Dengue Recognition System for Srilanka

By using machine learning

Name - Savindu Mahasen Ruhunuhewa

Cardiff ID - st20212770

ICBT ID - KD/BSCSE-CMU/03/02

**Introduction**

Dengue fever poses a persistent threat to public health in Sri Lanka, exacting a heavy toll on individuals and the healthcare system alike. With its tropical climate fostering the proliferation of the Aedes mosquito—the primary vector for dengue transmission—the country contends with recurrent outbreaks of the disease, precipitating widespread illness, economic strain, and societal disruption.

Conventional methods of diagnosing dengue hinge largely on clinical assessment, a process fraught with challenges due to the disease's variable and overlapping symptoms. This diagnostic complexity frequently results in delays in treatment initiation, heightening the risk of severe complications and mortality. Moreover, the sheer volume of dengue cases overwhelms healthcare facilities, exacerbating resource constraints and impeding effective disease management.

Recognizing the imperative for more efficient and precise diagnostic tools, there has been a burgeoning interest in harnessing machine learning technologies for dengue detection. Machine learning, a facet of artificial intelligence, holds promise in scrutinizing extensive and diverse datasets to unveil hidden patterns and insights. By leveraging this computational prowess, machine learning algorithms can facilitate early detection of dengue, stratification of risk, and allocation of resources, thereby ameliorating clinical outcomes and optimizing resource utilization.

This project endeavors to craft a bespoke machine learning-driven dengue detection system tailored explicitly to the Sri Lankan milieu. Through the integration of disparate datasets encompassing clinical, epidemiological, and environmental variables, our objective is to fashion predictive models adept at accurately discerning dengue cases and forecasting outbreaks. Furthermore, by harnessing real-time data streams and sophisticated analytics, our system aspires to furnish timely insights to healthcare practitioners and policymakers, enabling proactive interventions and targeted public health initiatives.

In collaboration with local healthcare authorities, academic institutions, and technological partners, this endeavor embodies a multidisciplinary endeavor to combat dengue in Sri Lanka.

By amalgamating expertise from the realms of healthcare, data science, and public health, we aspire to devise innovative solutions that not only augment dengue diagnosis and surveillance but also empower communities to mitigate the disease's impact. Ultimately, our vision is to forge a scalable and sustainable framework for dengue management that can be tailored and implemented across diverse contexts, paving the way for a healthier and more resilient Sri Lanka.

**Problem statement**

* Based on symptoms, machine learning existing applications or systems are fully created for health care professionals and doctors. At that moment there is no specific machine learning application or system is available for patient to detect the dengue by their own device.
* Lot of urban areas don’t have laboratory facilities for earlier detecting of dengue. Based on that urban area people need to go long distance for meet the doctor or health care professional to check the dengue or not. So at that time Dengue fever can be highly increase for them.

**Main Objectives**

Objective 1: Review Existing Machine Learning Applications for Dengue Detection Examine and analyze current machine learning systems used for detecting dengue fever. Assess the existing systems and limitations of current methods.

Objective 2: Explore Patient Needs and Challenges in Dengue Diagnosis Investigate the requirements and difficulties patients face regarding dengue diagnosis, particularly in urban settings where laboratory access is limited. Explore the obstacles patients encounter in obtaining prompt and accurate dengue diagnosis and their potential impact on public health.

Objective 3: Develop a User-Friendly Dengue Detection System Create a new machine learning-based dengue detection system tailored to patients' needs and suitable for use on their personal devices. Prioritize ease of use, accessibility, and convenience to ensure the system is suitable for a diverse range of patients.

Objective 4: Optimize Machine Learning Algorithms for Dengue Prediction Evaluate and compare various machine learning algorithms, such as support vector machines, decision trees, and random forests, for their effectiveness in predicting dengue fever. Investigate methods for improving model performance, such as tuning hyperparameters, selecting relevant features, and employing ensemble techniques.

Objective 5: Validate the Performance of the Dengue Detection System Thoroughly test and validate the developed dengue detection system using real patient data and clinical scenarios. Assess the system's accuracy, sensitivity, specificity, and predictive capabilities against established benchmarks and reference standards.

Objective 6: Offer Recommendations for System Enhancement and Future Development Identify areas where the dengue detection system can be improved based on symptoms.

**Literature Review**

**Introduction**

In now a days dengue is a common health issue in the world. Because dengue is spread though the word faster. Especially in srilanka. For an example in srilanka in may 2023 61,361 total dengue patients are reported(Sri Lanka: Dengue Outbreak - May 2023). Earlier days lot of laboratory testing methods were doing for detecting the dengue. For example Reverse transcriptase-polymerase chain reaction (RT-PCR), viral isolation and identification, nuclear acid sequence-based amplification (NASBA), IgM and IgG seroconversion, PCR testing are used. But these tests have several disadvantages. for example cost, complexity, and takes lot of time to get the result etc. Based on these disadvantages different systems, and machine learning techniques already created for dengue detection. But now a days most of these created machine learning systems or applications created for assisting to doctors and healthcare professional to identify the patient dengue. Which means there is no specifically created the any particular machine learning system or application for patient to dengue identify. In this literature review mainly focused on what are the already existing application, systems by using machine learning techniques and their problems. And also consider the other normal application, or systems. And their problems.

**Problems**

**Dengue prediction and diagnosis using machine learning techniques**

According to this “A dengue disease prediction and diagnosis model using sentiment analysis and machine learning algorithms” research paper, for the predicting the dengue KNN classifier, decision tree, random forest, Gaussian naive Bayes, and support vector classifier (SVC)(Gupta et al., DDPM: A dengue disease prediction and diagnosis model using sentiment analysis and machine learning algorithms 2023).

**The main limitations of this application**

Unbalance data set - when it comes to the dengue dataset. Where that contains positive cases of dengue may be rare compare to the negative cases.

Hyper parameter tuning problem - when using this models for increasing the model accuracy need to take lot of time to find out the best parameters.

Scalability - when using this KNN classifier, random, SVC are high expensive when we are dealing with larger amount of dataset (Jamwal & Bhatia, Prevalence of vector borne diseases in Jammu division, Jammu and Kashmir, India 2021).

**Classification of dengue application using machine learning techniques**

According to the “Classification of dengue using machine learning techniques” research paper, Simple Cart, C-4.5, Multi-layer perception algorithms used for classifying the dengue is infected or not (Sajana et al., Classification of dengue using Machine Learning Techniques 2018).

The main limitations of this models

* Overfitting - These simple cart, C-4.5, willing to overfitting, when trees becomes to the more deep, and complex. But this leads to generate the inaccurate predictions.
* Sensitive to small changes - these simple cart, C-4.5 algorithms highly respond to the training dataset. That leading to the different tree structure and potentially different predictive performance.
* Imbalance dataset - when it comes to the dengue dataset. Where that contains positive cases of dengue may be rare compare to the negative cases.
* Privacy and security problems - extracting the human sensitive information raise some privacy problems.
* Model accuracy - According to the “Dengue outbreaks prediction in bangaladesh perspective using distinct multi-layer perceptron NN and decision tree” research paper. Predicting accuracy of Multi-layer perceptron 68.5%. which means it not sufficient for using the predicting purpose (Khan et al., Dengue outbreaks prediction in Bangladesh perspective using distinct multilayer Perceptron NN and decision tree 2022).

**Presumptive diagnosis System for dengue fever by using machine learning**

Presumptive diagnosis system used the Decision Tree, Random Forest, Naive Bayes algorithms used for predicting the dengue fever. What ever this Random Forest , Naive Bayes, Decision Tree algorithms use for early detection of dengue in this system (Khan & Raza, Development and evaluation of a predictive diagnostic system for dengue fever using Machine Learning Techniques 2023).

The main Limitation of this system

* Less number of data points are available
* Accuracy of this models are very low
* Sensitivity - presumptive diagnosis system is lack with sensitivity for early stage of detecting (Khan & Raza, Development and evaluation of a predictive diagnostic system for dengue fever using Machine Learning Techniques 2023).

**Detection of dengue disease by using fused machine learning**

In this case use the PFDM model. In this PFDM model use two main ML-based procedures. For an example SVM and ANN. And this PFDM has two basic components. the training layer and the testing layer. In the training layer,

It has five steps (collecting, preprocessing, classification, efficiency, and machine-level combination). If the model fails to meet learning criteria, it is reassigned until satisfactory results are achieved. The outputs of ANN and SVM are then fused, and the trained model is stored in a cloud system. In the testing layer, data is obtained from a database, and preprocessed trained models are loaded from the cloud. The fused model is used to predict whether a Dengue diagnosis is progressive or destructive, with the predicted outcome compared to the actual result to measure accuracy (Al Nasar et al., Detection of dengue disease empowered with fused machine learning 2022).

The main limitation

* One important factor in classification is class imbalance, and classification efficiency cannot be entirely determined by accuracy rates alone (Al Nasar et al., Detection of dengue disease empowered with fused machine learning 2022).
* Feature engineering - SVM(Support vector machine) is heavily depend on feature engineering to identify the relevant patterns (Hoyos et al., Dengue models based on machine learning techniques: A systematic literature review 2021).
* Complexity and overfitting - when dealing with the noisy and small datasets. Overfitting can generate unreliable predictions (Hoyos et al., Dengue models based on machine learning techniques: A systematic literature review 2021).
* Computational resources - when using ANN it required the computational resources including high performance GPU, and memory (Chakraborty et al., Forecasting dengue epidemics using a hybrid methodology 2019).

**Decision support system for dengue detection**

The use of decision support systems (DSS) is essential for the early diagnosis of dengue. A viable method for creating DSS for dengue care is case-based reasoning (CBR), which uses clinical guidelines and historical cases to forecast a patient's present state based on their vital signs and symptoms (binti Mohd Zainee & Chellappan, A preliminary dengue fever prediction model based on vital signs and blood profile 2020).

Limitation of decision support system

* Imbalance dataset - which means dengue infected positive cases significantly are lower than negative cases.
* Data is limited - dengue dataset size is small .
* Complexity of system - Dengue has several symptoms. That symptoms can be overlapped with similar similar illness (Lopez et al., An intelligent decision support system to prevent and control of Dengue 2018).

**Solution**

According to the above mentioned machine learning systems, applications and normal systems like decision support system there are significant limitations are available. Base on that machine learning system, applications limitations are complexity of model and system, low accuracy of models, overfitting, unbalance dataset etc. And also above mention systems are mainly created for doctor and healthcare professional. Then some urban area people need to go long distance for meet the doctor or laboratories for detecting the dengue is infected or not. Based on that urban area people cannot be identify the dengue in earlier stage. These are the major problems existing in machine learning dengue systems and application. When it comes to the normal systems like decisions support system there is some issue like complexity of the system, data limitation, in balance dataset available.

