Available online at www.qu.edu.iq/journalcm
JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS
ISSN:2521-3504(online) ISSN:2074-0204(print)

Modeling Cholesterol Levels Using Machine Learning: A Study with the Framingham Heart Study Dataset

Yahya Albugg

Northern Technical University (NTU), iraq\nineveh\mosul\mosul jiddiyda, dr.yahya.albugg@ntu.edu.iq

ARTICLE INFO

Article history:
Received: 23/10/2025
Revised form:25/11/2025
Accepted : 04 /12/2025
Available online:30 /12/2025

Keywords:
Cholesterol
Predictive Model
Feature Importance
Framingham Heart Study
Classification
Risk Factors

ABSTRACT

Cholesterol levels are associated with many health risks, especially cardiovascular disease. Therefore, predicting an individual's cholesterol levels is important to avoid such complications. This paper explores the determinants of blood cholesterol level and builds a machine learning model to predict cholesterol levels using the Framingham Heart Study dataset. Factors such as age, body mass index, and glucose levels were analyzed. The results showed that these factors are the most influential in determining cholesterol levels. Random Forest achieved the highest accuracy in predicting cholesterol levels in the three-level case (78%) and the binary case (88%). These findings indicate that machine learning can effectively identify individuals at risk of elevated cholesterol and highlight the usefulness of focusing on the main influential factors in preventive healthcare and early risk assessment.

https://doi.org/ 10.29304/jqcsm.2025.17.42567

1. Introduction

Cholesterol is an essential biologically active substance for the human body and plays a role in many essential functions. However, increased cholesterol levels may lead to health problems, especially cardiovascular disorders [1][2][3]. Understanding the factors that fluctuate biochemical parameters, especially cholesterol, is important in the fields of community health and medical sciences [4][5]. Cholesterol appears in three main categories within cellular structures: Low-density lipoprotein (LDL), which is referred to as “bad” cholesterol, serves to transport excess cholesterol from the body [6-8]. On the other hand, substances such as low-density lipoprotein (LDL) or “bad” cholesterol are usually involved in the formation of cholesterol plugs on the arteries. Also, VLDL is involved in the accumulation of atheromatous plaques [8] [9]. Developing a model to predict the individual risk level for chronic non-communicable diseases including high cholesterol, which is strongly associated with changes in lifestyle and perceptions, is a major goal in the field of healthcare forecasting [10]. This prediction reflects the potential consequences on an individual's health and, implicitly, the potential to save effort and money within the community's health care system. Existing studies show that, utilizing machine learning methods, it is possible to predict the likelihood of a given patient being hospitalized using only the patient's socioeconomic and behavioral profiles; Without the need for clinical risk factors [11][12][13]. Cholesterol levels are assessed by conducting blood cholesterol tests [14]. Predicting cholesterol levels and subtypes in the absence of clear clinical symptoms

∗Corresponding author yahya

Email addresses: *dr.yahya.albugg@ntu.edu.iq*

Communicated by ‘sub etitor’

associated with hypercholesterolemia provides insight into the likelihood of developing cardiovascular disease [14] [15]. Therefore, ways to prevent or reduce high blood cholesterol levels are ways to reduce the chances of developing cardiovascular disease. Increased blood cholesterol levels are related to many factors such as: gender, age, family history of heart disease, dietary practices, obesity levels, physical activity, alcohol, smoking history, and diabetes [16][17][18][19]. Although many studies have applied machine learning to cholesterol estimation, most of them focus on specific lipid markers or clinical settings. There remains a gap in understanding how AI models can use common demographic and clinical variables to predict overall cholesterol levels in large population datasets such as FHS.

Based on the existing literature and the identified research gap, this study proposes the following research hypotheses: (1) demographic, behavioral, and clinical variables—particularly age, BMI, and glucose—are significant predictors of cholesterol levels in the Framingham Heart Study dataset; and (2) machine learning algorithms, especially ensemble models, can effectively predict cholesterol categories in both binary and multi-class settings. This paper proposes an analysis two main aspects: first, it performs an analysis of the determinants of cholesterol levels by feature importance ratio and identifies factors associated with increased cholesterol risk. Second, predict cholesterol levels using machine learning models on the FHS study dataset from case records. This paper is organized as follows: in Section 2, a review of the literature on cholesterol prediction and its associated factors is presented. Section 3 discusses the methodology used, including data preprocessing steps, feature selection techniques, and machine learning algorithms used. Section 4 presents the, experimental setup, including the parameters used for, training. The results of the experiments are discussed in Section 5. Section 6 discusses the implications of the results and potential limitations of the study. Finally, Section 7 concludes the paper by summarizing the main contributions and findings.

