Diabetes Detection without Visiting the Clinic (A Machine Learning Approach)

Bahatheq Tariq Ahmed S

Undergraduate, Computer Science, National University of Singapore, Singapore

Email: tariqbahatheq@outlook.com

Venkat Vishwanth Sreenivasan

Undergraduate, Computer Science, National University of Singapore, Singapore

Email: vishwanth.s22@u.nus.edu

https://colab.research.google.com/drive/1KBDNEQZNKeiISultOQSOcTMTJVq50C65

Abstract:

The research presented in this paper focuses on the application of machine learning techniques for early detection of diabetes, without the need for clinic-dependent data. Utilizing a dataset of 253,680 examples sourced from the Behavioral Risk Factor Surveillance System (BRFSS), the study employs a variety of machine learning models, including Decision Tree, Random Forest, XGBoost, Neural Networks, SVM, and Naive Bayes. The paper highlights the significance of early diabetes detection and the potential of machine learning in making this process more accessible and efficient. The dataset underwent extensive preprocessing, including under-sampling to address imbalance and feature engineering to enhance model performance. The paper meticulously discusses the employed preprocessing techniques, providing insights into the importance of handling data imbalance and feature selection in machine learning applications for healthcare. The neural network model emerged as the top-performing model, achieving an accuracy of 88.76%. This result underscores the potential of machine learning in diabetes detection. We believe that this is fruitful as most people will avoid visiting the clinic to check for diabetes because of costs and loss of time. In conclusion, whilst we believe that this approach is beneficial, we suggest that this model only to be used as a possible indicator with the need to visit the doctor to fully confirm the presence of diabetes.

Keywords: Diabetes Detection, Machine Learning, Neural Networks, Feature Engineering, Data Preprocessing, Responsible AI, Transparency, Model Evaluation, Early Detection.

1. Introduction

Diabetes is a chronic health condition that affects how the body converts food into energy. It is a significant public health concern, with nearly half a billion people worldwide suffering from it. In Singapore, one in three individuals is at risk of developing diabetes, and it is estimated that about one million Singaporeans will be living with the condition by 2050. (1)

For those affected, the implications can be severe, including a lifetime of daily medication or injections, and, in some cases, blindness, amputation, kidney dialysis, and premature death. (1)

Early detection of diabetes can help save patients' health by enabling them to take medications and adopt routines that mitigate the condition's effects. Therefore, we have decided to employ machine learning to detect diabetes as the target variable in individuals based on commonly known factors, eliminating the need for a clinic visit.

2. The Dataset

Selected Dataset: The chosen dataset is sourced from the Kaggle website (2) and classifies individuals as having diabetes or not based on 21 features. It comprises 253,680 examples, with 218,334 of the examples representing people without diabetes and 35,346 examples representing those with the condition. The dataset was collected from the Behavioral Risk Factor Surveillance System (BRFSS), an annual health-related telephone survey conducted by the CDC. The data is anonymized for privacy reasons, and the 2015 dataset was used for this project. Table 1 below provides descriptions of the 21 features.

Alternative Datasets: We evaluated three potential datasets. The second dataset included pre-diabetes as a third classification, while the third dataset was under-sampled based on the target variable. We selected our dataset because the second one was highly imbalanced, and we preferred to perform under-sampling that better suited our dataset instead of the third option.

HighBP	HighChol	CholCheck	BMI	Smoke	Stroke	HeartDisease
High Blood Pressure	High Cholesterol	cholesterol check in 5 years	Body Mass Index	Smoked at least 100 cigarettes in your entire life?	had a stroke.	had/have (CHD) or (MI)

PhysActivity	Fruits	Veggies	Alcohol	Healthcare	NoDocbcCost	GenHlth
physical activity in the past 30 days - not including job	Consume Fruit 1 or more times per day	Consume Vegetables 1 or more times per day	Heavy drinkers	Any kind of health care coverage	Couldn't see a doctor in the past 12 months due to cost	General health: scale of 1-5 1 = excellent 5 = poor

MentHlth	PhysHlth	DiffWalk	Sex	Age	Education	Income
Bad mental health during the past 30 days	Bad physical health during the past 30 days	difficulty walking or climbing stairs	0 = female 1 = male	13-level age category	Education level scale 1-6	Income scale 1-8

Table 1: Dataset features with descriptions

Feature Selection: We removed the high blood pressure and high cholesterol features from the dataset to make the model accessible for anyone from their home, without the need for clinic visits, as these features require lab test confirmation. While this decision might affect the model's accuracy, we believe it is worthwhile as it increases the model's accessibility and usefulness for individuals. People can then use the model as an indicator of potential diabetes risk, prompting them to seek treatment before the condition worsens.