Based on above mentioned machine learning systems, applications and normal systems like decision support systems problems or limitations, dengue detection system for patient is going to be implemented for patient by using decision tree, random forest, logistic regression algorithm as solution including the “grid\_search cv” , proper dataset and models parameter like random state.

The main reasons of this support vector machine model is used in dengue detection system are easily work with high dimensionality features, and memory efficient which means support vector machines don’t need all the data. It wants only near the decision boundary. And also support vector machines can be easily work with different data types by using the mathematical tricks called kernals(Nordin et al., The classification performance using support vector machine for endemic dengue cases 2020).

The reasons of decision tree algorithm used for dengue detection system are efficiency is very high. Which means by using this decision tree algorithm can be making the understandable decision. And also easy to apply this decision tree algorithm, and work with larger dataset (Jayasundara et al., Developing a decision support testing algorithm to detect severity level of Dengue 2017).

The reasons of random forest algorithm used for dengue detection system are efficiency is very high. Which means by using this random forest algorithm can be making the correct prediction. And also easy to apply this decision tree algorithm, and work with larger dataset (Katta et al., An efficient learning model selection for dengue detection 2022).

**Methodology**

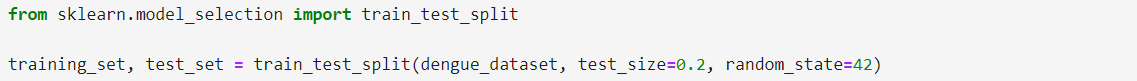
This research belongs to the both quantitative and qualitative aspects

**Data Collection**

Qualitative description - For the dengue detection system based on symptoms by using machine learning technique, data sets are received by websites and hospitals, and other medical institutions. Combine these datasets together and created the full dataset.

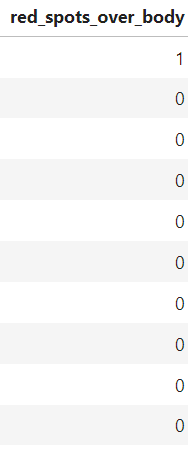
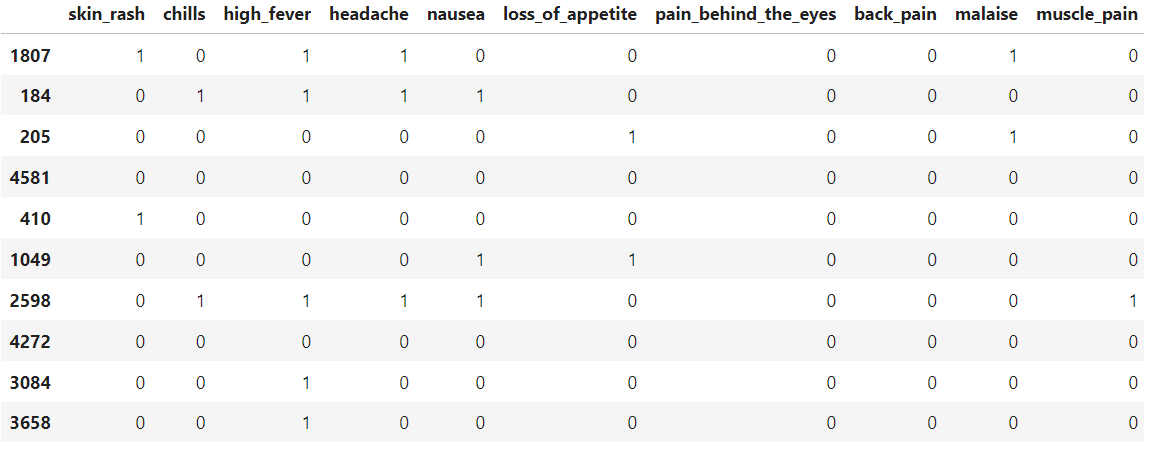
**Splitting the dataset**

Quantitative description - This dengue dataset is going to be divided into the two part by using “random splitting function” Training set and testing set. Training set size is 80%. Test set size is 20%. This training set is used for model training and testing is used for evaluate model performance.



**Data Cleaning**

Quantitative description - Checking the null values in created training and test dataset. According to this dengue prediction for srilanka by using machine learning, dengue dataset doesn’t have any particular null values. And also check the correlational coefficiency which means what are the important attributes for predicting the dengue. Based on that correlational coefficiency remove the unnecessary columns from the training and testing dataset.



**Model selection and explanation of machine learning models used**

For classifying this dengue, binary logistic regression, decision tree algorithm, random forest algorithms are the models going to be used.

Qualitative description  
Binary logistic regression - Binary logistic regression is chosen as a model for dengue prediction due to its simplicity and ease of interpretation. Unlike more complex models, logistic regression offers a clear understanding of the relationship between independent variables and the binary outcome of dengue presence. Its straightforward nature makes it particularly valuable in medical research, where comprehensibility is crucial. Logistic regression is also adept at handling large datasets efficiently, making it suitable for the extensive data often encountered in dengue prediction studies. Furthermore, it accommodates both continuous and categorical predictor variables, allowing for the incorporation of diverse data types commonly found in such studies. By estimating the probability of dengue presence based on various symptoms and risk factors, logistic regression serves as a valuable tool in the array of machine learning techniques for dengue prediction(Pampel, Logistic regression: A Primer 2021).

Qualitative description

Decision tree model - The main reasons of this decision tree model is going to be used in this dengue detection system are this model can be easily used, higher accuracy models can be created, and higher predictions can be produced (Huynh-Cam et al., Using decision trees and random forest algorithms to predict and determine factors contributing to first-year university students’ learning performance 2021).

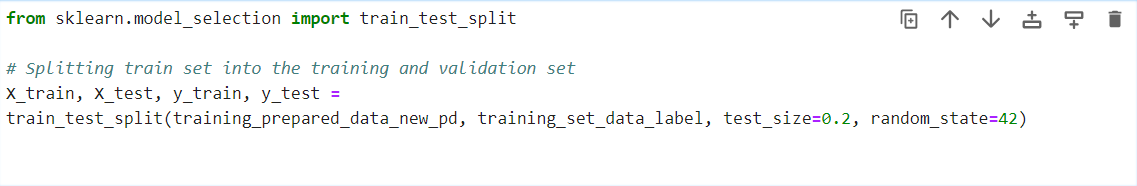
Qualitative description

Random forest model -

The main reasons of this random forest is used in this dengue detection system are increase the model accuracy, reduce the overfitting level, and handling the larger dataset (George & Ganesan, Advanced cuttlefish optimizer-random decision forest (ACORDF) based design of fractional order PID controller for higher-order time-delay system 2021).

**Model training**

After spited the prepared data set into the training and testing dataset, This training set split into the training and validation test again . Below image is the example



**Techniques**

Qualitative description

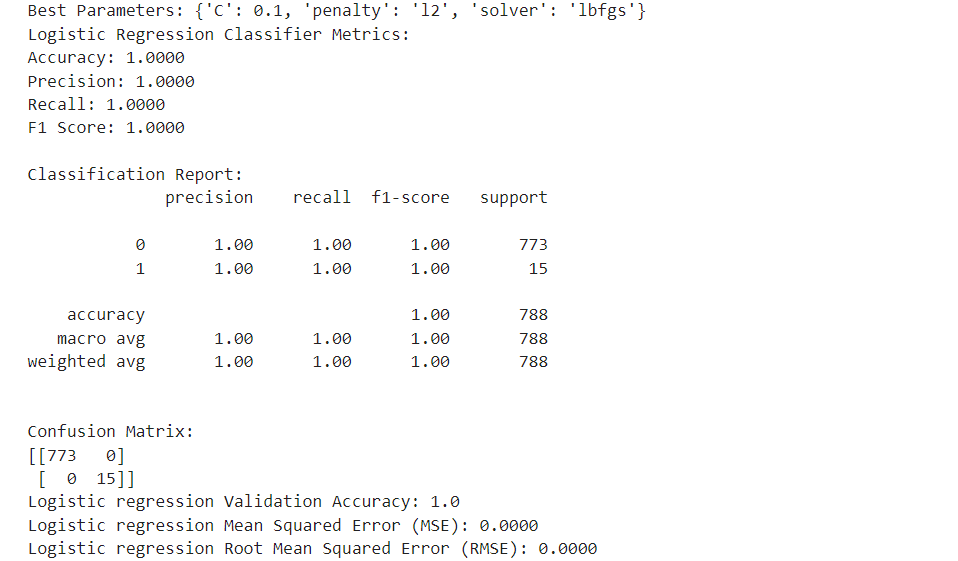
In our methodology, we utilize advanced techniques such as hyperparameter tuning and grid search CV to optimize the performance of our models. These techniques involve identifying the most effective parameters for binary logistic regression, random forest, and decision tree algorithms, thereby fine-tuning the models for dengue prediction. By systematically exploring the parameter space, we aim to enhance model accuracy and mitigate the risk of overfitting.

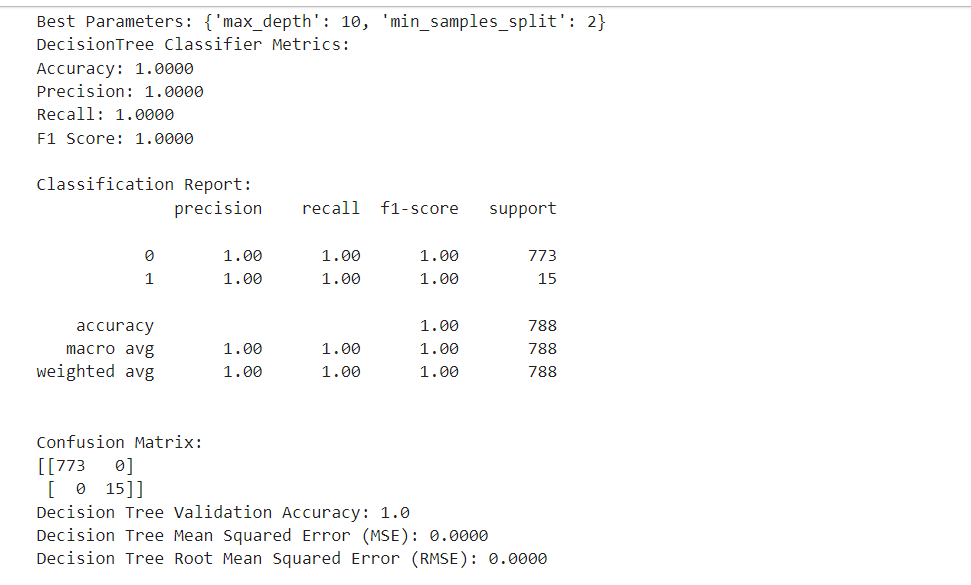
Additionally, we employ ensemble methods to further improve model robustness and generalization. Ensemble techniques involve combining predictions from multiple models, such as random forest, decision tree, and binary logistic regression that choose for dengue prediction. By leveraging the strengths of each individual model and mitigating their weaknesses, ensemble methods enable us to achieve superior predictive performance while reducing the likelihood of overfitting.

