ECUTBILDNING

Data Science
Thesis

A Simulation Study for Choosing Predictive Insurance Model

Cynthia Uzomba
Abel Shiferaw

Abstract

Background: Machine learning models in the regression analysis provide insight to determine the predictive insurance model.
Objective: The aim of the study is to suggest a predictive insurance model that predicts the cost of consultation to enhance the chance of achieving insurance for patients in Nigeria and hence increase doctor consultation via Carelyo across time.
Research steps: A literature review on regression and four selected regression models namely Multiple Linear Regression, Random Forest, K-Nearest Neighbors, Least Absolute Shrinkage and Selection Operator is presented moreover, the health care company Carelyo is introduced. The project was carried out in six stages, beginning with identifying the data needed for analysis and ending with using the populated data for analysis and modeling. The project used manual data entry as the data population method, and the populated data was validated for completeness, consistency, and correctness. The data was then used to develop an insurance model.
Findings: The performance of the models was assessed based on small mean square error (MSE) and high r-square score (R^2) provided that small MSE is prioritized.
Discussion: Finally, a suggestion was made to choose KNN as a fit insurance model to predict the target (cost of consultation) in the simulated dataset.

Key words: Regression, Multiple Linear Regression, Random Forest, K-Nearest Neighbors, Least Absolute Shrinkage and Selection Operator, Data Simulation, Mean Square Error, R-square Score, Model fitness.

Table of content

1. Introduction

Regression models are an important tool in data analysis, providing a way to establish the relationship between two or more variables. The models allow us to predict the outcome of a dependent variable based on the independent variables, enabling us to make informed decisions and identify trends. The process of regression analysis involves examining the relationship between a dependent variable and one or more independent variables. The dependent variable is the variable that is being predicted or explained, while the independent variable is the variable that is being used to make the prediction (Collins & Moons, 2019). For example, in a healthcare setting, we may want to predict a patient's health outcome based on their age, gender, and medical history. The dependent variable in this case would be the patient's health outcome, while the independent variables would be their age, gender, and medical history.

Regression models can be categorized into two main types: linear and nonlinear. Linear regression models assume a linear relationship between the dependent and independent variables, while nonlinear regression models assume a nonlinear relationship. The choice of regression model depends on the nature of the data being analyzed and the research question being asked (Chicco et al., 2021). In recent years, machine learning regression models have gained popularity due to their ability to handle large datasets and complex relationships between variables (Speiser et al., 2019). Random Forest, MLR (Multiple Linear Regression), KNN (K-Nearest Neighbors), and LASSO (Least Absolute Shrinkage and Selection Operator) are some of the popular machine learning regression models used in data analysis. Random Forest is a tree-based model that works by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees (Chowdhury & Turin, 2020). MLR is a linear model that assumes a linear relationship between the dependent and independent variables. KNN is a non-parametric algorithm that classifies new data points based on the most similar training examples (Leisman et al., 2020). LASSO is a model that selects a subset of the most important variables, shrinking the coefficients of less important variables to zero. Predictive model analysis is an important technique used in various industries, including healthcare, finance, and marketing, to forecast customer behavior, detect fraud, or identify trends.

Predictive models are built using data mining and machine learning algorithms and can be used to make predictions based on historical data. The models are trained using historical data and then used to make predictions on new data (Chukhrova & Johannssen, 2019). The accuracy of the predictions depends on the quality and quantity of the data used to train the model.

In general, regression models are an essential tool in data analysis, enabling us to predict the outcome of dependent variables based on independent variables. Machine learning regression models such as Random Forest, MLR, KNN, and LASSO are widely used in data analysis due to their ability to handle large datasets and complex relationships between variables (Wynants et al., 2020). Predictive model analysis is a powerful technique used to make predictions based on historical data, and it has many applications in various industries.