2. Literature Review

A number of researchers have presented methods to model cholesterol levels and associated factors taking advantage of machine learning algorithms. The study [20] dealt with the estimation of low-density lipoprotein cholesterol (LDL-C), in cases of high triglycerides or non-fasting cases. The results showed that machine learning models, specifically those using gradient boosting (LDL-CX) and neural networks (LDL-CN), achieved a higher correlation with directly measured LDL-C compared to traditional methods such as the Friedewald and Martin equations. The machine learning models also showed significantly less bias. , especially in high triglyceride scenarios. In [21], they proposed a model to predict total cholesterol levels using machine learning. To develop the models they used clinical and anthropometric data collected during weight loss interventions. Through cluster analysis, it identifies patient groups with common characteristics that may contain diagnostic information. The results demonstrate the potential of machine learning to predict cholesterol levels with low average absolute error rates. In [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 utilizes a significant dataset and different data mining and machine learning techniques in order to define the complicated relations between risk factors and developing early-stage hypertension. Importantly, the study identifies key predictors of hypertension such as age, body mass index, glucose levels, and C-reactive protein. This study demonstrates how machine learning can be used to define early disease prediction based on data.

Authors in [23], focus on the application of machine learning to the management of dyslipidaemia for high-risk patients receiving lipid lowering medications in a primary care setting. Machine learning algorithms were developed by the authors to support the management of lipids while extracting information from electronic health records (via natural language processing) regarding a patient's medication history. The study demonstrates that machine learning has the ability to identify suboptimal prescription behaviours, locate high-risk patients for closer monitoring, and provide evidence-based recommendations for therapeutic alternatives. However, the study focuses primarily on optimising medication management without examining the impact of other factors that might also affect the cholesterol levels of an individual. In [24], a method based on machine learning has been used to estimate LDL-C levels, and the paper discusses the effect of characteristics of the training datasets used for that purpose. The study demonstrates the significance of the training dataset characteristics for achieving credible estimates. In order to predict low-density lipoprotein cholesterol (VLDL-C), the authors [25] conducted an implementation study using interpretable machine learning techniques based on features of age, gender, and laboratory measures. They determined that the generalized linear model (GLM) yielded the best results among the different approaches. In paper [26], a model for predicting blood pressure and cholesterol was created using machine learning and multiple linear regression analysis (MRA.) The researchers indicated that machine learning models were successful in predicting outcomes for the presence of blood pressure and high cholesterol conditions. In [27], the authors analyzed the use of ensemble learning techniques to predict diseases such as diabetes and cholesterol. The study

emphasized that ensemble learning algorithms performed better than individual algorithms such as Adaboost, Random Forest, Bagging, Voting, and Stacking.

The existing studies suggest that there are possibilities in machine learning for predicting some of the factors, and possibly disease, related to cholesterol. However, there is little comprehensive understanding of how features can predict cholesterol levels and how far it could predict cholesterol levels based on daily features and behaviors as discussed in the FHS data set.

3. Materials and Methods

To predict cholesterol levels in the Framingham Heart Study (FHS), a set of machine learning models was used and tested. Initially, cholesterol levels were divided into levels and a set of models were developed. Finally, it was tested. This section describes the dataset, models used, and breakdowns of cholesterol levels.

3.1. Machine learning model

Various machine learning models have been used to analyze factors affecting blood cholesterol levels. Selecting the appropriate learning algorithm for a given dataset and case study is crucial. An experimental approach that involves testing and evaluating multiple algorithms is usually the preferred methodology to ensure the optimal approach is selected. To achieve this, a group of machine learning algorithms such as (Random Forest, Gradient Boosting, Support Vector Classifier, K-Nearest Neighbors, Decision Trees, Naive Bayes, and Multi-Layer Perceptron) used in the literature were tested, with the aim of determining the most appropriate algorithm. The dataset was divided into input features (X) and the target variable (y), where X represents the clinical and demographic factors used for prediction and y represents the cholesterol category. The data was then split into training and testing sets so the model learns from one portion of the data and is evaluated on unseen samples to measure its performance.

