, Jayanthi. (2023). Importance and Need of Medical Entomology and Medical Entomologist in Public Health. International Journal of Medical Sciences and Nursing Research. 3. 1-2. 10.55349/ijmsnr.20233312.

[5] Hidayat, Anas. (2023). MANAGEMENT OF INCREASING PUBLIC KNOWLEDGE ABOUT THE IMPORTANCE OF MEDICAL RECORDS IN HEALTH CARE FACILITIES. Jurnal Pengabdian Masyarakat Permata Indonesia. 3. 7-11. 10.59737/jpmpi.v3i1.218.

[6] Dickens, Brian & Sassanpour, Mana & Bischoff, Evan. (2023). The Effect of Chia Seeds on High-Density Lipoprotein (HDL) Cholesterol. Cureus. 15. 10.7759/cureus.40360.

[7] Zhu, Chen & Wu, Juan & Wu, Yixian & Guo, Wen & Lu, Jing & Zhu, Wenfang & Li, Xiaona & Xu, Nianzhen & Zhang, Qun. (2022). Triglyceride to high-density lipoprotein cholesterol ratio and total cholesterol to high-density lipoprotein cholesterol ratio and risk of benign prostatic hyperplasia in Chinese male subjects. Frontiers in Nutrition. 9. 10.3389/fnut.2022.999995.

[8] Katahira, Masahito & Imai, Shu & Ono, Satoko & Moriura, Shigeaki. (2023). Estimating Triglyceride Levels Using Total Cholesterol, Low-Density Lipoprotein Cholesterol, and High-Density Lipoprotein Cholesterol Levels: A Cross-Sectional Study. Metabolic syndrome and related disorders. 21. 10.1089/met.2023.0045.

[9] Wu, Shouling & Su, Xin & Zuo, Yingting & Chen, Shuohua & Tian, Xue & Xu, Qin & Zhang, Yijun & Zhang, Xiaoli & Wang, Penglian & He, Yan & Wang, Anxin. (2023). Discordance between remnant cholesterol and low-density lipoprotein cholesterol predicts arterial stiffness progression. Hellenic Journal of Cardiology. 10.1016/j.hjc.2023.05.008.

[10] Siddharth, Saurav & Farooq, Bilkisa & Kumar, Nirnay & Burhan, Mirza. (2023). Effect of Lifestyle in Female Infertility: A Review Based Study. International Journal for Research in Applied Science and Engineering Technology. 11. 1777-1783. 10.22214/ijraset.2023.56307.

[11] Hernández-Arango, Alejandro & Arias, María & Pérez, Viviana & Chavarría, Luis & Jaimes, Fabian & Mater, Alma. (2023). Prediction of the risk of adverse clinical outcomes with machine learning techniques in patients with chronic no communicable diseases.

[12] Lukyanenko, Roman & Maass, Wolfgang & Storey, Veda. (2022). Trust in artificial intelligence: From a Foundational Trust Framework to emerging research opportunities. Electronic Markets. 32. 3. 10.1007/s12525-022-00605-4.

[13] Ahuja, Abhimanyu. (2019). The impact of artificial intelligence in medicine on the future role of the physician. PeerJ. 7. e7702. 10.7717/peerj.7702.

[14] Lokpo, Sylvester & Laryea, Roger & Osei-Yeboah, James & Owiredu, William & Ephraim, Richard & Adejumo, Esther & Ametepe, Samuel & Appiah, Michael & Nogo, Peter & Affrim, Patrick & Precious Kwablah, Kwadzokpui & Abeka, Ohene Kweku. (2022). The pattern of dyslipidaemia and factors associated with elevated levels of non-HDL-cholesterol among patients with type 2 diabetes mellitus in the Ho municipality: A cross sectional study. Heliyon. 8. e10279. 10.1016/j.heliyon.2022.e10279.

[15] Verbeek, Rutger & Hoogeveen, Renate & Langsted, Anne & Stiekema, Lotte & Verweij, Simone & Hovingh, G. Kees & Wareham, Nicholas & Khaw, Kay-Tee & Boekholdt, S & Nordestgaard, Børge & Stroes, Erik. (2018). Cardiovascular disease risk associated with elevated lipoprotein(a) attenuates at low low-density lipoprotein cholesterol levels in a primary prevention setting. European heart journal. 39. 10.1093/eurheartj/ehy334.