2. The health care company: Carelyo

Carelyo is a health company founded with the aim of providing easy, affordable, and accessible healthcare services to people. The company was established by a team of medical professionals and software engineers who noticed the growing need for a better healthcare system.
One of Carelyo's primary goals is to revolutionize the healthcare industry by providing quality healthcare services to everyone, regardless of their socioeconomic status. The company aims to bridge the gap between healthcare and technology by using innovative solutions to deliver personalized and convenient healthcare services to its clients. Carelyo believes that healthcare should be a basic right, and everyone should have access to affordable healthcare services.
Carelyo offers a range of healthcare services, including easy consultation booking, drug prescription, hospital referencing, and laboratory integration. The company's consultation booking service allows patients to book a consultation with a medical professional in just five seconds. Patients can do this from wherever they are, making it convenient for them to receive medical care without having to leave their homes (SWEDCON, 2023). The company's drug prescription service allows patients to receive their medication instantly on their dashboard after their doctor prescribes it. Patients can buy their medication online with discounts or visit accredited pharmacies close to them. The company's hospital referencing service allows medical professionals to refer patients to the right specialist at a clinic or hospital nearest to them, saving patients time and money. Finally, the laboratory integration service helps identify health symptoms early, which is vital for preventing costly treatment of chronic illness, dependency on lifelong medication, or sudden death in the worst-case.
Carelyo's mission is to improve the quality of healthcare services by providing innovative solutions that make healthcare more accessible and affordable. The company believes that technology can play a significant role in improving the healthcare system, and it is committed to using technology to provide quality healthcare services to its clients. Carelyo is also committed to providing a safe and secure platform for its clients to receive medical care. The company uses advanced security measures to ensure that its clients' personal and medical information is protected. Carelyo's success can be attributed to its dedicated team of medical professionals and software engineers who work tirelessly to provide quality healthcare services to its clients. The company's team is made up of doctors, nurses, pharmacists, and other healthcare professionals who have years of experience in their respective fields. The software engineers are also experts in their field, and they use their knowledge and expertise to create innovative solutions that make healthcare more accessible and affordable.
In general, Carelyo is a healthcare company committed to improving the quality of healthcare services by providing innovative solutions that make healthcare more accessible and affordable. The company's mission is to bridge the gap between healthcare and technology by using technology to provide quality healthcare services to its clients. Carelyo offers a range of healthcare services, including easy consultation booking, drug prescription, hospital referencing, and laboratory integration. The company's success is due to its dedicated team of medical professionals and software engineers who work tirelessly to provide quality healthcare services to its clients.

3. Data Simulation

Data simulation is the process of generating or populating data in order to create a representative sample of data for analysis, testing, or training purposes (DeBruine & Barr, 2021). In the case data analysis at Carelyo internship, data simulation was used to populate the patient, doctor, and consultation tables in MySQL to create a dataset for analysis and model creation. There are several techniques for data simulation, such as random data generation, copying existing data, and synthetic data generation. In this case, the data was populated using Excel to create the required datasets in CSV format, which were then imported into MySQL. In general, steps involved in the data simulation process is explained as follows:

I. Identify the data needed for analysis:

The first step in data simulation is to identify the specific data required for analysis. This step involves careful consideration of the research or business problem at hand, the objectives of the analysis, and the expected outcomes. It's important to have a clear understanding of the variables that are most relevant to the analysis and any potential limitations or biases that may be present. For instance, in Carelyo internship, identifying the necessary data involved determining the patient, doctor, and consultation variables that were most important for creating an insurance model (Smolak et al., 2020). This required careful consideration of factors such as patient age, consultation time, amount paid by patient, and years of experience of the doctors.
Identifying the relevant data also involves determining the appropriate data sources and methods for collecting or generating the data. This may include manual data entry, data collection through surveys or interviews, or the use of existing databases or data repositories. Overall, the identification of data is a crucial step in the data simulation process, as it sets the foundation for the accuracy and effectiveness of the subsequent data simulation steps.

II. Determine how the data will be populated:

The second step in data simulation is to determine how the identified data will be populated. This step involves considering the available data population methods and selecting the most appropriate one for the specific data and analysis objectives. As highlighted above, there are several data population methods available, including manual data entry, copying existing data, random data generation, and synthetic data generation. Each method has its own advantages and limitations, and the appropriate method will depend on the specific data and analysis objectives. In this specific analysis, the data was populated using Excel to create the required datasets in CSV format. This involved manually inputting data into the Excel spreadsheet, as well as using formulas and copying and pasting data from other sources. The decision to use manual data entry was likely influenced by the limited size of the dataset and the availability of the necessary data sources (Hu et al., 2019). Other factors that may influence the data population method include the level of accuracy and completeness required, the available resources and tools, and the potential for bias or errors in the data. Overall, the determination of how the data will be populated is an important step in the data simulation process, as it impacts the accuracy, completeness, and suitability of the resulting dataset for analysis and modeling.