3. Data Preprocessing

3.1. Data Cleaning

The dataset selected has been pre-cleaned and has many features removed. Therefore, there are no missing data such as null values to consider and fix. However, the dataset is highly imbalanced where the minority class is diabetes label.

To overcome this imbalance, under-sampling was used. Under-sampling reduces the number of data points in the majority class till the majority class is approximately the same size as the minority class.

To be precise, NearMiss Algorithm was used to carry out the under-sampling by randomly removing data points in the majority class that are close to each other. Moreover, through this method, information loss is minimized as there still exists a data point that is quite similar to the one removed.

This resulted in a dataset of 70,692 entries with 35346 entries with diabetes label and 35346 entries with No Diabetes label.

3.2. Bias

The provided dataset was pre-reduced from 441,455 entries with 330 features, from the BRFSS, to 253,680 entries with 21 features. The selection criteria used to reduce the features might have been subjected to human bias. For example, in Representative Heuristic, one might incorrectly assume that BMI might be an important feature to judge diabetes based on stereotypes. However, it is not always true. As some might have obtained diabetes through genetics.

4. Data Visualization

Feature-Target Relationship: We visualized the relationship between each feature and the target variable using bar plots, as shown in Figure 1. Some correlations are readily apparent, such as the connection between BMI and diabetes, where both extremely low and high values are linked to the condition. General, physical, and mental health also show a strong association with diabetes, indicating that an unhealthy lifestyle is generally correlated with the disease. Furthermore, health conditions such as strokes, heart disease, or difficulty walking are also related to diabetes. We use these correlations to engineer features that will enhance the accuracy of our models.

Mutual Information Score (MI): A well-established metric for identifying relational information is the Mutual Information score. The advantages of MI include its ability to detect any relationship type (whereas correlation only detects linear relationships), ease of use and interpretation, computational efficiency, robust theoretical foundation, and resistance to overfitting[3]. The mutual information scores for the features are presented in Figure 2.