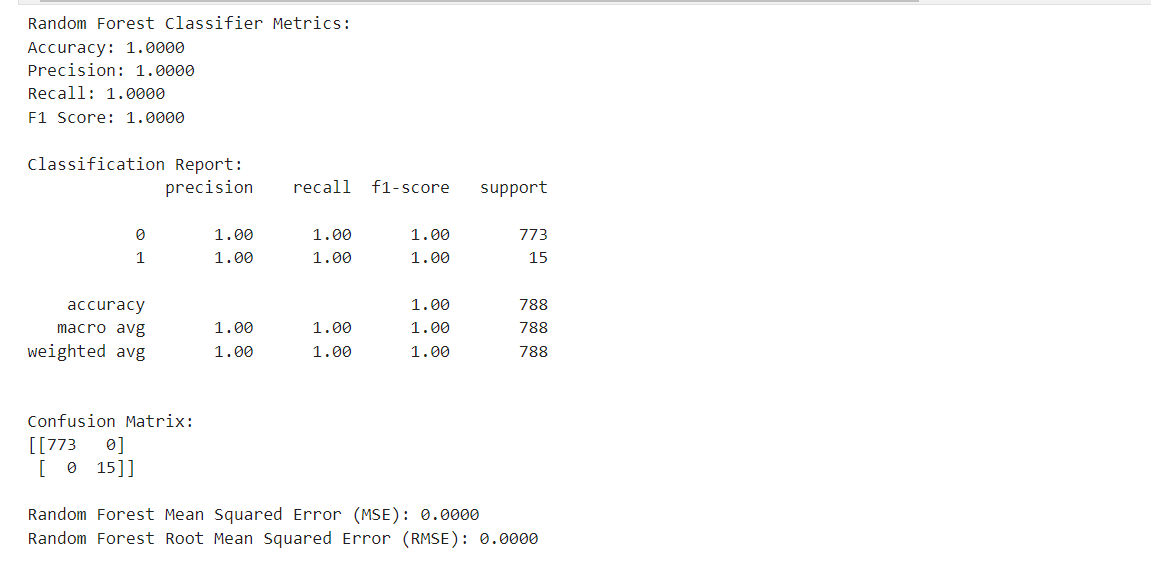
Overall, our approach integrates these advanced techniques within our methodology to optimize model performance, enhance accuracy, and ensure robust dengue prediction capabilities.

**Model results and model evaluation**

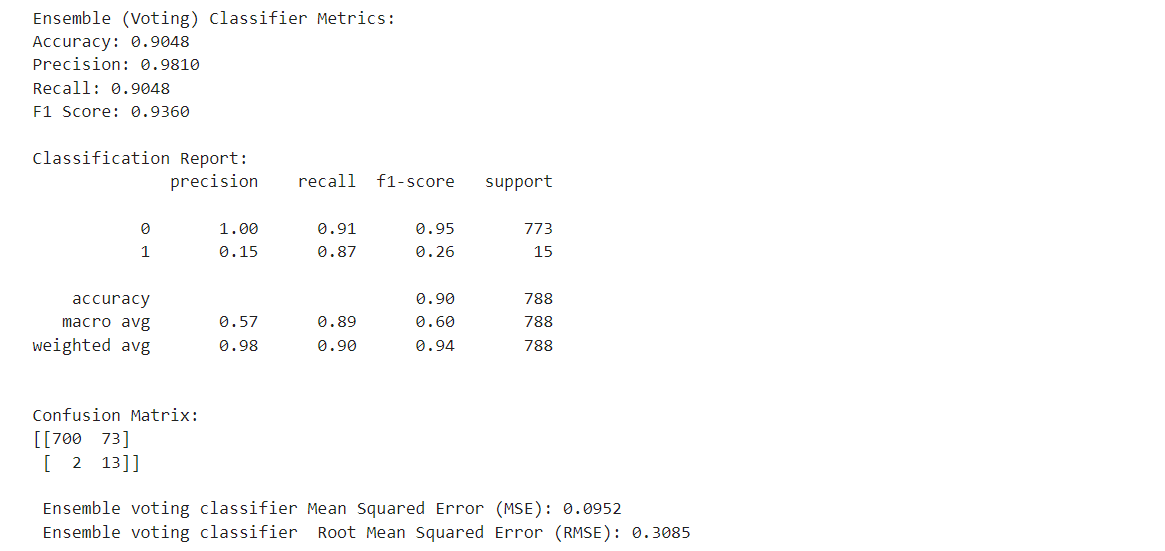
According to this dengue prediction project after doing the grid search cv for hyper parameter tuning, models metrices (accuracy, F1 score, precision, recall, F1 score) and model evaluations are shown below images.

****

****

****

After doing the grid search cv and hyperparameter tuning model overfitting issue doesn’t fix. Because above random forest, decision tree, binary logistic regression accuracies are still 1.0(100%). So then to avoid that overfitting issue use the ensemble technique by combining the random forest, decision tree, binary logistic regression model. Below images is the example for using the ensemble technique and accuracy, precision, recall, F1 score after using the ensemble technique.

****

**Limitations of the methods have been identified in the methodology**

Limited Data Availability and Quality: One of the primary challenges in developing predictive models for dengue detection is the availability and quality of data. While data are collected from websites, hospitals, and medical institutions, there may be limitations in terms of completeness, accuracy, and representativenes. To overcome this limitation, efforts should be made to enhance data collection mechanisms, collaborate with relevant stakeholders to improve data sharing, and employ data augmentation techniques to expand the dataset size.

Imbalanced Dataset: Imbalance in the dataset, where the number of dengue-positive cases may be significantly lower than the dengue-negative cases, can impact the performance of the predictive models. This imbalance can lead to biased predictions and reduced sensitivity in detecting dengue cases. To address this issue, techniques such as oversampling of the minority class (dengue-positive cases) or using algorithms that are robust to class imbalance should be considered.

Model Interpretability: While complex machine learning models like random forest and may offer high predictive accuracy, they often lack interpretability, making it challenging to understand the underlying factors contributing to dengue prediction. To enhance interpretability, simpler models like logistic regression and decision trees are utilized in conjunction with more complex models. Additionally, techniques such as feature importance analysis and model visualization can aid in understanding the decision-making process of the models.

Generalization to Other Regions: The predictive models developed in this research may be tailored to the specific context of Sri Lanka, raising concerns about their generalizability to other regions with different environmental, demographic, and healthcare characteristics. To improve generalizability, future research should involve validation of the models across diverse geographical regions and populations, incorporating region-specific features and adjusting model parameters accordingly.

**Analysis and Discussion:**

In the analysis and discussion of dengue detection for Sri Lanka using machine learning models such as binary logistic regression, decision tree, and random forest, the interpretation of results provides valuable insights into the effectiveness and limitations of these models.

The models demonstrated promising performance in accurately predicting dengue presence based on symptom data. The binary logistic regression model, known for its simplicity and interpretability, offered a clear understanding of the relationship between symptoms and dengue presence. Decision tree models showed the potential for easy interpretation and high prediction accuracy, while random forest models enhanced accuracy by reducing overfitting and handling larger datasets effectively.

However, despite their effectiveness, the models also exhibited certain limitations. While binary logistic regression provides simplicity and interpretability, it may struggle with capturing complex relationships in the data. Decision tree models, although easy to understand, are prone to overfitting, especially with deeper trees. Random forest models, while robust, can be computationally expensive and challenging to interpret due to their ensemble nature.

The implications of these findings for dengue detection and public health are significant. Accurate prediction of dengue presence using machine learning models can facilitate early detection and intervention, helping public health authorities allocate resources more effectively and mitigate the spread of the disease. By leveraging these models, healthcare practitioners can prioritize high-risk areas for targeted interventions, such as mosquito control measures or public awareness campaigns.

Addressing the research questions or hypotheses posed in this study provides further insights into the applicability of machine learning models for dengue detection in Sri Lanka. Future research directions may include exploring the use of more advanced machine learning techniques, such as deep learning or ensemble methods, to improve prediction accuracy further. Additionally, integrating additional data sources, such as environmental or demographic factors, could enhance the models' predictive capabilities and support more comprehensive dengue surveillance efforts.

**Solution design and implementation**

In the design and implementation of machine learning models for dengue detection in Sri Lanka, careful attention must be given to ensure their seamless integration into existing surveillance systems and their potential impact on public health practices. Choosing appropriate models, such as binary logistic regression, decision trees, and random forests, requires considering their suitability for the specific needs and constraints of dengue surveillance. Each model presents its own advantages and challenges, necessitating thorough evaluation based on factors like interpretability, scalability, and computational requirements.

To integrate these models effectively into dengue surveillance systems, it is crucial to establish interoperability with current data collection, processing, and reporting mechanisms. This may entail developing standardized data interfaces, utilizing cloud-based solutions for scalability, and implementing real-time data integration pipelines to facilitate timely and accurate dengue risk assessments. Additionally, robust validation and evaluation frameworks are essential to continually monitor model performance and adapt to changing dengue dynamics and epidemiological trends.

Several challenges, including data quality issues, model interpretability concerns, and resource limitations, must be addressed for the real-world deployment of machine learning models in dengue detection. Strategies such as data preprocessing, feature engineering, and model explainability techniques can help overcome these challenges and enhance the reliability and trustworthiness of deployed models. Moreover, fostering collaboration among public health authorities, researchers, and local communities through stakeholder engagement and capacity building initiatives is crucial to maximize the impact of machine learning-based dengue surveillance efforts.

The potential impact of these models on public health practices in Sri Lanka is substantial, offering the opportunity to complement traditional surveillance methods with data-driven insights for more targeted and effective dengue control strategies. Leveraging the predictive capabilities of machine learning models enables public health authorities to identify high-risk areas, allocate resources efficiently, and implement proactive interventions to curb the spread of dengue fever, thus alleviating the burden on healthcare systems. Ultimately, successful integration and deployment of these models have the potential to revolutionize dengue surveillance and control efforts in Sri Lanka, leading to improved public health outcomes and greater resilience to dengue outbreaks.

**Research questions and Hypotheses**

**Normality testing**

Normality testing means check whether data is normally distributed or not. Below images are examples for normality testing.

