Machine Learning for Disease Prediction

Nadia Paquin
Gustavo Gyotoku
Victoria Haley

June 16, 2023
IST707

Introduction

The issue of diagnostic error within the medical field has proven to be a significant and expensive problem, with far-reaching consequences for both healthcare providers and patients. Research, such as the Harvard Medical Practice Study conducted in hospitals across New York State, has revealed that diagnostic errors constitute a substantial portion of medical errors and rank as the second largest cause of adverse events (Leape et al.). Misdiagnosis can have severe consequences, exacerbating patients' conditions, delaying proper treatment for their ailments, and potentially leading to unnecessary complications. Furthermore, these errors are financially burdensome for both the patient and the caregiver– misdiagnoses represent the second leading cause of malpractice suits against hospitals (Bartlett). Addressing this issue is of paramount importance, as emphasized by Nancy W. Dickey, MD, former president of the American Medical Association, who asserted at the 1998 Annenberg Conference on Patient Safety that a zero percent error rate is the only acceptable standard (Arkes). While achieving a flawless diagnosis record may seem like an ambitious objective, it is undoubtedly a worthwhile goal to pursue, given the high costs, detrimental consequences, and hardships faced by patients as a result of misdiagnosis.

As such, our project aimed to develop a robust machine learning model for predicting a likely prognosis based on reported symptoms, offering significant implications for the healthcare industry. With numerous variables involved in disease diagnosis, accurate prediction models based on symptoms and health factors are critical. Our objectives were to build a machine learning model for accurate diagnosis prediction, identify the most important reported symptoms to guide prioritization, and demonstrate the impact of our technology in real-world problem-solving.

To ensure reliable and valid results, our project followed a structured experimental design. We selected the “Disease Prediction using Machine Learning" dataset from Kaggle, containing diverse patient records with comprehensive symptoms and prognoses. We performed data preprocessing, including exploratory data analysis (EDA), data validation, data normalization, and dataset splitting. We explored various machine learning algorithms, such as decision trees, random forest, k-Nearest Neighbors (KNN), and Naive Bayes, and selected the best-performing model for disease prediction. Additionally, feature importance analysis helped identify influential symptoms for prognosis prediction and informed future assessments.

Following a rigorous and systematic approach, our experimental design aimed to draw meaningful conclusions and make predictions in medical diagnostics. In the subsequent sections, we will present our findings, including the performance of our machine learning models, the significance of symptoms in prognosis prediction, and recommendations for improving diagnostic accuracy and reducing errors in healthcare. By addressing the challenges of disease

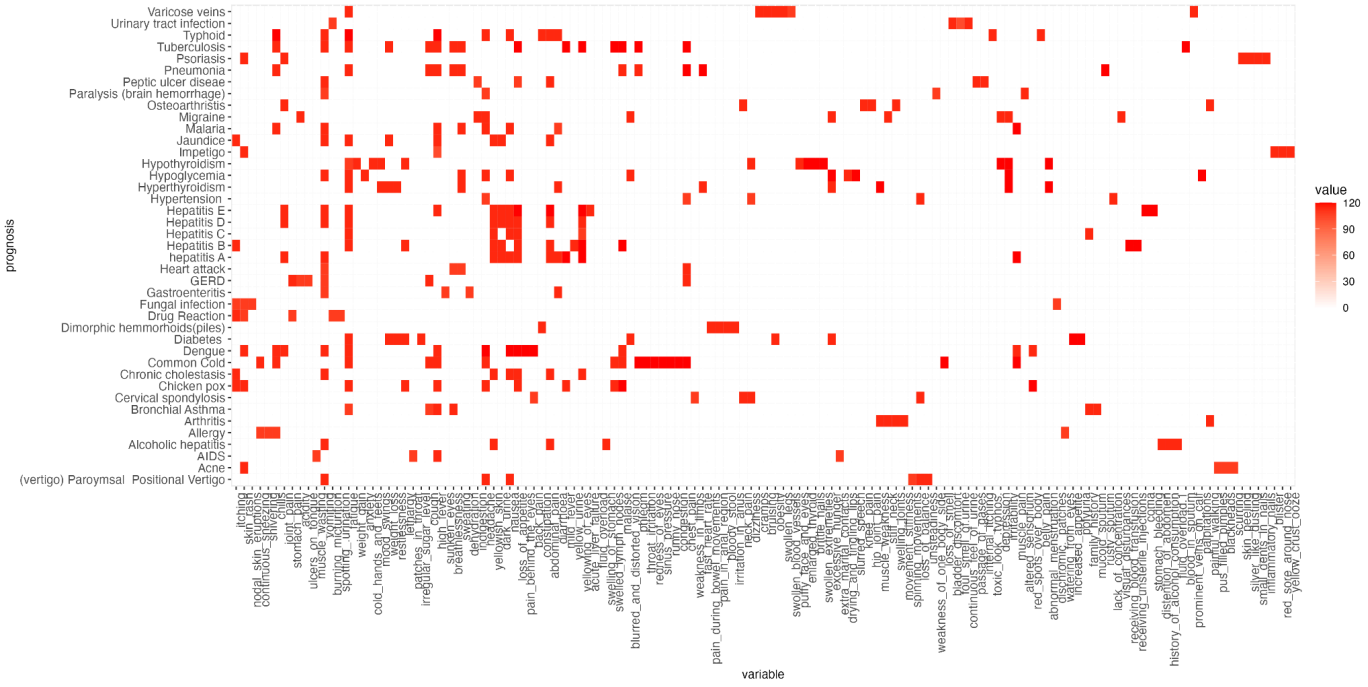
diagnosis through machine learning, our project strives to enhance healthcare practices, improve diagnostic accuracy, and ultimately enhance patient care and outcomes.