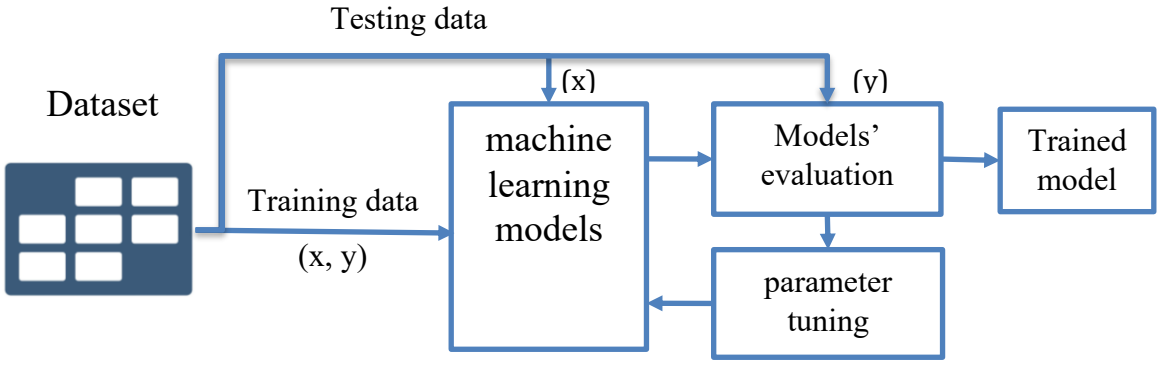

Fig. 1 - framework of machine learning models

Random Forest (RF) is a state-of-art machine learning algorithm that 'bootstraps' several decision trees to enhance the generalisations ability of the resulting model and thereby handle over fitting of data. It is especially used in classification problems [28-29]. Gradient Boosting (GB) is another type of ensemble method that construct decision trees in sequential manner in which each tree tries to minimize the error made in the prior trees [30]. Support Vector Classifier (SVC) is a form of supervised learning that can handle classification. This method operates by identifying the plat that has the greatest distance between various classes and hence it is appropriate for both the binary and multiclass classification issues [31-32]. The K-Nearest Neighbors (KNN) is easy as it is one of the simplest classifiers. Of the K-nearest neighbors, it categorises a point by the majority class which the point belongs to. It is particularly useful when working in the local affine space of the patterns in the data [33][34]. Decision Trees (CT) is one of the simple types of ML. It divided the data into branches according to a feature value and make decision tree structures [35][29].
Naive Bayes (NB) It is a probabilistic classifier of a rather type that provides predictions based on the Bayes theorem [36]. Multi-Layer Perceptron (MLP) is a type of ANN which is layered and comprises of multiple nodes also known as neurons. Classification is one of the possible tasks it can be used in, as well as in many others [37][38]. These models were trained for the purpose of predicting cholesterol levels given the dataset given with different attributes like gender, age, education, smoking history, blood pressure.

3.2. Feature importance calculation

For each classification setting (three-level and binary cholesterol categories), feature importance scores were derived from the Random Forest model. The Random Forest computes an importance score for each feature based on the mean decrease in Gini impurity: at every split in every tree, the reduction in impurity attributed to a given feature is accumulated and weighted by the number of samples in the corresponding node. After training, the total contribution of each feature is averaged over all trees in the forest and normalised so that the importance scores across all features sum to 1.

3.3. Data and preprocessing

The models were developed and tested in this paper using data from the Framingham Heart Study (FHS) [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 (Accurate Definition)	Type
male	Sex of participant (1 = male, 0 = female)	Binary
age	Age of participant at exam	Continuous
education	Education level (1–4 categorical scale)	Categorical
currentSmoker	Whether participant currently smokes	Binary
cigsPerDay	Number of cigarettes smoked per day	Continuous
BPMeds	Using blood pressure medication at baseline	Binary
prevalentStroke	Presence of stroke prior to the baseline exam	Binary
prevalentHyp	Presence of hypertension at baseline	Binary
diabetes	Physician-diagnosed diabetes	Binary
totChol	Total cholesterol (mg/dL)	Continuous
sysBP	Systolic blood pressure (mmHg)	Continuous
diaBP	Diastolic blood pressure (mmHg)	Continuous
BMI	Body Mass Index (kg/m²)	Continuous
heartRate	Heart rate (beats per minute)	Continuous
glucose	Fasting glucose level (mg/dL)	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 that were less than or equal to 200 mg/dL were termed 'Desirable', values between 201 mg/dL and 239 mg/dL were termed 'Borderline high', and values greater than or equal to 240 mg/dL were classified as 'High' [40]. The threshold limit for the three different types of cholesterol level was selected based on the National Library of Medicine (NIH) [41]. In the case of binary classification, 'normal' was defined as values less than or equal to 200 mg/dL and 'high' classified values in the 201 mg/dL to 239 mg/dL range.