III. Choose the data population method:

The third step in data simulation is to choose the appropriate data population method based on the data requirements and analysis objectives. This involves considering the advantages and

disadvantages of each data population method and selecting the method that best meets the needs of the analysis. The most common data population methods include manual data entry, copying existing data, random data generation, and synthetic data generation (Tawfik et al., 2019). Each method has its own unique strengths and limitations, and the choice of method will depend on factors such as the size and complexity of the dataset, the desired level of accuracy and completeness, and the available resources and tools.

The decision to use manual data entry to populate the dataset may have been influenced by the limited size of the dataset and the availability of the necessary data sources. However, for larger and more complex datasets, using random or synthetic data generation may be more efficient and accurate. Other factors that may influence the choice of data population method include the potential for bias or errors in the data, the need to preserve data privacy and security, and the potential impact on the accuracy and reliability of the resulting analysis and models. Overall, the choice of data population method is a critical step in the data simulation process, as it directly impacts the quality and suitability of the resulting dataset for analysis and modeling.

IV. Populate the Data:

The fourth step in data simulation is to populate the dataset using the selected data population method. This involves executing the chosen method to generate or input data into the appropriate fields in the dataset. In Carelyo internship, after choosing manual data entry as the data population method, the actual population process would involve entering data such as patient age, consultation time, amount paid by patient, and years of experience of the doctors into the Excel spreadsheet. This would have been done in a systematic and organized way to ensure the accuracy and completeness of the dataset. During the data population process, it's important to check for errors, inconsistencies, and outliers that may affect the accuracy and reliability of the dataset (Smolak et al., 2020). Any errors or inconsistencies should be corrected to avoid issues later in the analysis and modeling process.

After populating the dataset, it's also important to ensure that it meets the requirements and expectations of the analysis objectives. This may involve conducting initial exploratory analysis to identify any patterns or trends in the data and making adjustments or modifications to the dataset as needed. Overall, the population of the data is a crucial step in the data simulation process, as it directly affects the accuracy and suitability of the dataset for subsequent analysis and modeling.

V. Validate the Data:

The fifth step in data simulation is to validate the populated data to ensure that it meets the quality and accuracy standards required for the analysis and modeling involves checking the data for completeness, consistency, and correctness, and addressing any issues that may be identified. The validation process may involve a combination of manual and automated techniques. For example, manual validation may involve reviewing the dataset for any missing or inconsistent data and verifying the accuracy of the data against known sources. Automated validation techniques may include using algorithms or statistical tests to identify any anomalies or inconsistencies in the dataset. After identifying any issues or errors in the dataset, the next step is to address them. This may involve correcting errors in the data, removing outliers, or imputing missing data. The goal is to ensure that the dataset is as complete and accurate as possible, and that it meets the requirements and expectations of the analysis objectives (Greasley & Edwards, 2021). Overall, the validation of the populated data is a critical step in the data simulation

process, as it directly affects the reliability and validity of subsequent analysis and modeling. It is important to devote sufficient time and resources to this step to ensure that the resulting dataset is of the highest quality and accuracy possible.

VI. Use the Populated Data for Analysis and Modeling:

The sixth and final step in data simulation is to use the populated data for analysis and modeling. This involves applying the appropriate data analysis techniques and modeling tools to extract insights and information from the dataset. In my Carelyo internship, this involved using Jupyter Notebook to perform exploratory data analysis and create models for the insurance policy which is covered in the next topic (DeBruine & Barr, 2021). The insights and information gained from this analysis can then be used to make informed decisions and optimize the performance of the insurance policy.

It's important to note that the analysis and modeling process is iterative and ongoing and may require revisiting previous steps in the data simulation process as new insights and information are gained. For example, if issues or errors are identified in the dataset during the analysis process, it may be necessary to go back to the validation step to correct the issues and improve the accuracy and reliability of the dataset. Overall, the use of the populated data for analysis and modeling is a critical step in the data simulation process, as it enables the extraction of meaningful insights and information that can drive informed decision-making and optimization. It's important to ensure that the analysis and modeling process is transparent, replicable, and grounded in sound data practices to ensure the reliability and validity of the resulting insights and decisions.