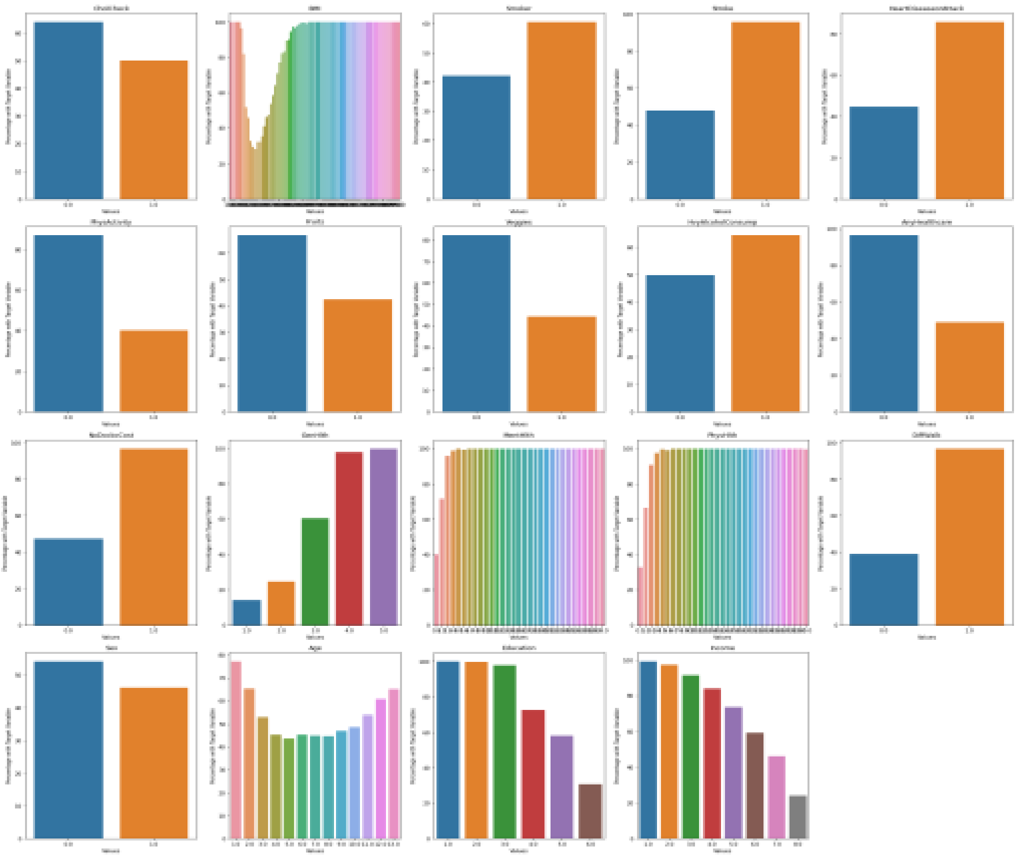

Figure 1: The bar plots show the percentages of people that have the target variable based on each value of the feature.

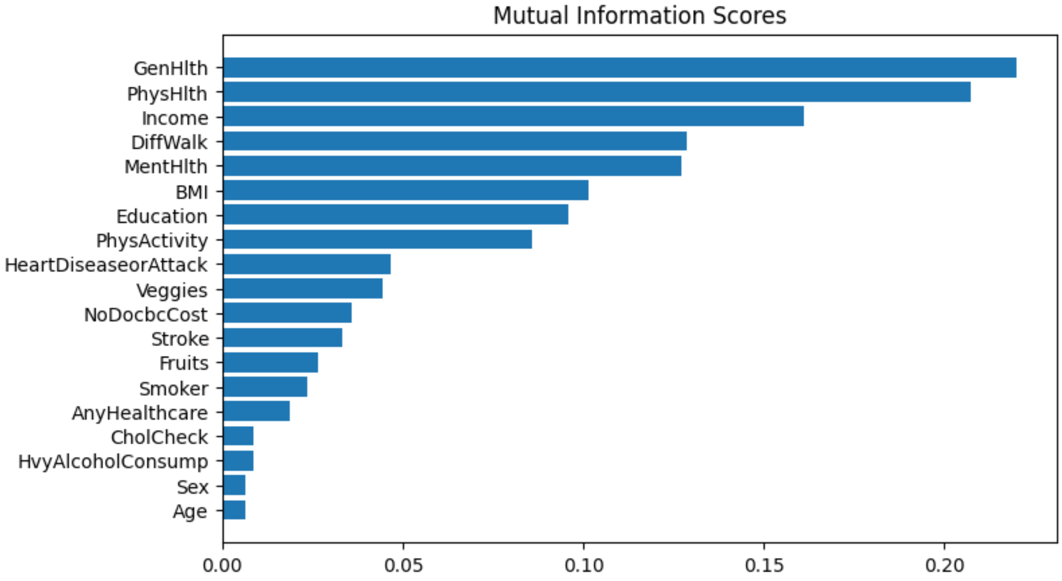

Figure 2: Mutual Information scores

5. Feature Engineering

Based on the relational information gathered from Figures 1 and 2, we experimented with combining new features that fall under categories such as physical health and physical activity, healthy lifestyle, diet, etc. We also explored exponentiating some features to examine potential non-linear relationships. The features we added are shown in Table 2:

PhysHA	PhysHlth + PhysActivity
GenAge	GenHlth * Age
class	Income * Education
BMIwalk	BMI * DiffWalk

Table 2: Feature engineering

Additionally, we experimented with removing some features based on their MI scores. However, this approach yielded slightly worse results. Therefore, we retained the original set of features along with the engineered ones to improve our model's performance. By incorporating these engineered features, we aimed to enhance the accuracy and predictive power of our machine-learning models.

Moreover, we split the dataset into 3 segments, Training, cross-validation, and Testing. This helps to overcome overfitting issues.

6. Model Selection and Performance Metrics

6.1. Model Selection

The target variable used for prediction is categorical. Hence, classification algorithms are required for prediction. We selected the most popular algorithms to determine the best algorithm that has the highest prediction accuracy.

6.2. Performance Metrics

To judge the performance of the classification algorithms, we opted to use Accuracy. Accuracy indicates the ratio of labels correctly predicted. Due to the nature of the dataset selected being balanced, Accuracy performs exceptionally well in this instance. Furthermore, this metric allowed us to better fine-tune the algorithm. A high accuracy would indicate that the model has a higher chance of predicting the label correctly and hence have good performance.