Significant level (α) = 0.05 (5%)

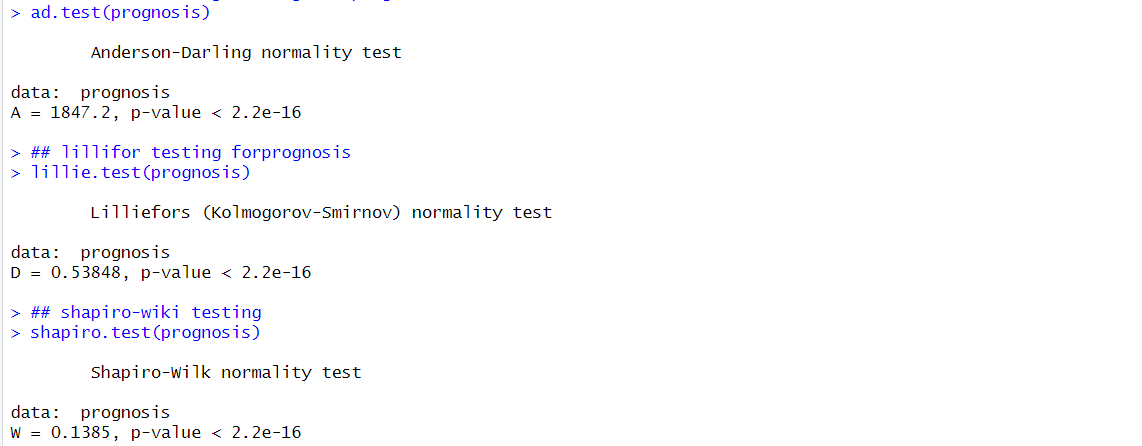
Confidence level =0.95(95%)

Variable = prognosis

(\*) Whether or not prognosis is normally distributed?

H0: prognosis values are normally distributed

H1: prognosis values are not normally distributed



According to the above three test p value is 2.2e-16. which means p-value is less than the alpha value (0.05). conclusion of this test is reject the null hypotheses and accept the alternative hypothesis. Which means prognosis is not normally distributed.

Significant level (α) = 0.05 (5%)

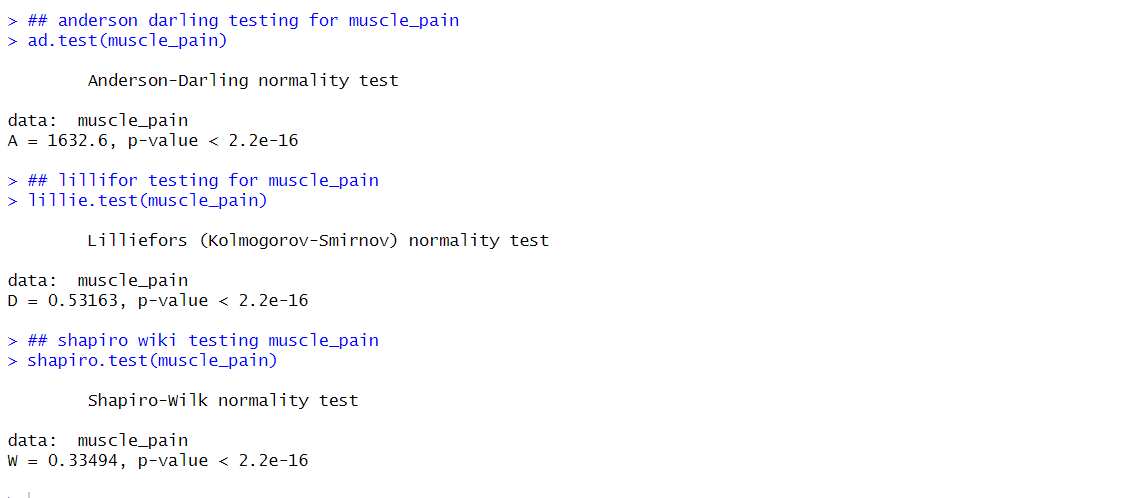
Confidence level =0.95(95%)

Variable = muscle\_pain

(\*) Whether or not muscle\_pain is normally distributed?

H0: muscle\_pain values are normally distributed

H1: muscle\_pain values are not normally distributed



According to the above three test p value is 2.2e-16. which means p-value is less than the alpha value (0.05). conclusion of this test is reject the null hypotheses and accept the alternative hypothesis. Which means muscle\_pain is not normally distributed.

Significant level (α) = 0.05 (5%)

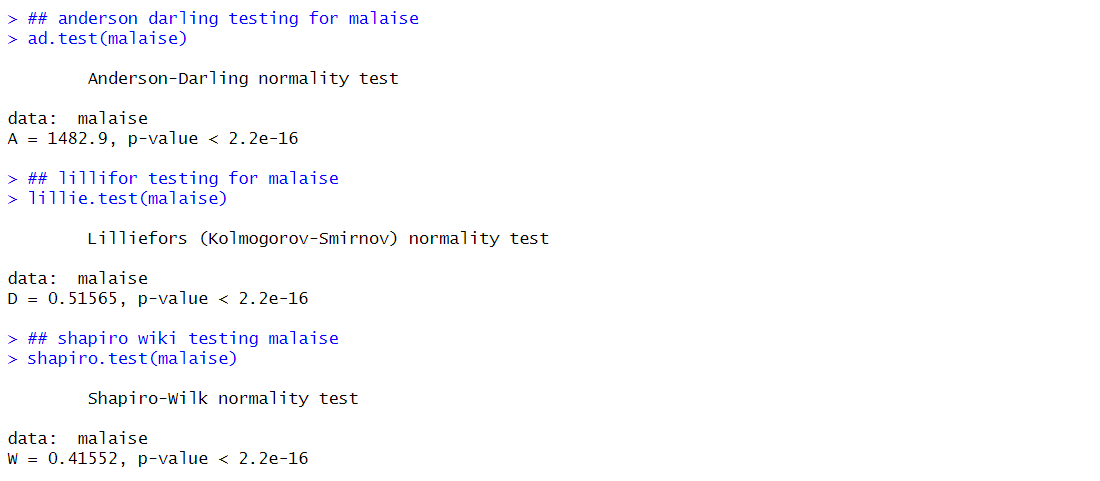
Confidence level = 0.95(95%)

Variable = malaise

(\*) Whether or not malaise is normally distributed?

H0: malaise values are normally distributed

H1: malaise values are not normally distributed



According to the above three test p value is 2.2e-16. which means p-value is less than the alpha value (0.05). conclusion of this test is reject the null hypotheses and accept the alternative hypothesis. Which means malaise is not normally distributed.

Significant level (α) = 0.05 (5%)

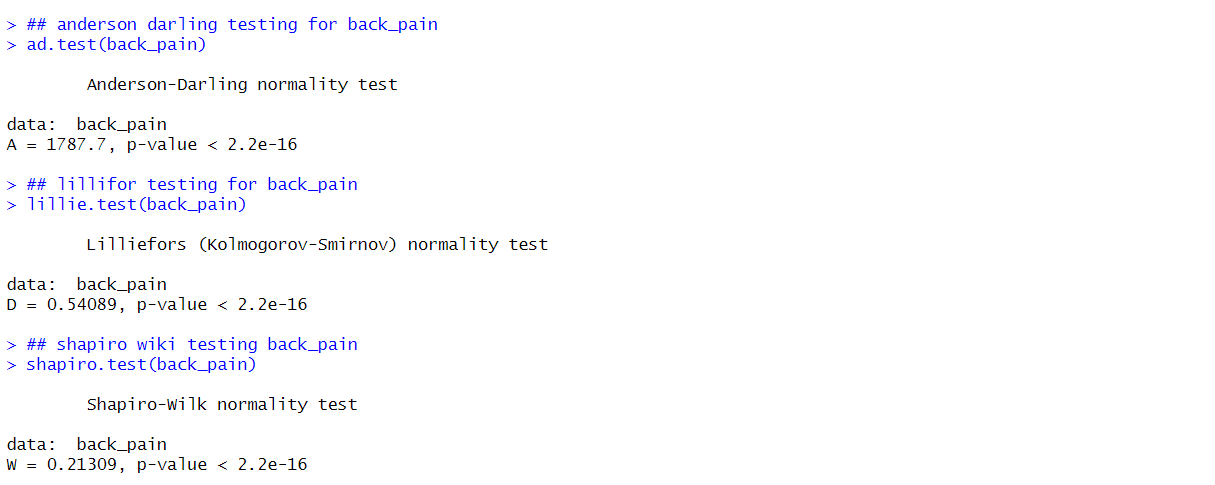
Confidence level =0.95(95%)

Variable = back\_pain

(\*) Whether or not back\_pain is normally distributed?

H0: back\_pain values are normally distributed

H1: back\_pain values are not normally distributed



According to the above three test p value is 2.2e-16. which means p-value is less than the alpha value (0.05). conclusion of this test is reject the null hypotheses and accept the alternative hypothesis. Which means malaise is not normally distributed.

Significant level (α) = 0.05 (5%)

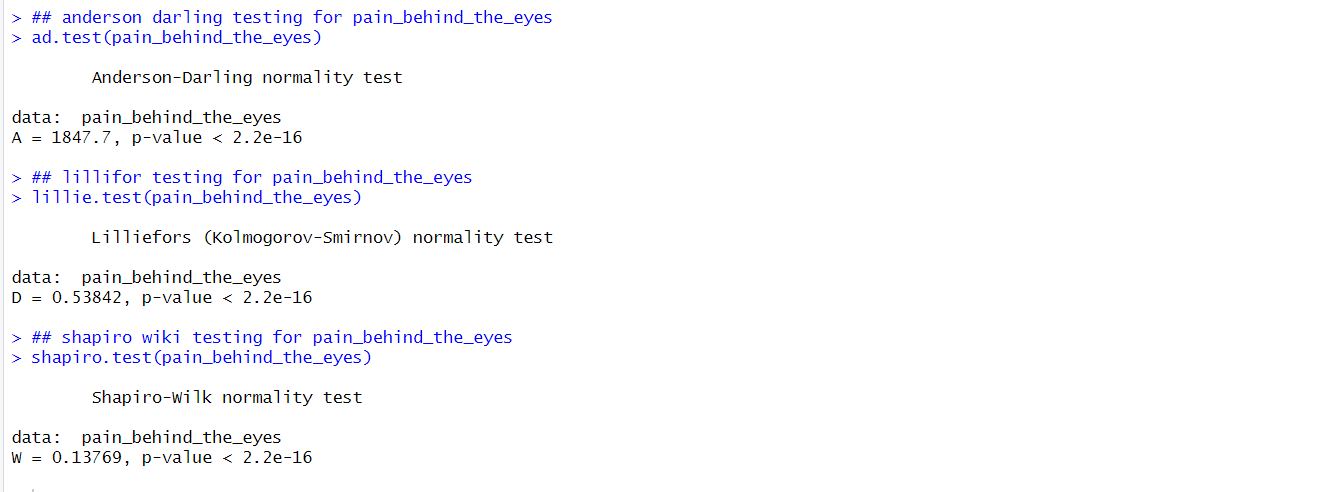
Confidence level =0.95(95%)

Variable = pain\_behind\_the\_eyes

(\*) Whether or not pain\_behind\_the\_eyes is normally distributed?