4. Predictive Analysis

4.1. Pre-process: Data preparation

At the early stage of data preparation there were three populated datasets namely patient.csv with 250 rows and 29 columns, doctor.csv with 214 rows and 22 columns and consultation.csv with 520 rows and 17 columns. Next up, duplicates and missing values were checked for each of the populated dataset, fortunately there were no duplicates. Missing values were dropped since imputing medical data without proper medical examination would be inappropriate.

From an existing feature (date_of_birth) in patient dataset a new variable (age) was created and from the existing variable (graduation_date) in the doctor dataset a new variable (year_of_experience) was generated.

The simulated data is generated by merging the cleaned patient, doctor, and consultation datasets with list of explanatory variables namely consult_id, patient_anatomy, patient_consult_status, patient_community, patient_marital_status, patient_has_children, patient_has_dependent, patient_num_of_children, patient_num_of_dependents, patient_age, patient_gender, consultant_doctor_experience and response variable namely consult_cost.

At the final stage of data preparation, all the categorical variables in the simulated dataset were transformed to numerical features using dummy transformation and was summarized with tables and distribution plots. In addition, the scale of correlation among the explanatory variables and response variable was studied via correlation matrix plot moreover, log transformation of the response variable was carried out to reduce outlier and mean square error in the latter analysis while, the explanatory variables were standardized via robust scaling.

4.2. Method: Modeling

For the prediction of the target (consult_cost), the four machine learning models (RF, MLR, KNN, and LASSO) explained above were considered with an 80% train and 20% simulated dataset split. The strength of the predictive models was evaluated based on the value of mean square error (MSE) and r-square score (R^2) in such a way that the model with smallest MSE (closer to zero) and highest R^2 (closer to 1) is chosen as suitable insurance model.
Remark: Model with small MSE is preferred (prioritized) despite high R^2 score.

4.3. Results

The empirical findings are represented in the form of tables and graphs as follows.

Table 1. Mean square error (MSE) of the predictive models for train and test simulated dataset.

	MLR	RF	KNN	LASSO
train _MSE	0.048099	0.001089	0.0	0.059861
test _MSE	0.053425	0.011827	0.200511	0.060166

Fig 1. Horizontal bar graphs showing MSE for train and test simulated dataset.

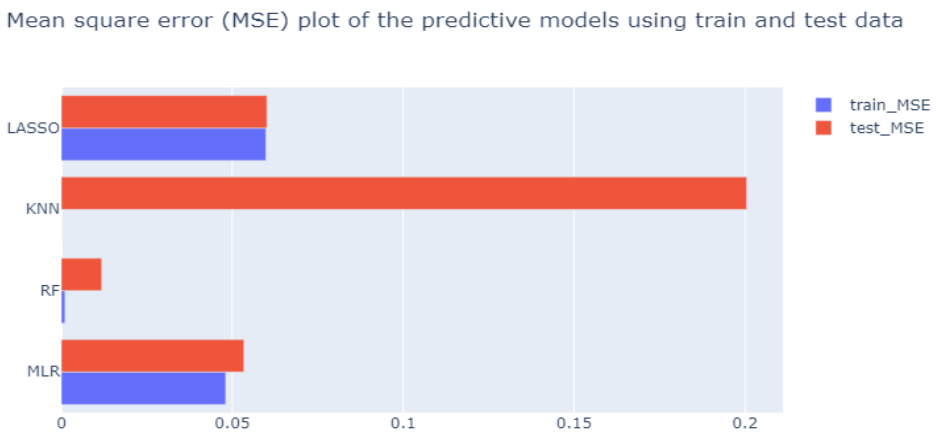
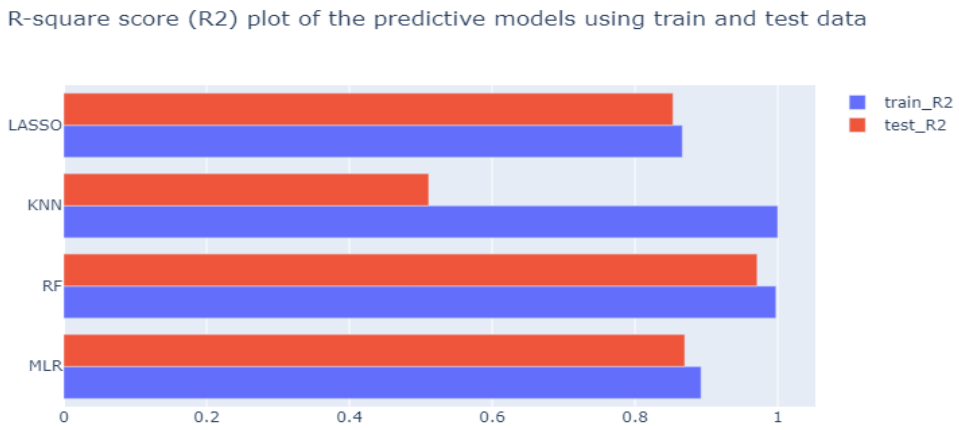


Table 2. R-square score (R^2) of the predictive models for train and test simulated dataset.