[16] Shao, Zeguo & Xiang, Yuhong & Zhu, Yingchao & Fan, Aiqin & Zhang, Peng. (2020). Influences of Daily Life Habits on Risk Factors of Stroke Based on Decision Tree and Correlation Matrix. Computational and Mathematical Methods in Medicine. 2020. 1-12. 10.1155/2020/3217356.

[17] Schmidt, Gilda & Schneider, Christina & Gerlinger, Christoph & Endrikat, Jan & Gabriel, Lena & Ströder, Russalina & Müller, Carolin & Juhasz-Böss, Ingolf & Solomayer, Erich-Franz. (2020). Impact of body mass index, smoking habit, alcohol consumption, physical activity and parity on disease course of women with triple-negative breast cancer. Archives of Gynecology and Obstetrics. 301. 10.1007/s00404-019-05413-4.

[18] Chua, Shiao & Yovich, Steven & Hinchliffe, Peter & Yovich, John. (2023). Male Clinical Parameters (Age, Stature, Weight, Body Mass Index, Smoking History, Alcohol Consumption) Bear Minimal Relationship to the Level of Sperm DNA Fragmentation. Journal of Personalized Medicine. 13. 759. 10.3390/jpm13050759.

[19] Kuan, Valerie & Warwick, Alasdair & Hingorani, Aroon & Tufail, Adnan & Cipriani, Valentina & Burgess, Stephen & Sofat, Reecha & Fritsche, Lars & Igl, Wilmar & Cooke Bailey, Jessica & Grassmann, Felix & Sengupta, Sebanti & Bragg-Gresham, Jennifer & Burdon, Kathryn & Hebbring, Scott & Wen, Cindy & Gorski, Mathias & Kim, Ivana & Cho, David & Heid, Iris. (2021). Association of Smoking, Alcohol Consumption, Blood Pressure, Body Mass Index, and Glycemic Risk Factors With Age-Related Macular Degeneration: A Mendelian Randomization Study. JAMA Ophthalmology. 139. 10.1001/jamaophthalmol.2021.4601.

[20] Oh, Gyu & Ko, Taehoon & Kim, Jin-Hyu & Lee, Min & Choi, Sae & Bae, Ye & Kim, Kyung & Lee, Hae-Young. (2022). Estimation of low-density lipoprotein cholesterol levels using machine learning. International Journal of Cardiology. 352. 10.1016/j.ijcard.2022.01.029.

[21] Garcia-D'Urso, Nahuel & Climent i Pérez, Pau & Sanchez, Miriam & Martí, Ana & Guilló, Andrés & Azorin-Lopez, Jorge. (2022). A Non-Invasive Approach for Total Cholesterol Level Prediction Using Machine Learning. 10.1109/ACCESS.2022.3178419.

[22] Liao, Pen-Chih & Chen, Ming-Shu & Jhou, Mao-Jhen & Chen, Tsan-Chi & Yang, Chih-Te & Lu, Chi-Jie. (2022). Integrating Health Data-Driven Machine Learning Algorithms to Evaluate Risk Factors of Early Stage Hypertension at Different Levels of HDL and LDL Cholesterol. Diagnostics. 12. 1965. 10.3390/diagnostics12081965.

[23] Krentz, Andrew & Haddon-Hill, Gabe & Zou, Xiaoyan & Pankova, Natalie & Jaun, André. (2023). Machine Learning Applied to Cholesterol-Lowering Pharmacotherapy: Proof-of-Concept in High-Risk Patients Treated in Primary Care. Metabolic syndrome and related disorders. 21. 10.1089/met.2023.0009.

[24] Hidekazu, Ishida & Nagasawa, Hiroki & Yamamoto, Yasuko & Fujigaki, Hidetsugu & Doi, Hiroki & Saito, Midori & Ishihara, Yuya & Fujita, Takashi & Ishida, Mariko & Kato, Yohei & Kikuchi, Ryosuke & Matsunami, Hidetoshi & Takemura, Masao & Ito, Hiroyasu & Saito, Kuniaki. (2023). Dataset dependency of low-density lipoprotein-cholesterol estimation by machine learning. Annals of clinical biochemistry. 45632231180408. 10.1177/00045632231180408.

[25] Uysal, Ilhan & Caliskan, Cafer. (2023). Prediction of VLDL Cholesterol Value with Interpretable Machine Learning Techniques. 10.1007/978-3-031-08637-3_6.

[26] Chaudhuri, Avijit. (2023). Prediction of Blood Pressure and Cholesterol By Machine Learning Technique. international journal of engineering technology and management sciences. 7. 10.46647/ijetms.2023.v07i02.007.