Methods

About the data:

The dataset used in this report, titled "Disease Prediction using Machine Learning," was sourced from Kaggle. It encompasses a comprehensive collection of health information for patients, including an array of symptoms and their corresponding prognoses. Structured as a binary matrix, each row signifies a patient, while each column denotes a specific symptom. With a total of 4920 observations (patients) and 132 variables (symptoms), the dataset provides a substantial and diverse sample to explore diagnostic patterns. Notably, the dataset's final column represents the corresponding prognosis, offering 41 distinct outcomes to study and analyze. This dataset serves as a valuable resource for investigating the potential applications of machine learning in disease prediction and diagnostic decision-making.

Upon examining the dataset, several noteworthy observations come to light. First, it becomes evident that each prognosis is associated with multiple symptoms, underscoring the complexity and multifaceted nature of disease manifestation. Furthermore, it is apparent that certain symptoms are common across different prognoses, indicating their non-specificity and potential overlap in various medical conditions. This emphasizes the importance of considering a wide range of symptoms and employing sophisticated algorithms to accurately predict diagnoses. Additionally, the starkness of symptom prevalence within the dataset is striking. The high incidence rates, illustrated by the intense red color on the symptom-prognosis association map (see figure below), make it relatively straightforward to determine strong associations. However, in real-world scenarios, setting association thresholds or employing tiered categories to capture varying degrees of association may be necessary. Fortunately, in this binary dataset, such nuanced considerations are not required, simplifying the analysis process. Overall, these observations shed light on the intricacies and challenges involved in leveraging this dataset for disease prediction and highlight the need for careful analysis and interpretation when applying machine learning techniques to real-world healthcare scenarios.



Analyses:

During the initial analysis, the focus was on trying different techniques to gain insights. The first algorithm used was KNN, which aimed to identify groups that were close to each other based on proximity. After running the code and running a grid search for the optimal number of groups, KNN identified 10 as the optimal number of groups. However, this was far from the 43 diseases that were already identified in the dataset. Given the result, the methodology was revisited with the concern that it was due to some oversight. The obvious culprit was the fact that the columns were all boolean values of 1s and 0s. Such values do not work well with the default parameter of the Euclidean distance.

A correlation test was administered to the dataset to see any relationship between certain variables. The initial thought that certain symptoms would be close to each other and the disease associated with them would also correlate. This was not the case as previously stated, boolean values do not have linear relationships, so a correlation test would not draw any concrete results. During this time in data exploration, the “hail-mary” approach of throwing the data into different tests without forming a good process resulted in insightful results and wasted time.

Furthermore, before identifying one of many issues with the methodology, a baseline was set for the classification models. The goal was to try different models to see if they could accurately identify the diseases. If not, the accuracy could be increased through tuning methods. For the data validation portion of the process, the most important metric was the overall accuracy. Given that the dataset was discrete in nature, the results would not generate many false positives or true positives. Even if the classifier did, it would not be significant enough to track.

In creating the process, one of the important points was that our test should be able to be replicated by others. For this reason, the “seed” was set to “seed(123).” This attribute would

recreate the results. Another important issue that came to light was the run speeds. To prevent unworkable runspeeds, the “makeCluster” function was utilized. This function creates other “workers” assigned to a CPU core. They would then work in parallel in running the functions provided - in this case the classifiers decision tree, naive bayes, randomforest, and KNN model.

Thus, the group decided to first utilize the decision tree classifier to identify the variables that would have the most impact on the test dataset. The idea was that these symptoms would be useful in increasing the accuracy. The “varImp” function from the “rpart” library reported the top most impactful variables by a numeric value. These variables were saved to an object which was then fed to the remaining classifiers Naive Bayes, Random Forest and KNN. The KNN classification model contained the original data without the top significant variables. As expected, the KNN model resulted in a perfect overall score, a sign that the dataset was overfitting. Moreover, the algorithm selection had the greatest impact. The KNN model had the lazy approach, as it did not assume anything about the data, such as correlation, distribution, or other characteristics. This proved to be useful as the data itself was not overly simplistic and had very little complexity.