H0: pain\_behind\_the\_eyes values are normally distributed

H1: pain\_behind\_the\_eyes values are not normally distributed



According to the above three test p value is 2.2e-16. which means p-value is less than the alpha value (0.05). conclusion of this test is reject the null hypotheses and accept the alternative hypothesis. Which means pain\_behind\_the\_eyes is not normally distributed.

Significant level (α) = 0.05 (5%)

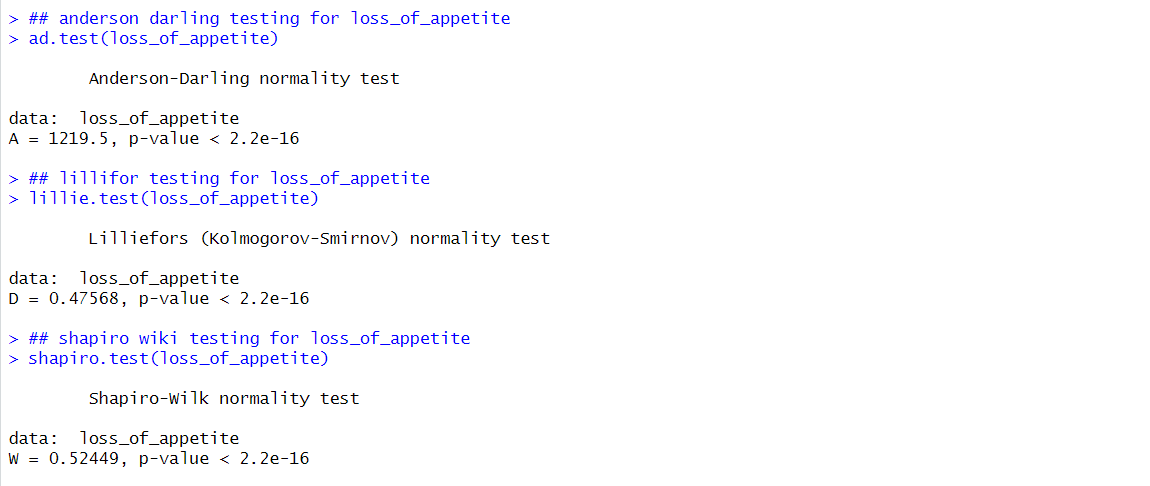
Confidence level =0.95(95%)

Variable = loss\_of\_appetite

(\*) Whether or not loss\_of\_appetite is normally distributed?

H0:loss\_of\_appetite values are normally distributed

H1:loss\_of\_appetite values are not normally distributed



According to the above three test p value is 2.2e-16. which means p-value is less than the alpha value (0.05). conclusion of this test is reject the null hypotheses and accept the alternative hypothesis. Which means loss\_of\_appetite is not normally distributed.

Significant level (α) = 0.05 (5%)

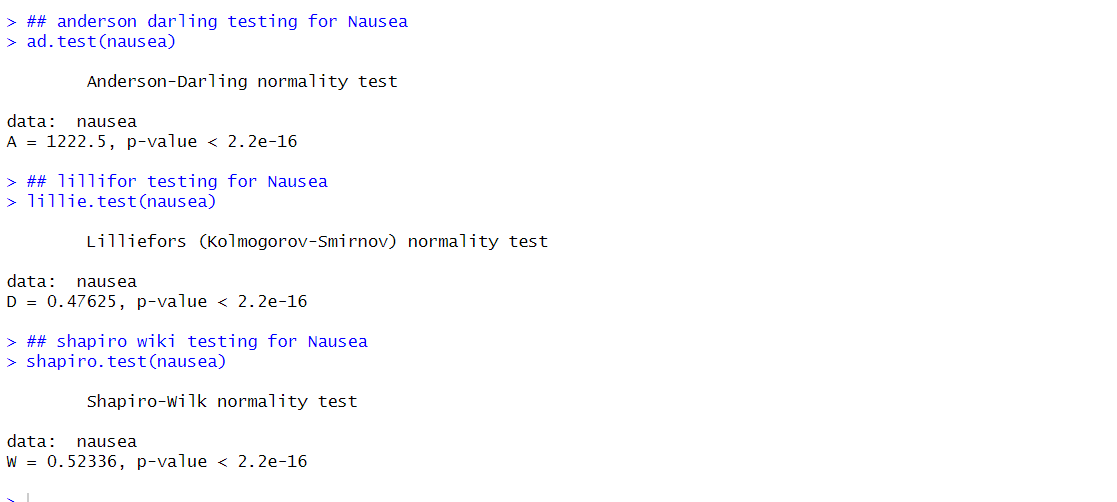
Confidence level =0.95(95%)

Variable = nausea

(\*) Whether or not nausea is normally distributed?

H0:nausea values are normally distributed

H1:nausea values are not normally distributed



The statistical analysis reveals a p value of 2.2e-16, indicating a significant deviation from the normal distribution for the variable "nausea." With this p-value being well below the conventional significance level of 0.05, According to this normality tests of nausea, reject the null hypothesis and accept the alternative hypothesis. Consequently, it is evident that "nausea" does not adhere to a normal distribution pattern.

Significant level (α) = 0.05 (5%)

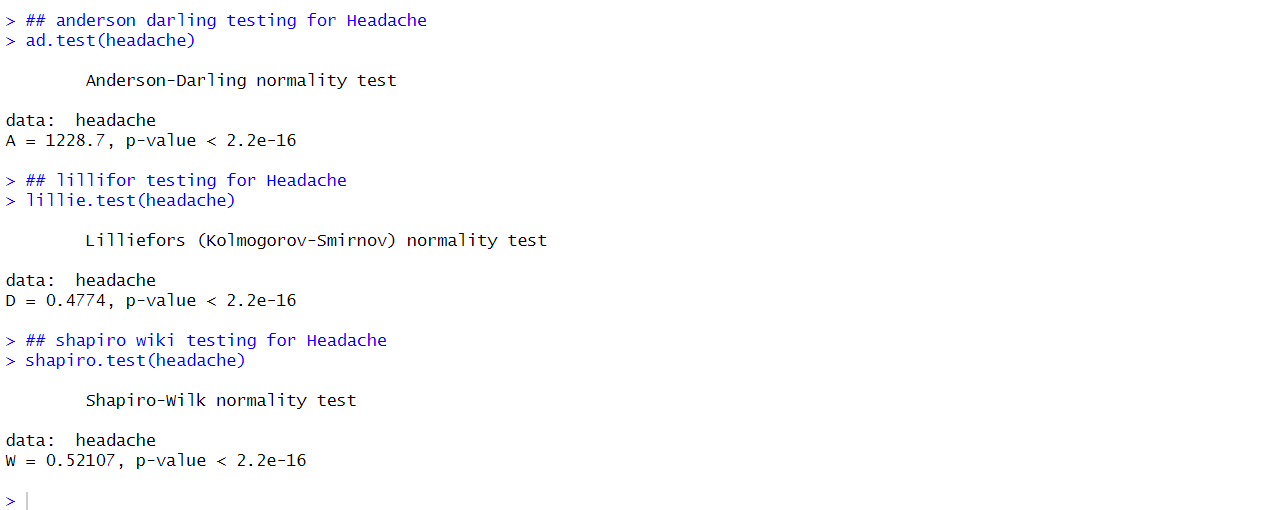
Confidence level =0.95(95%)

Variable = headache

(\*) Whether or not headache is normally distributed?

H0: headache values are normally distributed

H1: headache values are not normally distributed



According to the above three test p value is 2.2e-16. which means p-value is less than the alpha value (0.05). conclusion of this test is reject the null hypotheses and accept the alternative hypothesis. Which means headache is not normally distributed.

Significant level (α) = 0.05 (5%)

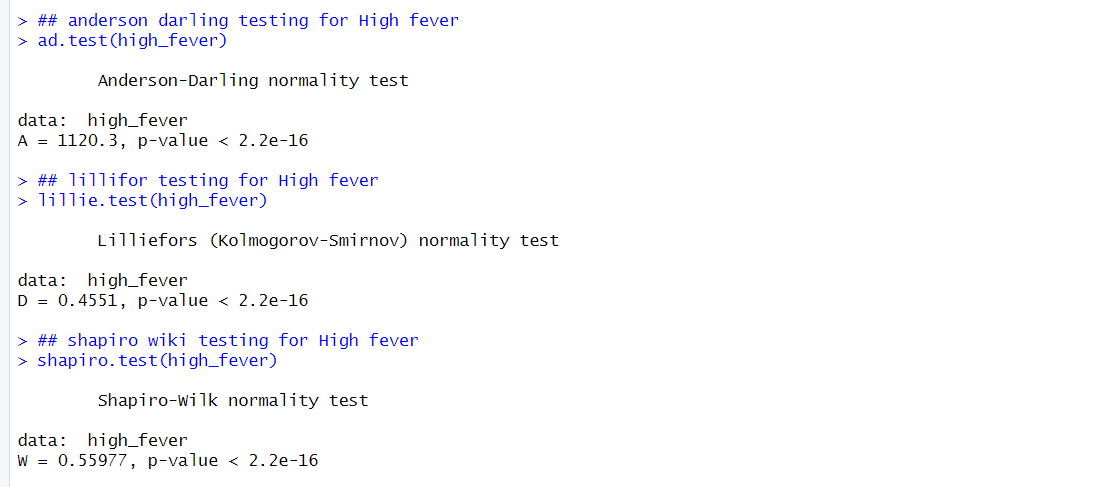
Confidence level =0.95(95%)

Variable = high\_fever

(\*) Whether or not high\_fever is normally distributed?

H0: high\_fever values are normally distributed

H1: high\_fever values are not normally distributed



Based on the statistical tests conducted, the obtained p value of 2.2e-16 is significantly lower than the predetermined alpha value of 0.05. This indicates that we reject the null hypothesis and accept the alternative hypothesis. Therefore, it can be concluded that the variable "high\_fever" is not normally distributed.