The FHS dataset contains several variables with missing values. In this study, missing records were handled by removing any rows that contained missing values in the selected predictors or in the target variable, in order to ensure consistency of the training samples and avoid introducing bias from imputation.

Table 2 - Total Cholesterol Level Classification

Total Cholesterol Level	Class
Less than 200mg/dL	Desirable
201-239 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.

4. 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. To ensure reproducibility and to avoid model overfitting, all machine-learning algorithms were trained using the default hyperparameters available in the scikit-learn implementation.

4.1. Feature Importance

Feature importance is analyzed to identify the variables that have an impact on blood cholesterol levels. Feature importance has the purpose of identifying the most effective input variables in providing an estimate of cholesterol levels and is of importance to the healthcare profession and researchers in general as well as to identify the input features for a prediction model. During classification, cholesterol levels were first divided into three categories namely; “Desirable”, “Borderline High” and “High” based on certain thresholds. The average importance scores according to this three-level classification are shown in Figure 2 based on the selected features.

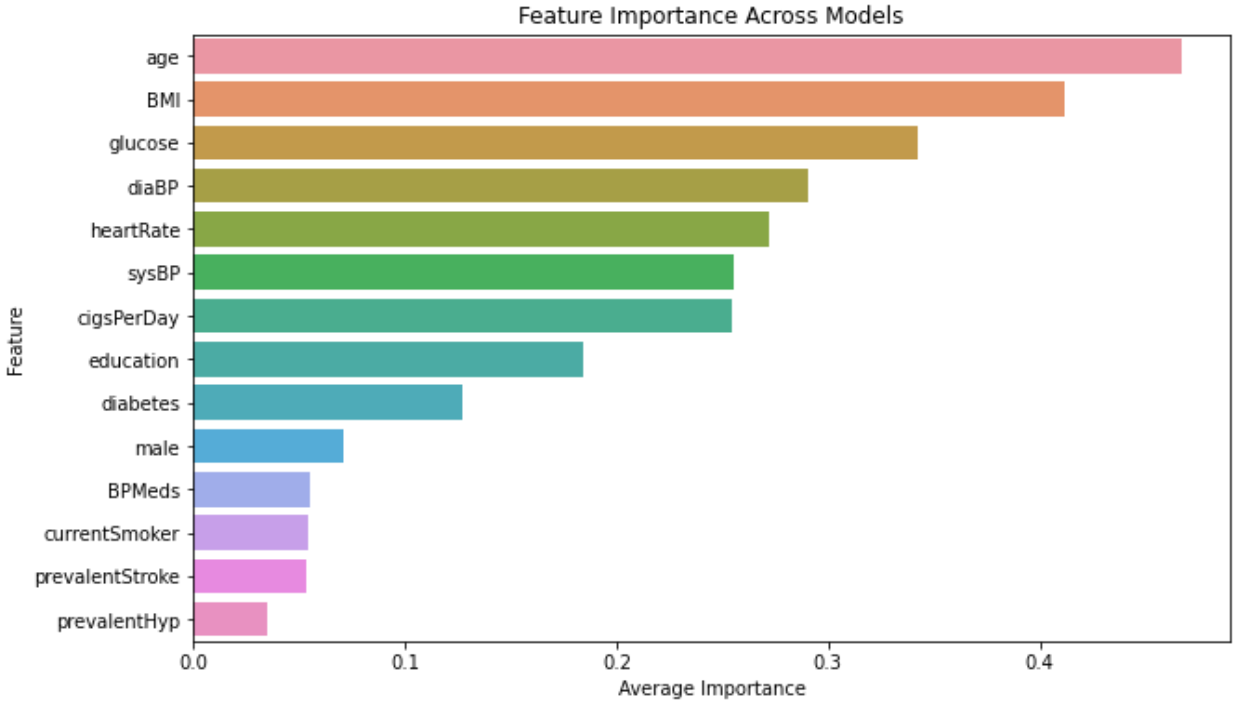

Fig. 2 - Average Importance based on Three-level Classification

As depicted in fig 1, age, BMI, and glucose levels are very influential variables for predicting cholesterol levels. They are among the most significant predictors of whether cholesterol levels fall within the “desirable,” “borderline high,” or “high” categories. In addition, factors such as diastolic blood pressure (diaBP), heart rate, systolic blood pressure (sysBP), and average number of cigarettes smoked per day (cigsPerDay) showed importance in this classification.