[27] .R, Karthikeyan & Geetha, P & E., Ramaraj & Ar, Karthikeyan. (2022). Prediction Of Diabetes And Cholesterol Diseases Based On Ensemble Learning Techniques. 9. 491.

[28] Nath Boruah, Arpita & Biswas, Saroj & Bandyopadhyay, Sivaji. (2022). Transparent rule generator random forest (TRG-RF): an interpretable random forest. Evolving Systems. 14. 10.1007/s12530-022-09434-4.

[29] Latif, Sohaib & Fang, Xian & Arshid, Kaleem & Almuhaimeed, Abdullah & Imran, Azhar & Alghamdi, Mansoor. (2023). Analysis of Birth Data using Ensemble Modeling Techniques. Applied Artificial Intelligence. 37. 10.1080/08839514.2022.2158273.

[30] Dissanayake, Kaushalya & Md Johar, Md Gapar. (2023). Two-level boosting classifiers ensemble based on feature selection for heart disease prediction. Indonesian Journal of Electrical Engineering and Computer Science. 32. 381-391. 10.11591/ijeecs.v32.i1.pp381-391.

[31] Ahmed, Md & Shefaq, Fatima. (2022). A Study on Machine Learning and Supervised and Deep Learning Algorithms to Predict the Risk of Patients: Ten Year Coronary Heart Disease. International Journal of Privacy and Health Information Management. 9. 12. 10.4018/IJPHIMT.305127.

[32] Zapata, Ruben & Huang, Shu & Morris, Earl & Wang, Chang & Harle, Christopher & Magoc, Tanja & Mardini, Mamoun & Loftus, Tyler & Modave, Francois. (2023). Machine learning-based prediction models for home

discharge in patients with COVID-19: Development and evaluation using electronic health records. PLOS ONE. 18. e0292888. 10.1371/journal.pone.0292888.

[33] Handa, Disha & Saraswat, Kajal. (2022). Comparative Analysis of KNN Classifier with K-Fold Cross-Validation in Acoustic-Based Gender Recognition.

[34] Pal, Osim. (2021). Skin Disease Classification: A Comparative Analysis of K-Nearest Neighbors (KNN) and Random Forest Algorithm. 1-5. 10.1109/ICECIT54077.2021.9641120.

[35] Wernigg, Robert & Wernigg, M.. (2022). A case study for assessing the utility of a decision tree based learning algorithm in mental health inpatient care quality management. European Psychiatry. 65. S171-S171. 10.1192/j.eurpsy.2022.454.

[36] Mustamin, Nurul & Aziz, Firman & Firmansyah, Firmansyah & Ishak, Pertiwi. (2023). Classification Of Maternal Health Risk Using Three Models Naive Bayes Method. IJCCS (Indonesian Journal of Computing and Cybernetics Systems). 17. 395. 10.22146/ijccs.84242.

[37] Mentes, Nurettin & Çakmak, Mehmet & KURT, MEHMET. (2023). Estimation of Service Length with the Machine Learning Algorithms and Neural Networks for Patients Who Receiving Home Health Care. Evaluation and Program Planning. 1. 102324. 10.1016/j.evalprogplan.2023.102324.

[38] Dayal, Karan & Shukla, Manmohan & Mahapatra, Satyasundara. (2023). Disease Prediction Using a Modified Multi-Layer Perceptron Algorithm in Diabetes. EAI Endorsed Transactions on Pervasive Health and Technology. 9. 10.4108/eetpht.9.3926.

[39] Tsao, Connie & Vasan, Ramachandran. (2015). Cohort Profile: The Framingham Heart Study (FHS): Overview of milestones in cardiovascular epidemiology. International Journal of Epidemiology. 44. 1800-1813. 10.1093/ije/dyv337.

[40] Rustamov, Zahiriddin & Rustamov, Jaloliddin & Zaki, Nazar & Turaev, Sherzod & Sultana, Most & Tan, Jeanne & Balakrishnan, Vimala. (2023). Enhancing Cardiovascular Disease Prediction: A Domain Knowledge-Based Feature Selection and Stacked Ensemble Machine Learning Approach. 10.21203/rs.3.rs-3068941/v1.

[41] “Cholesterol levels: Medlineplus medical test,” MedlinePlus, https://medlineplus.gov/lab-tests/cholesterol-levels/ (accessed Nov. 6, 2023).