Significant level (α) = 0.05 (5%)

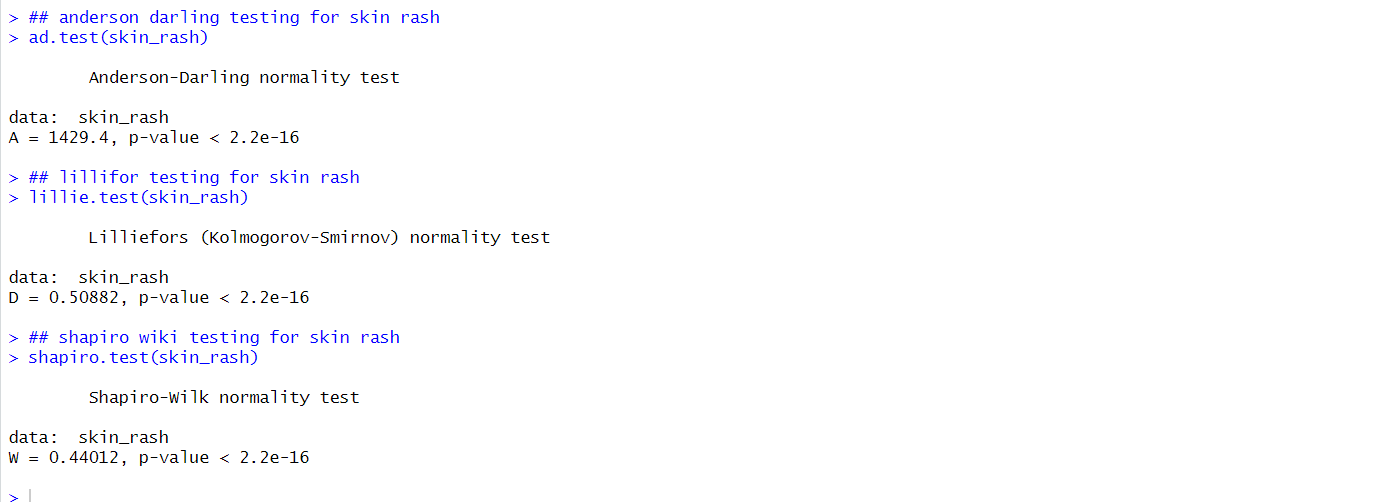
Confidence level =0.95(95%)

Variable = skin\_rash

(\*) Whether or not skin\_rash is normally distributed?

H0: skin\_rash values are normally distributed

H1: skin\_rash values are not normally distributed



Based on the results of the three tests conducted, with a p value of 2.2e-16, which is significantly lower than the predetermined alpha value of 0.05, According to above tests reject the null hypothesis and accept the alternative hypothesis. Consequently, it can be inferred that the variable "skin\_rash" is not normally distributed.

Significant level (α) = 0.05 (5%)

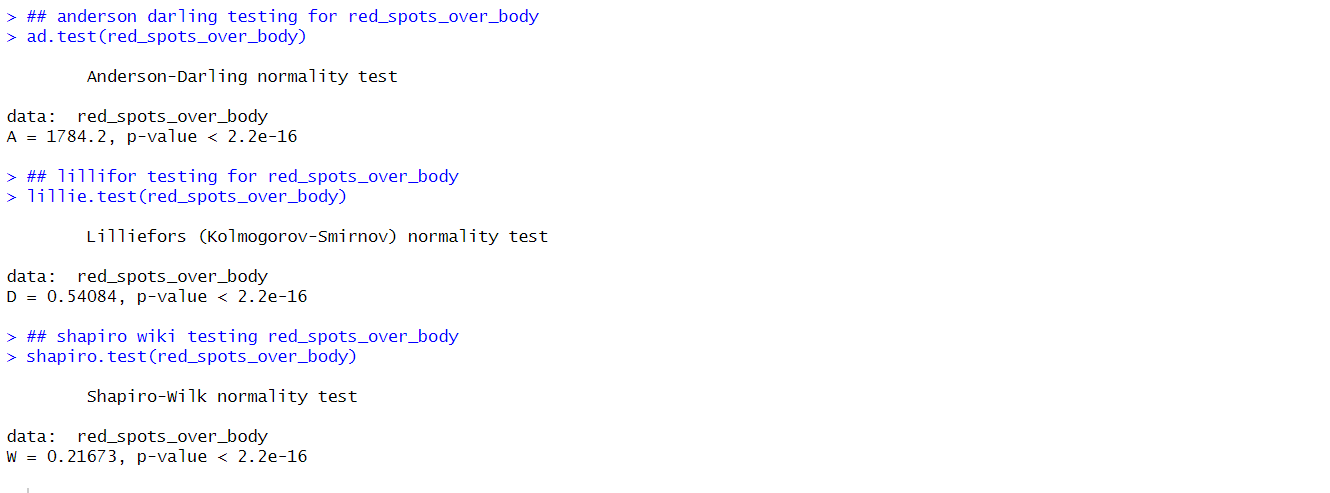
Confidence level =0.95(95%)

Variable = red\_spots\_over\_body

(\*) Whether or not red\_spots\_over\_body is normally distributed?

H0: red\_spots\_over\_body values are normally distributed

H1: red\_spots\_over\_body values are not normally distributed



Based on the statistical analysis conducted on the dataset, the obtained p value of 2.2e-16 is significantly lower than the predetermined alpha value of 0.05. This result leads to the rejection of the null hypothesis and the acceptance of the alternative hypothesis. In practical terms, this suggests that the variable "red\_spots\_over\_body" does not follow a normal distribution.

**Correlational analysis**

In according to the above dengue dataset, based on the normality testing, data are not normally distributed. So then, to do the correlational analysis, spearman method suitable. Correlational analysis is means check the correlational coefficiency between the dependent(y) and independent(x) variables. Which means identify the what are the important attributes that directly affected to dengue prediction based on this project. Below images are example for correlational analysis of dengue detection.

Sample field: muscle\_pain, prognosis

Significant level (α) = 0.05

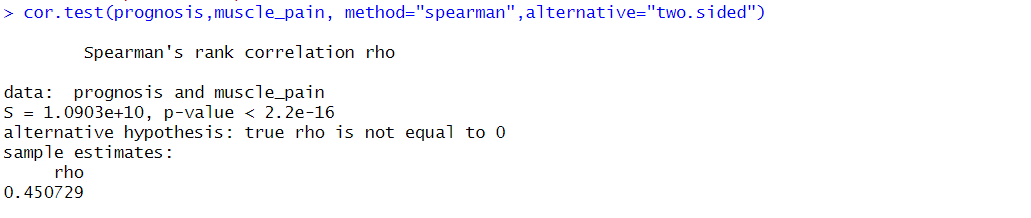
Confidence level = 95%

Let ῤ = the true population correlation coefficient between Temperature and Crop\_Yield

(\*) Whether or not there is a correlation found between muscle\_pain and prognosis?

H0: There is no correlation (ῤ =0) between muscle\_pain and prognosis

H1: There is a correlation (ῤ ≠0) between muscle\_pain and prognosis



At a significance level of 5%, there is statistically significant support for a moderate positive correlation (ῤ = 0.45) between muscle pain and prognosis, as determined by Spearman rank correlation-tests (< 2.2e-16) analyzing sample correlations. This suggests the necessity of conducting regression analysis to establish the regression model for further investigation.

Sample field: malaise, prognosis

Significant level (α) = 0.05

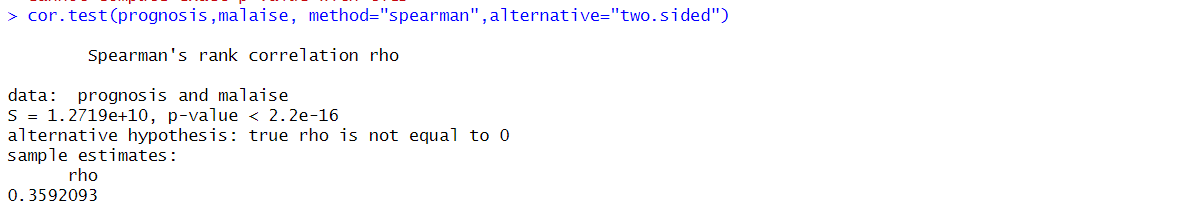
Confidence level = 95%

Let ῤ = the true population correlation coefficient between malaise and prognosis

(\*) Whether or not there is a correlation found between malaise and prognosis?

H0: There is no correlation (ῤ =0) between malaise and prognosis

H1: There is a correlation (ῤ ≠0) between malaise and prognosis



At a 5% significance level, statistically significant support is found for a moderate positive correlation (ῤ = 0.36) between malaise and prognosis, as confirmed by Spearman rank correlation-tests (< 2.2e-16) analyzing sample correlations. Therefore, there's a need for regression analysis to establish the regression model for further investigation.

Sample field: back\_pain, prognosis

Significant level (α) = 0.05

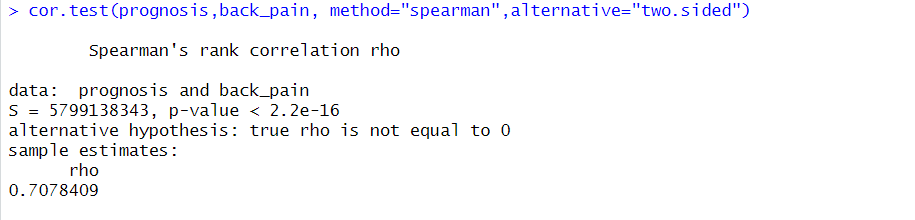
Confidence level = 95%

Let ῤ = the true population correlation coefficient between back\_pain and prognosis

(\*) Whether or not there is a correlation found between back\_pain and prognosis?

H0: There is no correlation (ῤ =0) between back\_pain and prognosis

H1: There is a correlation (ῤ ≠0) between back\_pain and prognosis



A strong positive correlation (ῤ = 0.71) is found between back pain and prognosis, with statistical significance at a 5% level, as confirmed by Spearman rank correlation-tests (< 2.2e-16) analyzing sample correlations. This suggests the need for regression analysis to establish the regression model for further exploration.

Sample field: pain\_behind\_eyes, prognosis

Significant level (α) = 0.05

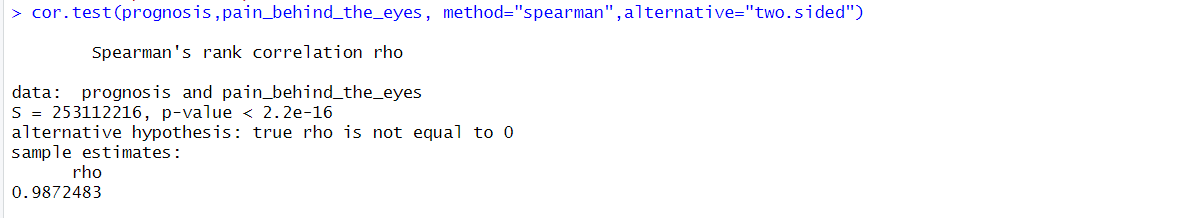
Confidence level = 95%

Let ῤ = the true population correlation coefficient between pain\_behind\_eyes and prognosis

(\*) Whether or not there is a correlation found between pain\_behind\_eyes and prognosis?