With respect to a two-level classification of “normal” and “high” cholesterol levels, figure 3 presents the average importance scores for features selected for the children in this classification.

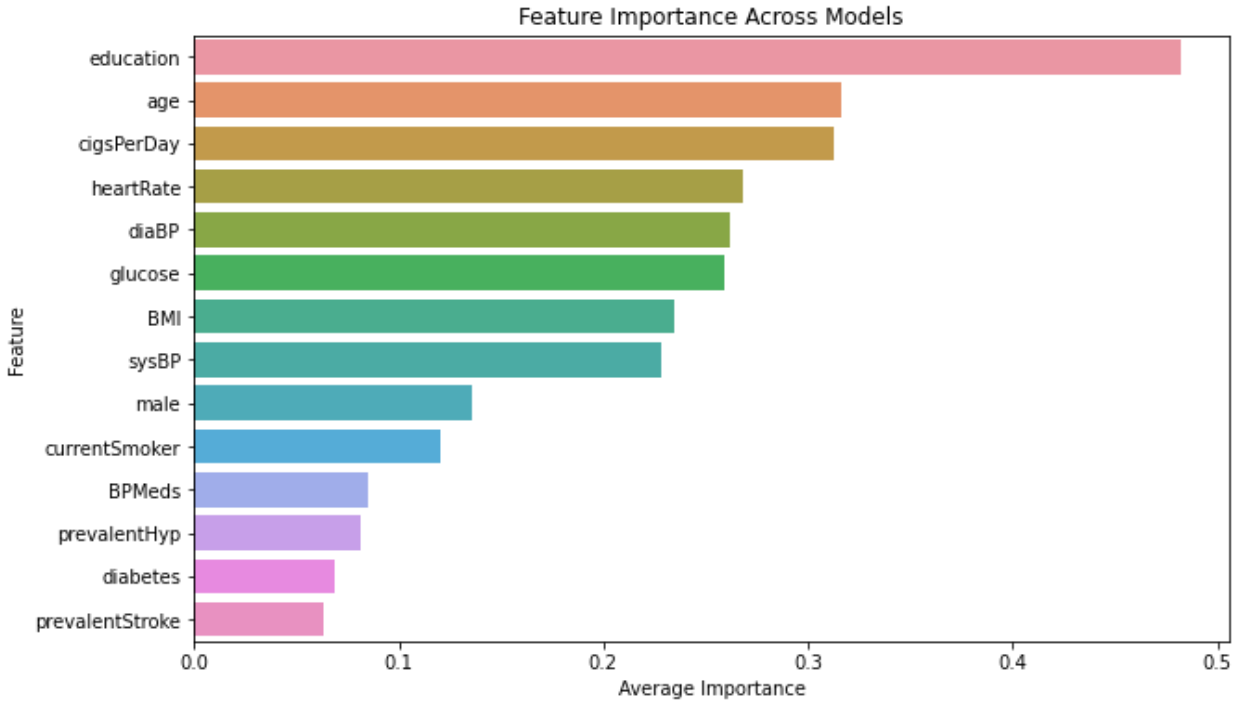

Fig 3 - Average Importance based on Two-level Classification

Figure 2 illustrates that education, age, and average numbers of cigarettes smoked per day (cigsPerDay) are the most influential features to differentiate between the "normal" and "high" cholesterol classification. These findings highlight the importance of education level in the prediction of an individual at risk for "high" cholesterol levels.

4.2. Model performance

The predictive abilities of the chosen machine learning models to forecast cholesterol readings were compared against one another using precision, recall, F1 Score and Accuracy (ACC) metrics across both 3 classifications and 2 classifications (shown in Table 3).