7. Models

7.1 Decision trees

The Model: Decision trees are reliable models for classification tasks such as this one. Furthermore, they are considered white-box models due to their interpretability.

Performance: Utilizing the default parameters for the decision tree from the sklearn library, we achieved an accuracy score of 83.60% on the validation dataset. This serves as a solid foundation for our model. However, further tuning is possible. The most critical parameters affecting decision trees are the error function and maximum leaf nodes. Automatically tuning the model on these attributes using the validation set yields an accuracy score of 86.28%. After feature engineering, the model delivers a score of 86.62%. The testing or generalization accuracy is 86.81%.

Limitations and Advantages: Unfortunately, decision trees are sensitive to changes in data, as slight alterations in the data can modify the tree's structure. On the other hand, decision trees are interpretable and computationally efficient.

7.2. Random Forest

The Model: The Random Forest model effectively compensates for the sensitivity of decision trees. As a tree ensemble, it uses sampling with replacement when constructing each tree, making it more robust.

Performance: Using the default parameters for the Random Forest from the sklearn library, we achieved an accuracy score of 86.00% on the validation dataset, which is already better than the default decision tree. Nonetheless, further tuning is possible. The most critical parameters affecting random forests are the error function, maximum leaf nodes, tree depth, and tree count. Automatically tuning the model on these attributes using the validation set yields an accuracy score of 88.98%. After feature engineering, the model delivers a score of 88.86%. The testing or generalization accuracy is 88.59% using the random forest without feature engineering as the validation accuracy was better.

Limitations and Advantages: Although random forests are more robust, there is still a chance of misrepresenting the dataset during random sampling with replacement, as some data might never be selected.

7.3. XGBoost

The model: XGBoost stands for "Extreme Gradient Boosting".(3) XGBoost is a decision tree ensemble that works together to compute the final prediction by summing up the predictions of multiple trees. Moreover, through the gradient boosting ensemble technique, weak prediction models, such as decision trees, provide a good overall prediction model. (4)

Performance: Using the default parameter settings, the model was able to give an accuracy of 89.00%. Hyperparameter tuning improved this accuracy to 89.16% on the validation data. Feature engineering further improved this to 89.19%. The generalization accuracy is 88.55%. The tuned parameters are shown in Table 3. (5)

Name	Best Value	Description
Colsample_bytree	0.3256299	Specify the fraction of columns to subsample when constructing each tree
Eta	0.8240329	Shrinks the feature weights to make boosting more conservative and prevent overfitting
gamma	8.3293339	The minimum split loss required to make a partition on a leaf node of the tree. A lower value makes the model less conservative (More likely to have False Positive and identify more True Positive predictions)
Max_depth	3789.0	Maximum depth of the tree. Increased depth makes the model more complex and likely to overfit
Min_child_weight	0.0	Determines the partitioning of the leaf node by comparing its minimum sum of instance weight. A lower value makes the model less conservative.
Reg_alpha	0.6574576	L1 regularization term on weights
Reg_lambda	7.2628099	L2 regularization term on weights

Table 3: XGBoost parameters

Limitations and Advantages: XGBoost excels in classification problems, however, it is not good at regression problems due to the use of decision trees that underperform with continuous inputs. An advantage is the ensemble technique used which allows individual models to help correct each other as opposed to all models training in isolation which might result in them making the same errors giving a better overall prediction.

7.4. Neural Networks

The Model: Neural networks are powerful and versatile models that can be used for a variety of tasks, including classification problems such as this one. They consist of interconnected layers of nodes or neurons that can learn complex patterns and relationships within the data.

Performance: Using a basic feedforward neural network with default parameters, with 3 layers and 16 neurons we trained the model for 100 epochs and obtained an accuracy score of 88.31% on the validation dataset. This provides a strong baseline for our model. However, we can further optimize the architecture and parameters. Key factors that affect neural networks include the number of layers, number of neurons per layer, activation functions, learning rate, and optimization algorithm. Tuning the model on these attributes using the validation set yields an accuracy score of 89.07%. The network architecture is shown in Table 4. The final layer is a linear layer because it avoids computational error. (6) After feature engineering, the model delivers a score of 89.34%. The testing or generalization accuracy is 88.76%.