H0: There is no correlation (ῤ =0) between pain\_behind\_eyes and prognosis

H1: There is a correlation (ῤ ≠0) between back\_behind\_eyes and prognosis



A strong positive correlation (ῤ = 0.99) exists between pain behind the eyes and prognosis, with statistical significance at a 5% level, as determined by Spearman rank correlation-tests (< 2.2e-16) analyzing sample correlations. These findings suggest the need for regression analysis to establish the regression model for further investigation.

Sample field: loss\_of\_appetite, prognosis

Significant level (α) = 0.05

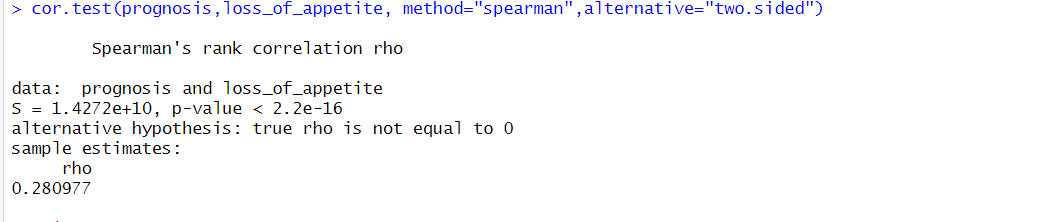
Confidence level = 95%

Let ῤ = the true population correlation coefficient between loss\_of\_appetite and prognosis

(\*) Whether or not there is a correlation found between loss\_of\_appetite and prognosis?

H0: There is no correlation (ῤ =0) between loss\_of\_appetite and prognosis

H1: There is a correlation (ῤ ≠0) between loss\_of\_appetite and prognosis



There is a statistically significant moderate positive correlation (ῤ = 0.28) between loss of appetite and prognosis, supported at a 5% significance level by Spearman rank correlation-tests (< 2.2e-16) analyzing sample correlations. Thus, further investigation via regression analysis is needed to establish the regression model.

Sample field: nausea, prognosis

Significant level (α) = 0.05

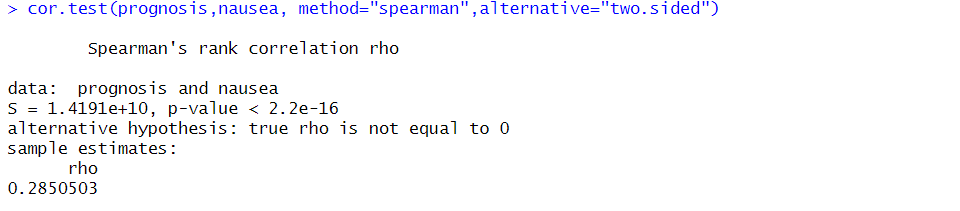
Confidence level = 95%

Let ῤ = the true population correlation coefficient between nausea and prognosis

(\*) Whether or not there is a correlation found between nausea and prognosis?

H0: There is no correlation (ῤ =0) between nausea and prognosis

H1: There is a correlation (ῤ ≠0) between nausea and prognosis



There is statistically significant support, at a 5% significance level, for a moderate positive correlation (ῤ = 0.29) between nausea and prognosis, as evidenced by Spearman rank correlation-tests (< 2.2e-16) analyzing sample correlations. Consequently, further investigation through regression analysis is required to determine the regression model.

Sample field: headache, prognosis

Significant level (α) = 0.05

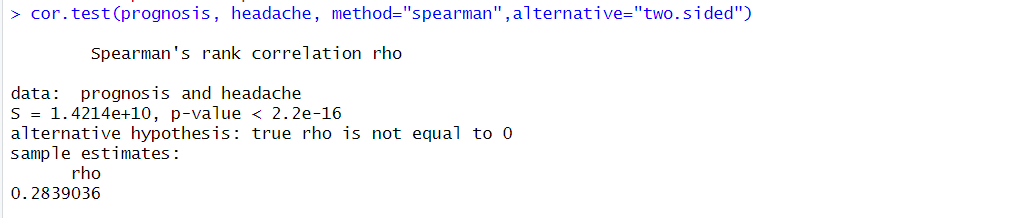
Confidence level = 95%

Let ῤ = the true population correlation coefficient between headache and prognosis

(\*) Whether or not there is a correlation found between headache and prognosis?

H0: There is no correlation (ῤ =0) between headache and prognosis

H1: There is a correlation (ῤ ≠0) between headache and prognosis



Statistically significant support at a 5% significance level is found for a moderate positive correlation (ῤ = 0.28) between experiencing headaches and prognosis, as determined by Spearman rank correlation-tests (< 2.2e-16) analyzing sample correlations. Consequently, there's a need for regression analysis to establish the regression model for further exploration.

Sample field: high\_fever, prognosis

Significant level (α) = 0.05

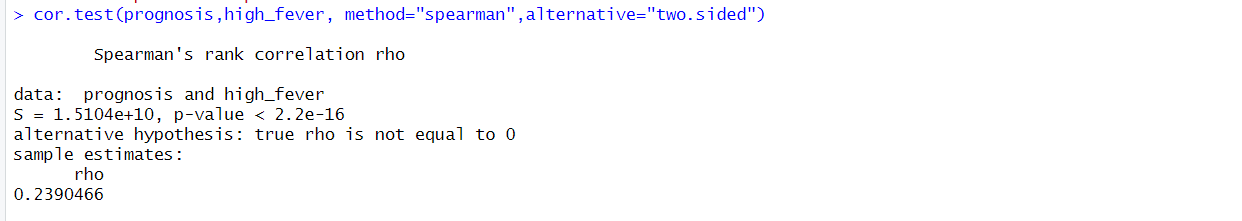
Confidence level = 95%

Let ῤ = the true population correlation coefficient between high\_fever and prognosis

(\*) Whether or not there is a correlation found between high\_fever and prognosis?

H0: There is no correlation (ῤ =0) between high\_fever and prognosis

H1: There is a correlation (ῤ ≠0) between high\_fever and prognosis



At a significance level of 5%, statistically significant support is found for a weak positive correlation (ῤ = 0.24) between high fever and prognosis, as indicated by Spearman rank correlation-tests (< 2.2e-16) analyzing sample correlations. Consequently, there is a requirement for regression analysis to establish the regression model for further investigation.

Sample field: skin\_rash, prognosis

Significant level (α) = 0.05

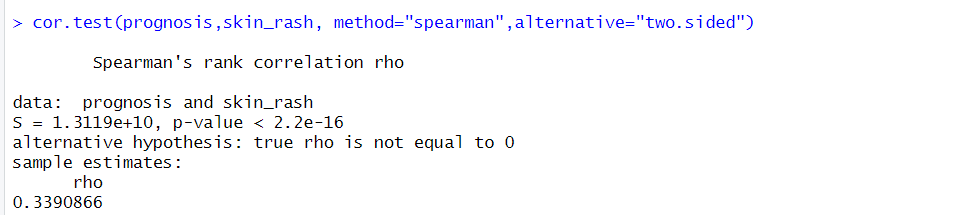
Confidence level = 95%

Let ῤ = the true population correlation coefficient between skin\_rash and prognosis

(\*) Whether or not there is a correlation found between skin\_rash and prognosis?

H0: There is no correlation (ῤ =0) between skin\_rash and prognosis

H1: There is a correlation (ῤ ≠0) between skin\_rash and prognosis



There is a moderate positive correlation (ῤ = 0.34) observed between the presence of a skin rash and prognosis, which is statistically significant at the 5% level according to Spearman rank correlation tests (< 2.2e-16). These findings suggest the need for regression analysis to develop a regression model for further examination.

Sample field: red\_spots\_over\_body, prognosis

Significant level (α) = 0.05

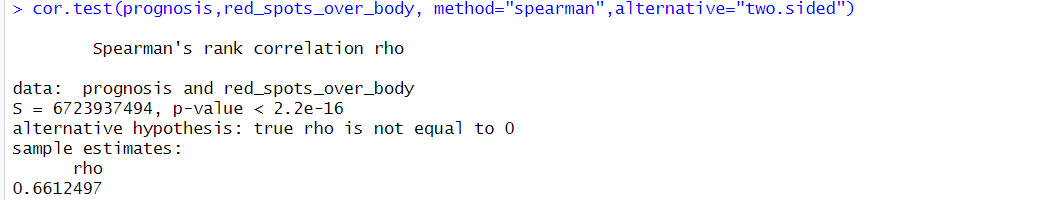
Confidence level = 95%

Let ῤ = the true population correlation coefficient between red\_spots\_over\_body and prognosis

(\*) Whether or not there is a correlation found between red\_spots\_over\_body and prognosis?

H0: There is no correlation (ῤ =0) between red\_spots\_over\_body and prognosis

H1: There is a correlation (ῤ ≠0) between red\_spots\_over\_body and prognosis



A moderate positive relationship (ρ = 0.66) exists between the presence of red spots on the body and the prognosis, with statistical significance at the 5% level, determined by Spearman rank correlation tests (p < 2.2e-16). These results indicate the need for additional exploration through regression analysis to develop a predictive model for prognosis based on the occurrence of red spots on the body.