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 information provided in Table 3, Random Forest had a combination of both the highest ACC and REC values for the first level ("high") cholesterol classification category indicating that the Random Forest model predicted high

cholesterol patients efficiently. Additionally, Gradient Boosting had the greatest balance between precision and recall, with the highest F1 Score indicating the highest degree of accuracy in being able to identify high cholesterol patients. Support Vector Classifier achieved the highest precision for identifying patients as having “high” cholesterol, therefore supporting the claim of this model providing the most precise selection of patients with high cholesterol levels.

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 cholesterol classification task with two levels, the performance across all models was improved, with Support Vector Classifier (SVC) and K-Nearest Neighbors (KNN) demonstrating high precision stating that an individual has "high" cholesterol, indicating that these models did well in detecting an individual with high cholesterol. The Random Forest (RF) and Gradient Boosting (GB) models achieved high F1 scores on their individual classes and for both the "desirable" and "high" categories due to having a relatively high degree of precision as well as recall rates being approximately equal between the two classes for most instances. Conversely, the Naive Bayes (NB) model showed much lower levels of recall for high-value cholesterol, meaning it produced many more false negatives than all of the other models.

5. Discussion

The variable influence assessment in this project produced important conclusions regarding what affects cholesterol. Across both three and two levels of cholesterol classification, age, BMI, and glucose were shown to have the largest impact on cholesterol. Healthcare providers can use this information to develop interventions directed toward specific individuals with high cholesterol. For example, many individuals classified as having high glucose or BMI will likely benefit the most from lifestyle changes to improve cholesterol. This serves as an example of how understanding the variability between people and applying personalized approaches to medicine will enable healthcare professionals to make better informed decisions concerning cholesterol management.

The machine-learning analyses of predicting cholesterol levels based on classification tasks revealed very different comparative performance metrics, depending on the scenarios involved. Some machine-learning algorithms demonstrated good metrics (i.e., precision, recall, and F1 score) for predicting cholesterol levels in some scenarios, while in other scenarios, as indicated by some of the highest precision metrics, support vector classifier (SVC) and k-nearest neighbours (KNN) performed better. Because there were significant differences, a potential opportunity exists to further refine these models and the subsequent intervention in the context of healthcare settings.

The predictive outcomes derived from the machine-learning models have significant clinical implications for treatment decisions. Being able to identify those patients most likely to fall into the “high” cholesterol category allows healthcare providers to implement earlier targeted interventions (i.e., lifestyle modification, dietary counselling, or additional laboratory investigations) before the development of complications such as atherosclerosis or cardiovascular disease. The strong correlation of age, body mass index (BMI), and glucose levels with model predictions corresponds with current clinical data. Thus, the models utility to stratify patients based on predicted cholesterol levels supports preventive medicine initiatives, can be used to improve the care process, assist with resource allocation, and help to prioritise patients who are at greater risk of dyslipidaemia and cardiovascular events.

6. Conclusion

The investigation found a range of factors to consider for informing the measurement of cholesterol levels and for developing models. Age, body mass index, and blood glucose level were the three most important predictors of cholesterol

regardless of whether the prediction model was built on two or three different cholesterol values. Therefore, these three predictors should be included in any health-related efforts or any personalized medical efforts that would aim to reduce risk for cardiovascular disease. When comparing the models for predicting cholesterol levels, it became apparent that there was a wide range of performance for each of the different types of models used (e.g., Random Forest and Gradient Boosting performed better than other models for predicting cholesterol levels with this application, while SVCs and K-nearest Neighbors performed well for some cholesterol levels). As such, it is clear that continued development of predictive models will enhance the capability of future models to provide accurate cholesterol prediction.

References:

[1] Huff, Trevor & Brandon Boyd & Ishwarlal Jialal. (2023). Physiology, Cholesterol. In StatPearls. StatPearls Publishing. Available from: 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] AL-Shamdeen, Muna & Ramo, Fawziya. (2024). PERFORMANCE EVALUATION FOR FACE MASK DETECTION BASED ON MULT MODIFICATION OF YOLOV8 ARCHITECTUREOCENA WYDAJNOŚCI WYKRYWANIA MASKI NA TWARZY NA PODSTAWIE WIELU MODYFIKACJI ARCHITEKTURY YOLOV8. Informatyka, Automatyka, Pomiary w Gospodarce i Ochronie Środowiska. 14. 89-95. 10.35784/iapgos.6056.
[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).