	MLR	RF	KNN	LASSO
train_R^2	0.892662	0.99757	1.0	0.866414
test_R^2	0.869761	0.971167	0.511192	0.853327

Fig 2. Horizontal bar graphs showing R^2 for train and test simulated dataset.



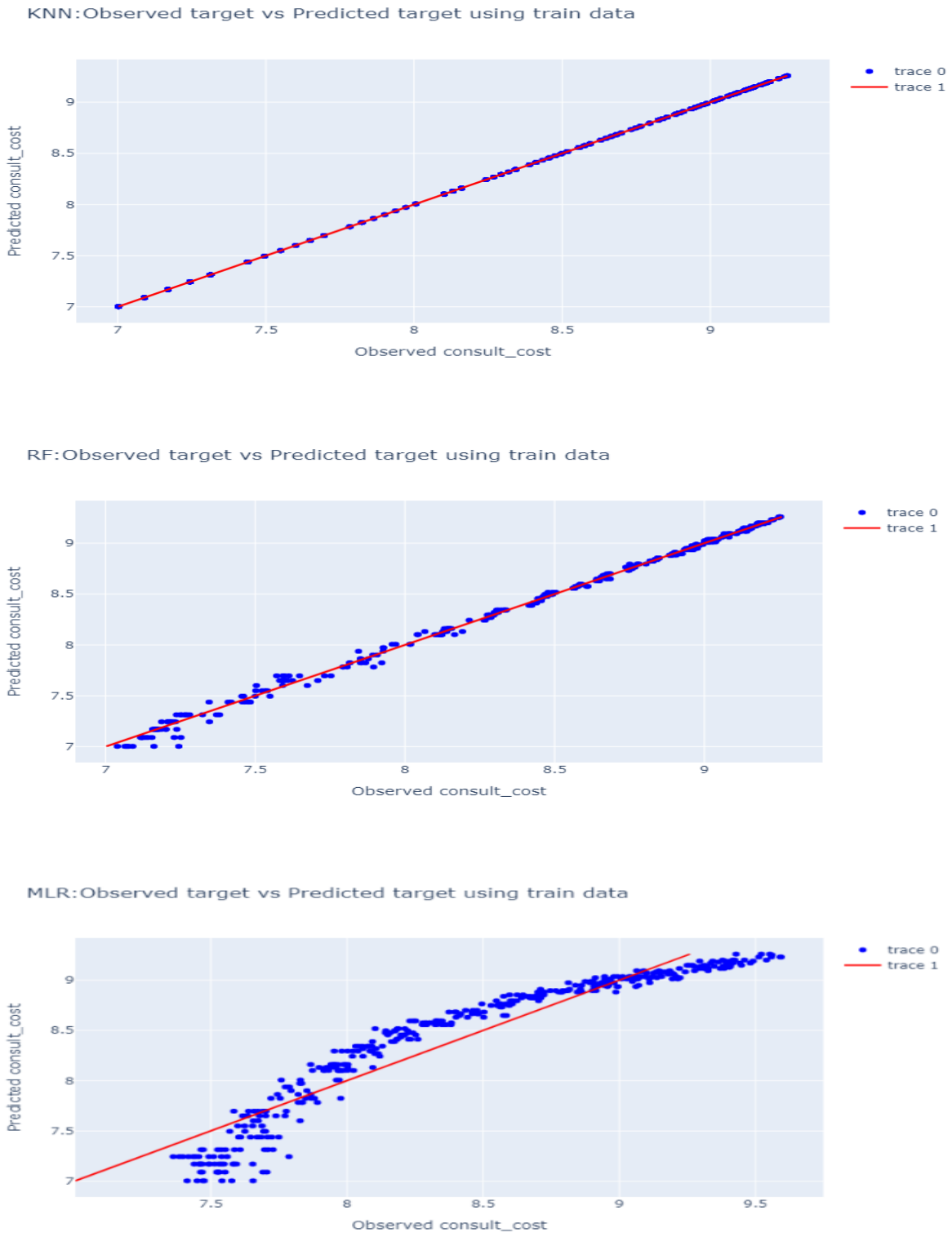
4.3.1. Mean Square error (MSE)