**Regression analysis**

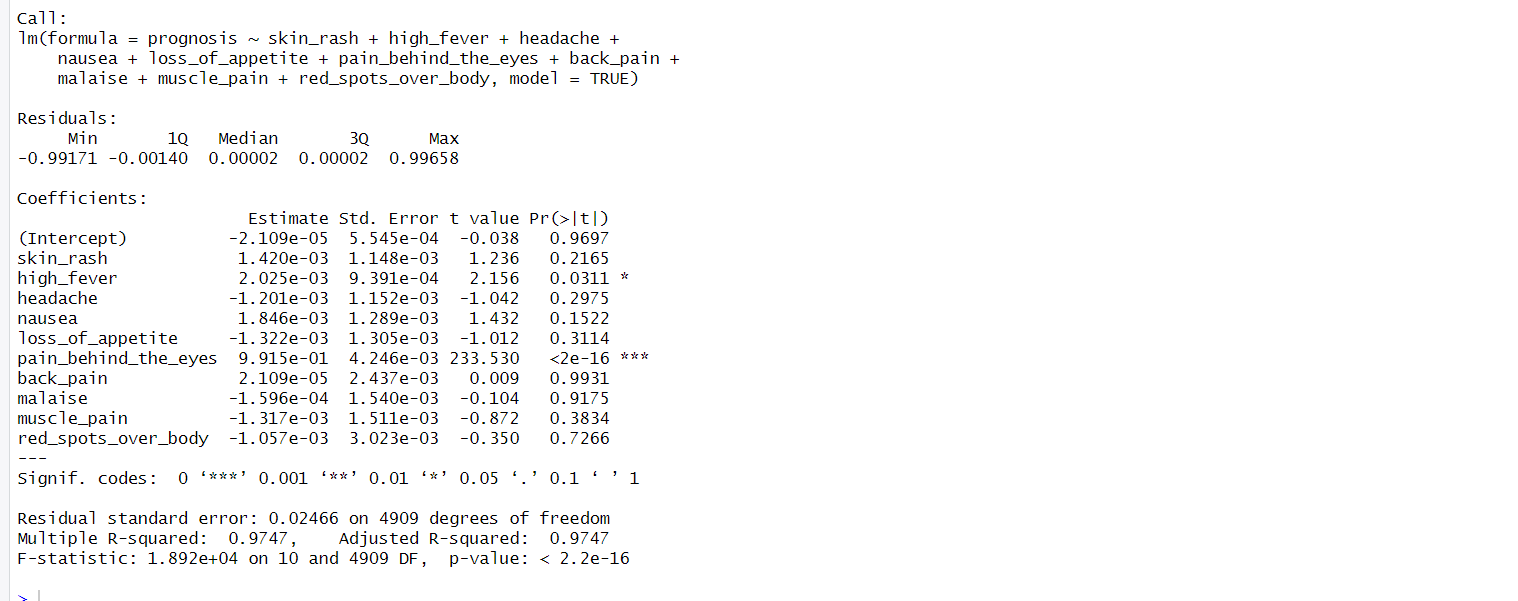
When it comes to the regression analysis there two type are there

1. Simple linear regression analysis
2. Multi linear regression analysis

When considering the simple linear and Multi linear model, Multi linear model give the best results. Because but when it comes real world scenario every problem are are associated with multiple factors. For an example this dengue prediction can be don based on multiple factors(symptoms) not only one symptom. Then fore regression analysis Multi linear model is suitable.

**Multi Linear Model**

++++ +



Final multi-linear model

= -2.109e-05 + 1.420e-03\*1+(-1.201e-03)\*0+1.846e-03\*0+(-1.322e-03)\*0+9.915e-01\*0+2.109e-05\*0+ (-1.596e-04)\*0+ (-1.317e-03)\*1+(-1.057e-03\*0)

= -9.8728961798

According to the multiple linear model developed for predicting the prognosis based on symptoms such as skin rash, high fever, headache, nausea, loss of appetite, pain behind the eyes, back pain, malaise, muscle pain, and red spots over the body:

The coefficients estimate the change in the predicted prognosis for a one-unit change in each predictor variable, while holding all other predictors constant. Notably, "pain\_behind\_the\_eyes" exhibits a highly significant effect on prognosis, with a coefficient estimate of 0.9915 and a large t-value of 233.530 (p < 0.001), indicating a strong positive association.

Other symptoms such as "high\_fever" also show a statistically significant impact on prognosis, with a coefficient estimate of 0.0025 (p = 0.0311), suggesting that an increase in high fever is associated with an increase in prognosis.

However, some symptoms like "skin\_rash," "headache," "nausea," "loss\_of\_appetite," "back\_pain," "malaise," "muscle\_pain," and "red\_spots\_over\_body" do not appear to have a statistically significant effect on prognosis, as indicated by their higher p-values.

Overall, the model explains a large portion of the variability in prognosis, with an R-squared value of 0.9747. This suggests that approximately 97.47% of the variance in prognosis can be accounted for by the included symptoms. Additionally, the adjusted R-squared value remains the same at 0.9747, indicating that the model's explanatory power is not affected by the number of predictors.

The residuals, representing the differences between the observed and predicted values of prognosis, exhibit a small range from -0.99171 to 0.99658, with a residual standard error of 0.02466, indicating a good fit of the model to the data.

**Future development and recommendation**

Currently, the dengue prediction system focuses on identifying whether an individual is infected with dengue fever at an early stage. However, future iterations of the system aim to expand its capabilities to categorize the specific serotype of the dengue virus, namely dengue virus serotype 1 (DENV-1), dengue virus serotype 2 (DENV-2), dengue virus serotype 3 (DENV-3), and dengue virus serotype 4 (DENV-4). This expansion is based on research findings reported in Nature news in 2022(Nature new, 2022). By incorporating these features into the system, it can provide more detailed and precise information about the specific type of dengue virus affecting an individual, enabling more targeted and effective treatment strategies.

**Conclusion**

In conclusion, this research explores the application of machine learning algorithms for dengue detection in Sri Lanka, a country where the disease poses a significant public health challenge. Through the utilization of various machine learning techniques such as classification algorithms, feature selection, and ensemble methods, aimed to develop accurate predictive models for early detection of dengue cases.

Based on the findings demonstrate that machine learning models, particularly those based on binary logistic regression, decision tree classifier, and ensemble method such as random forest exhibit promising performance in dengue prediction.

These models leverage diverse sets of features including environmental factors, demographic data, and historical dengue incidence rates to accurately classify individuals at risk of contracting dengue fever.

Furthermore, feature selection techniques such as Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA) have been instrumental in identifying the most relevant predictors for dengue detection, thereby enhancing the efficiency and Interpretability of the models.

However, it is essential to acknowledge several limitations and challenges encountered during the course of this study. These include the availability and quality of data, the dynamic nature of environmental factors influencing dengue transmission, and the need for real-time data integration to enhance the predictive accuracy of the models.

Despite these challenges, the results of this research hold significant implications for public health authorities and policymakers in Sri Lanka. By leveraging machine learning algorithms for dengue detection, it is possible to improve the early warning systems, allocate resources more efficiently, and implement targeted interventions to mitigate the burden of dengue fever in the country.

Moving forward, future research directions may involve the integration of additional data sources such as satellite imagery, climate forecasts, and social media data to enhance the predictive capabilities of the models. Additionally, the development of user-friendly and scalable predictive tools tailored to the needs of healthcare practitioners and decision-makers could facilitate the adoption of machine learning-based approaches in routine dengue surveillance and control efforts.

In conclusion, while there are challenges to be addressed, the application of machine learning in dengue detection offers tremendous potential to advance our understanding of disease dynamics and ultimately improve public health outcomes in Sri Lanka.

**References**

*Sri Lanka: Dengue Outbreak - May 2023* (no date) *ReliefWeb*. Available at: https://reliefweb.int/disaster/ep-2023-000084-lka (Accessed: 28 February 2024).

Pruszczyk, P. (2018) ‘Diagnosis: Clinical prediction rules and laboratory tests’, *ESC CardioMed*, pp. 2758–2761. doi:10.1093/med/9780198784906.003.0658.

Gupta, G. *et al.* (2023) ‘DDPM: A dengue disease prediction and diagnosis model using sentiment analysis and machine learning algorithms’, *Diagnostics*, 13(6), p. 1093. doi:10.3390/diagnostics13061093.

Jamwal, N. and Bhatia, S. (2021) ‘Prevalence of vector borne diseases in Jammu division, Jammu and Kashmir, India’, *Journal of Vector Borne Diseases*, 58(1), p. 28. doi:10.4103/0972-9062.321746.

Khan, Md.A. *et al.* (2022) ‘Dengue outbreaks prediction in Bangladesh perspective using distinct multilayer Perceptron NN and decision tree’, *Health Information Science and Systems*, 10(1). doi:10.1007/s13755-022-00202-x.

Sajana, T. *et al.* (2018) ‘Classification of dengue using Machine Learning Techniques’, *International Journal of Engineering &amp; Technology*, 7(2.32), p. 212. doi:10.14419/ijet.v7i2.32.15570.

Khan, J.R. and Raza, S.K. (2023) *Development and evaluation of a predictive diagnostic system for dengue fever using Machine Learning Techniques* [Preprint]. doi:10.21203/rs.3.rs-2473833/v2.

Al Nasar, M.R. *et al.* (2022) ‘Detection of dengue disease empowered with fused machine learning’, *2022 International Conference on Cyber Resilience (ICCR)* [Preprint]. doi:10.1109/iccr56254.2022.9996009.

Hoyos, W., Aguilar, J. and Toro, M. (2021) ‘Dengue models based on machine learning techniques: A systematic literature review’, *Artificial Intelligence in Medicine*, 119, p. 102157. doi:10.1016/j.artmed.2021.102157.

Chakraborty, T., Chattopadhyay, S. and Ghosh, I. (2019) ‘Forecasting dengue epidemics using a hybrid methodology’, *Physica A: Statistical Mechanics and its Applications*, 527, p. 121266. doi:10.1016/j.physa.2019.121266.

binti Mohd Zainee, N. and Chellappan, K. (2020) ‘A preliminary dengue fever prediction model based on vital signs and blood profile’, *2016 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES)* [Preprint]. doi:10.1109/iecbes.2016.7843530.

Lopez, D. *et al.* (2018) ‘An intelligent decision support system to prevent and control of Dengue’, *Journal of Ambient Intelligence and Humanized Computing* [Preprint]. doi:10.1007/s12652-018-0829-9.

Jayasundara, S.D.P., Perera, S.S.N. and Rathnayaka, N.S. (2017) ‘Developing a decision support testing algorithm to detect severity level of Dengue’, *European Scientific Journal*, 13(9). doi:10.19044/esj.2017.v13n9p137.

Rocha, F.P. and Giesbrecht, M. (2022) ‘Machine learning algorithms for dengue risk assessment: A case study for São Luís do Maranhão’, *Computational and Applied Mathematics*, 41(8). doi:10.1007/s40314-022-02101-z.

Katta, M. *et al.* (2022) ‘An efficient learning model selection for dengue detection’, *Smart Innovation, Systems and Technologies*, pp. 439–453. doi:10.1007/978-981-16-9873-6\_40.

Haque, M.A., Shivaprasad, G. and Guruprasad, G. (2021a) ‘Passenger Data Analysis of Titanic using machine learning approach in the context of chances of surviving the disaster’, *IOP Conference Series: Materials Science and Engineering*, 1065(1), p. 012042. doi:10.1088/1757-899x/1065/1/012042.

George, T. and Ganesan, V. (2021) ‘Advanced cuttlefish optimizer-random decision forest (ACORDF) based design of fractional order PID controller for higher-order time-delay system’, *Journal of Engineering, Design and Technology* [Preprint]. doi:10.1108/jedt-11-2020-0451.

Nordin, N.I. *et al.* (2020) ‘The classification performance using support vector machine for endemic dengue cases’, *Journal of Physics: Conference Series*, 1496(1), p. 012006. doi:10.1088/1742-6596/1496/1/012006.

(2022) *Nature news*. Available at: https://www.nature.com/scitable/topicpage/dengue-viruses-22400925/ (Accessed: 20 April 2024).

Pampel, F.C. (2021) *Logistic regression: A Primer* [Preprint]. doi:10.4135/9781071878729.