Optimizer Function	Learning rate	Loss Function	Epochs
Adam	0.0001	Sparse Categorical Cross entropy	100
Layer	Neurons	Activation function	Dropout
L1	32	ReLU	0
L2	64	ReLU	0
L3	128	ReLU	0
L4	50	ReLU	0
L5	25	ReLU	0
L6	12	ReLU	0
L7	7	ReLU	0
L8	2	Linear	0

Table 4: Final neural network architecture

Limitations and Advantages: One major drawback of neural networks is their black-box nature, which makes them difficult to interpret compared to decision trees or random forests. Additionally, they can be more computationally intensive and time-consuming to train, especially for large datasets and complex architectures.

On the other hand, neural networks are capable of modelling complex relationships and can often achieve higher performance than other algorithms when properly tuned and trained. Furthermore, their flexibility allows for various architectures and activation functions to be tailored to the specific problem at hand. Moreover, a significant advantage is that they can be used for transfer learning meaning we can keep adding data to the model continuously without having to retrain the model.

7.5. SVM

The Model: Support Vector Machine (SVM) is well suited for classification problems where it finds a hyperplane in the (Num of features – 1) plane to aid in categorizing the data. (7) Additionally, it is using the Radial Bias Function (RBF) that uses the distance between the data points.

Performance: This model was able to achieve an accuracy of 88.00%. This was further improved to 89.20% by tuning the parameters. With c = 44.1658 and gamma = 74.3120 where c is the regularization parameter and gamma is the kernel coefficient (8) and feature engineering this slightly reduced to 89.19%. The generalization accuracy is 79.26%.

Limitations and Advantages: A huge drawback is the performance of SVM in large datasets. When trained on an initial large dataset, SVM took 50 minutes whereas other models only took a few minutes. Additionally, SVM doesn’t perform well if imbalanced data is used as this causes the hyperplane to be closer to the minority class resulting in wrong classification. Some advantages of SVM are its ability to handle high dimensional data as well as its robustness to noise in the data.

7.6. Naive Bayes

The Model: Naive Bayes applies the Bayes theorem while assuming that all features are conditionally independent of each other. Hence, the name naive. Due to feature independence, using the Bayes theorem allows the model to use lesser parameters to calculate and make classifications. (9)

Performance: This model was able to achieve an accuracy of 87.41%. By tuning the hyperparameter, alpha, an additive smoothing parameter to aid in the zero-probability problem. With alpha = 0.033446, this model has an accuracy of 87.48% on validation data. This was slightly reduced to 87.18% using feature engineering. The generalization accuracy of this model is 86.78%.

Limitation: Due to the conditional independence assumption, this model is not suitable to make probability estimations. Moreover, in a dataset that has features that are dependent on each other, this model's assumption would make it a poor choice as a classifier.

Moreover, Naïve Bayes has a zero-probability problem, where any feature not present in the training set will be automatically assigned a probability of zero, which prevents this model from making any predictions regarding that feature. (10)

One advantage is the ability of the model to train and predict fast due to the independence assumption made as well as the simple computations done to make predictions.

8. Conclusion

8.1. Best Model

By comparing the generalized accuracy of the models, Neural Network has the best-generalised accuracy of 88.76%, followed by Random Forest with 88.59%, then XGBoost with 88.55%, then Decision Tree with 86.81%, then Naive Bayes with 86.78% and finally SVM with 79.56%.

8.2. Responsible AI

Privacy: This model ensures privacy through the use of a de-identified dataset for training as well as not collecting personal particulars from the users for prediction. Hence, individual rights to privacy are well preserved.

Transparency: Despite the dataset used to train the Neural Network as well as the code for the neural network being released to the public, Neural Network is a Black Box Algorithm that doesn't provide the user with transparency of the decision that leads to the predictions. However, if transparency is necessary, the Decision Tree can be used as it can provide the actual tree used to make the decision.

Regardless of the transparency provided by the Decision tree, the general public would not be able to understand and use the decision tree to their advantage and hence, we would be trading off transparency for higher accuracy as wrong classification can affect the user more than the user having more transparency.

8.3. Limitations

There are three limitations present in this model. Firstly, due to the use of accuracy as an evaluation metric, the dataset has to be balanced. In the real world, some medical datasets have large imbalances, such as diabetes. Therefore, training this model to classify such datasets would result in a poor evaluation of the performance.