From Table 1., KNN model performed best with, 0.0, MSE when predicting the target with train dataset which revealed the consistency (convergence in probability of estimate target to actual target with larger sample size) of KNN with the training simulated dataset. RF also performed well in terms of consistency with MSE of, 0.001089, using the training dataset. MLR and LASSO are less consistent as compared to RF with relatively higher train MSE of, 0.048099, and, 0.059861, respectively.
Again, from Table 1., it is showed that KNN is exposed to overfitting as it performed significantly weak for the test dataset. Moreover, RF seems better related to overfitting problem than KNN. In addition, MLR and LASSO showed a very slight overfitting (or no overfitting) as both models provided nearly the same prediction for both the train and test simulated dataset.

4.3.2. R-square score (R^2)

From Table 2., An r-square score (R^2 = 1.0) of the KNN model for the training simulated dataset shows that all the variation of the target (spread of observed target around mean target) is explained (predicted) by the variability (variation) of the explanatory variables in the KNN model while, for the test dataset in KNN, R^2 = 0.511192, showing that only approximately 51% of the variability in the target is explained by the variability of independent variables which is significantly less compared to the explained variation in the train dataset. For the remaining three models there is no significant difference in terms of explained variability for the train and test dataset. In general, fitness of the models that predicted the target (consul_cost) with the train simulated dataset is shown below.

Fig 3. The fitness of the four predictive models for train simulated dataset.



Discussion

For suggesting the predictive insurance model among the four regression models, the empirical findings of the current study and previous related studies was considered. The reason for considering earlier machine trained predictive analysis is to overcome the limitation this study faced in connection with available real data. Hence, by recalling the performance of KNN model in predicting the target with proper train dataset and the fitness of KNN seen in the result section of the this study, KNN can be suggested the most suitable predictive insurance model.
In general, the results of this project have significant implications for Carelyo's insurance policy, as the insights gained can be used to optimize performance and make informed decisions regarding patient care and treatment.

References

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Computer Science, 7, e623. https://doi.org/10.7717/peerj-cs.623

Chowdhury, M. Z. I., & Turin, T. C. (2020). Variable selection strategies and its importance in clinical prediction modelling. Family Medicine and Community Health, 8(1), e000262. https://doi.org/10.1136/fmch-2019-000262

Chukhrova, N., & Johannssen, A. (2019). Fuzzy regression analysis: Systematic review and bibliography. Applied Soft Computing, 84, 105708. https://doi.org/10.1016/j.asoc.2019.105708

Collins, G. S., & Moons, K. G. M. (2019). Reporting of artificial intelligence prediction models. The Lancet, 393(10181), 1577–1579. https://doi.org/10.1016/S0140-6736(19)30037-6

DeBruine, L. M., & Barr, D. J. (2021). Understanding Mixed-Effects Models Through Data Simulation. Advances in Methods and Practices in Psychological Science, 4(1), 2515245920965119. https://doi.org/10.1177/2515245920965119

Greasley, A., & Edwards, J. S. (2021). Enhancing discrete-event simulation with big data analytics: A review. Journal of the Operational Research Society, 72(2), 247–267. https://doi.org/10.1080/01605682.2019.1678406

Hu, T., Tang, T., & Chen, M. (2019). Data Simulation by Resampling—A Practical Data Augmentation Algorithm for Periodical Signal Analysis-Based Fault Diagnosis. IEEE Access, 7, 125133–125145. https://doi.org/10.1109/ACCESS.2019.2937838

Leisman, D. E., Harhay, M. O., Lederer, D. J., Abramson, M., Adjei, A. A., Bakker, J., Ballas, Z. K., Barreiro, E., Bell, S. C., Bellomo, R., Bernstein, J. A., Branson, R. D., Brusasco, V., Chalmers, J. D., Chokroverty, S., Citerio, G., Collop, N. A., Cooke, C. R., Crapo, J. D., … Maslove, D. M. (2020). Development and Reporting of Prediction Models: Guidance for Authors From Editors of Respiratory, Sleep, and Critical Care Journals. CriticalCare Medicine, 48(5), 623–633. https://doi.org/10.1097/CCM.0000000000004246

Smolak, K., Rohm, W., Knop, K., & Siła-Nowicka, K. (2020). Population mobility modelling for mobility data simulation. Computers, Environment and Urban Systems, 84, 101526. https://doi.org/10.1016/j.compenvurbsys.2020.101526

Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. Expert Systems with Applications, 134, 93–101. https://doi.org/10.1016/j.eswa.2019.05.028

Tawfik, G. M., Dila, K. A. S., Mohamed, M. Y. F., Tam, D. N. H., Kien, N. D., Ahmed, A. M., & Huy, N. T. (2019). A step by step guide for conducting a systematic review and meta-analysis with simulation data. Tropical Medicine and Health, 47(1), 46. https://doi.org/10.1186/s41182-019-0165-6

Wynants, L., Calster, B. V., Collins, G. S., Riley, R. D., Heinze, G., Schuit, E., Albu, E., Arshi, B., Bellou, V., Bonten, M. M. J., Dahly, D. L., Damen, J. A., Debray, T. P. A., Jong, V. M. T. de, Vos, M. D., Dhiman, P., Ensor, J., Gao, S., Haller, M. C., … Smeden, M. van. (2020). Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal. BMJ, 369, m1328. https://doi.org/10.1136/bmj.m1328

SWEDCON. (2023). Carelyo. https://carelyo.ng

Appendices

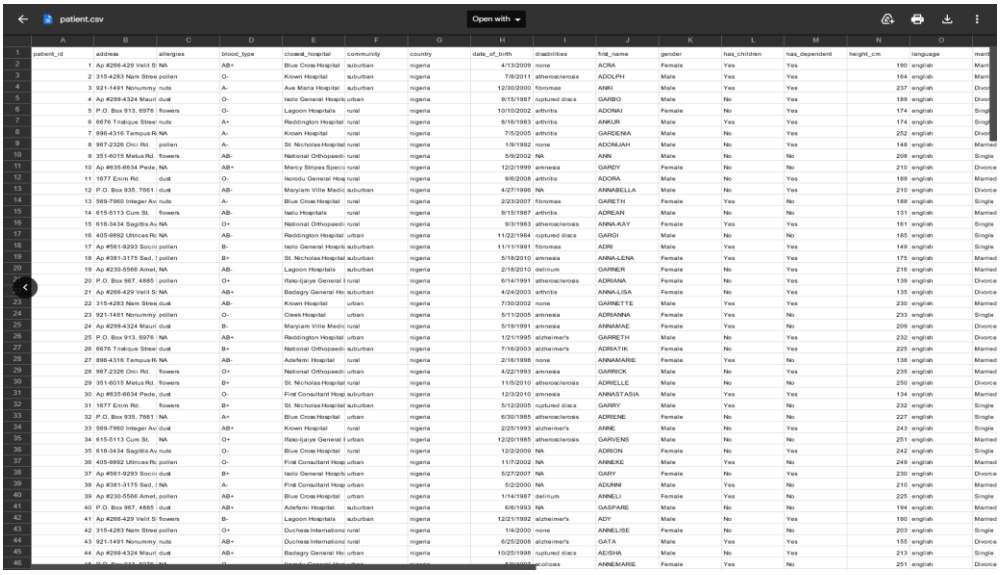
1. Thesis GitHub

   abelshif/Data-Science__Thesis-Project (github.com)

2. **Definition of the explanatory and response variables**

   consult_id: Consultaion id.
   patient_anatomy: Patient body part that examination via consultation.
   patient_consult_status: The status of patient consultation.
   patient_community: The community the patient lives in.
   patient_marital_status: The marital status of the patent
   patient_has_children: Wheather the patient has children or not.
   patient_has_dependent: Wheather the patient has someone to support or not.
   patient_num_of_children: Number of patients children
   patient_num_of_dependents: Number of people whom the patient supports(s).
   patient_age: Age of patient.
   patient_gender: Gender of patient.
   consultant_doctor_experience: Work experience by the consultant medical doctor.
   consult_cost: consultation cost.

3. **The Populated patient dataset**



4. ER model for the populated patient dataset