Secondly, the dataset provided consists of details of Americans only. This results in the model wrongly anchoring on select features to make predictions specific to Americans. Applying this model to the population might result in a higher probability of wrong predictions. For example, BMI would be a good indicator of diabetes for Americans however, in Singapore, sugar intake would be a good indicator of diabetes. Therefore, it is not recommended to generalize the predictions for the sample to the population that consists of people around the world.

Finally, our lack of experience in the medical field prevents us from truly ensuring that this model is trained and predicting diabetes correctly. Having sufficient knowledge can allow us to choose features more accurately and perform better feature engineering. Therefore, this model might not be truly accurate in predicting diabetes and would require the consultation of actual doctors to test the presence of diabetes.

Despite these limitations, we would still use this model to help predict diabetes in addition to visiting a doctor for a second opinion when the model predicts the presence of diabetes.

9. Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

10. Ethical Consideration:

This research prioritizes the ethical standards integral to scientific endeavors, particularly in the realm of medical research. Herein, we outline the key ethical considerations that were adhered to:

1. **No Human Experiments**: At no point did this study involve direct experiments on humans by the authors.
2. **Data Sources**: The data used in this research is sourced from a Kaggle challenge. It's crucial to understand that all data used was anonymized and void of any personal identifiers, ensuring the privacy and confidentiality of the responders.
3. **Patient Consent & Approvals**: While the authors did not directly conduct experiments or collect data, it's implicit that the original data collectors sought necessary consent from responders or their guardians.

4. **Protocols Followed**: The research strictly followed data handling and analysis protocols to ensure the integrity of the results. Furthermore, while the models and findings show promise, it's crucial to emphasize their supplementary role in medical diagnosis. Decisions based on these findings should be made with caution, in tandem with expert judgment.
5. **Transparency & Openness**: The research aims to contribute to the broader scientific community. As such, efforts have been made to ensure transparency in methodology, findings, and potential limitations. This open approach facilitates peer review and collective advancements in the field.

By adhering to these principles, this research aims to be both scientifically rigorous and ethically responsible, ensuring that advancements made contribute positively to patient care and the broader medical community.

11. References

1. MOH | News Highlights [Internet]. [cited 2023 Apr 7]. Available from: https://www.moh.gov.sg/news-highlights/details/speech-by-mr-ong-ye-kung-minister-for-health-at-world-diabetes-day-2021#:~:text=Locally%2C%20one%20in%20three%20individuals,will%20be%20living%20with%20diabetes
2. Diabetes Health Indicators Dataset | Kaggle [Internet]. [cited 2023 Apr 7]. Available from: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset
3. Introduction to Boosted Trees — xgboost 1.7.5 documentation [Internet]. [cited 2023 Apr 6]. Available from: https://xgboost.readthedocs.io/en/stable/tutorials/model.html
4. Gradient boosting. In: Wikipedia [Internet]. 2023 [cited 2023 Apr 6]. Available from: https://en.wikipedia.org/w/index.php?title=Gradient_boosting&oldid=1146274757
5. XGBoost Parameters — xgboost 2.0.0-dev documentation [Internet]. [cited 2023 Apr 6]. Available from: https://xgboost.readthedocs.io/en/latest/parameter.html#parameters-for-tree-booster
6. Google Colaboratory [Internet]. [cited 2023 Apr 7]. Available from: https://colab.research.google.com/github/tensorflow/docs/blob/master/site/en/tutorials/quickstart/beginner.ipynb#scrollTo=he5u_okAYS4a&line=1&uniqifier=1
7. 1.4. Support Vector Machines [Internet]. scikit-learn. [cited 2023 Apr 7]. Available from: https://scikit-learn/stable/modules/svm.html

8. sklearn.svm.SVC [Internet]. scikit-learn. [cited 2023 Apr 7]. Available from: https://scikit-learn/stable/modules/generated/sklearn.svm.SVC.html
9. 1.9. Naive Bayes [Internet]. scikit-learn. [cited 2023 Apr 7]. Available from: https://scikit-learn/stable/modules/naive_bayes.html
10. Jayaswal V. Laplace smoothing in Naïve Bayes algorithm [Internet]. Medium. 2020 [cited 2023 Apr 7]. Available from: https://towardsdatascience.com/laplace-smoothing-in-na%C3%AFve-bayes-algorithm-9c237a8bdece

Doi: https://doi.org/10.52132/Ajrsp.e.2023.55.2