Results

As stated, we utilized a large Kaggle dataset and employed a variety of machine learning techniques to predict diagnoses. The accuracies achieved by each method, along with brief descriptions, are as follows:

Decision trees construct a hierarchical structure of decision rules based on the provided dataset. The decision tree algorithm yielded a prediction accuracy of 74%. While this approach demonstrated promising results, it may be sensitive to data imbalance and prone to overfitting. Further fine-tuning could enhance its diagnostic capabilities.

Naive Bayes classifiers are probabilistic models that assume independence among features. The Naive Bayes algorithm exhibited a considerably higher accuracy of 89%. This technique's ability to handle large and complex datasets, along with its assumption of feature independence, contributed to its success in predicting diagnoses.

Random Forest is an ensemble learning method that combines multiple decision trees. Each tree is trained on a different subset of the data and features, and predictions are aggregated to generate the final result. This approach mitigates the risk of overfitting and improves generalization. Employing the Random Forest algorithm resulted in a remarkable accuracy of 93%. Random Forest demonstrated significant potential in accurately classifying and predicting diagnoses, leveraging the power of an ensemble of decision trees.

KNN is a non-parametric algorithm that classifies instances based on their similarity to other instances in the dataset. It identifies the k-nearest neighbors and assigns a class label based on the majority vote. Among the machine learning techniques employed, KNN achieved the highest accuracy of 99%. The exceptional performance of KNN in diagnosing various conditions can be attributed to its ability to find the most similar instances in the dataset and make predictions based on local patterns.

These results underscore the effectiveness of machine learning techniques in diagnosing medical conditions based on the provided dataset. Notably, KNN emerged as the most accurate method, closely followed by Random Forest and Naive Bayes. The Decision Tree approach, while demonstrating potential, may benefit from further refinement to improve its diagnostic accuracy. These findings highlight the value of leveraging machine learning algorithms in healthcare settings to enhance diagnostic accuracy and improve patient outcomes.

Conclusions

In conclusion, our disease prediction model has demonstrated impressive efficiency in predicting the provided data within the current patient population. It has delivered accurate results, demonstrating its potential as a valuable tool in the healthcare industry. This highlights the potential of machine learning in assisting disease prediction and diagnosis based on symptom data. Accurate prediction of common diseases can significantly impact the patient outcomes and healthcare decision-making.

However, it is important to note that our dataset covers only 43 diseases, while there are at least 10,000 documented diseases, according to the Washington Post (2016). This emphasizes the need for additional data collection and analysis to enhance the model's effectiveness across a broader population.

While our model performs well within the dataset, its ability to predict outcomes for the broader population cannot be guaranteed. The observed decrease in efficacy when predicting conditions outside of our patient population suggests a potential bias. This highlights the need for further investigation and improvements to ensure a more comprehensive and inclusive model.

Moving forward, we recommend focusing on practical applications and research insights to maximize the impact of our disease prediction model. By incorporating the model into clinical decision support systems, healthcare professionals can make more accurate and timely decisions, improving diagnosis and treatment planning. Additionally, leveraging the model's predictive capabilities can help identify high-risk patients and prioritize early interventions, potentially leading to improved patient outcomes and reduced healthcare costs.

Furthermore, it is crucial to assess the similarity or dissimilarity between the distributions of the training and testing sets using statistical analysis such as ANOVA. This insight will guide potential adjustments in the modeling process, ensuring the model's reliability and generalizability across diverse patient populations.

In summary, our disease prediction model demonstrates proficiency within the current patient population, but further improvements are needed for wider applicability. By addressing the model's limitations, incorporating it into clinical decision support systems, prioritizing early interventions, and conducting thorough statistical analyses, we can enhance its effectiveness and contribute to improved healthcare practices and patient outcomes. Our work presents the

potential of machine learning in the medical field and sets the stage for future research and collaboration in this important area.

References

1. Leape L, Brennan TA, Laird N, et al. The nature of adverse events in hospitalized patients. Results of the Harvard Medical Practice Study II. N Engl J Med. 1991;324:377–84.
2. Bartlett EE. Physicians' cognitive errors and their liability consequences. J Healthcare Risk Manage. Fall 1998:62–9.
3. Arkes H. Why medical errors can't be eliminated: uncertainties and the hindsight bias. Chron Higher Educ. May 19, 2000.
4. Kessler, G. (2016, November 17). Are there really 10,000 diseases and just 500 “cures”? The Washington Post. https://www.washingtonpost.com/news/fact-checker/wp/2016/11/17/are-there-really-10000-diseases-and-500-cures/

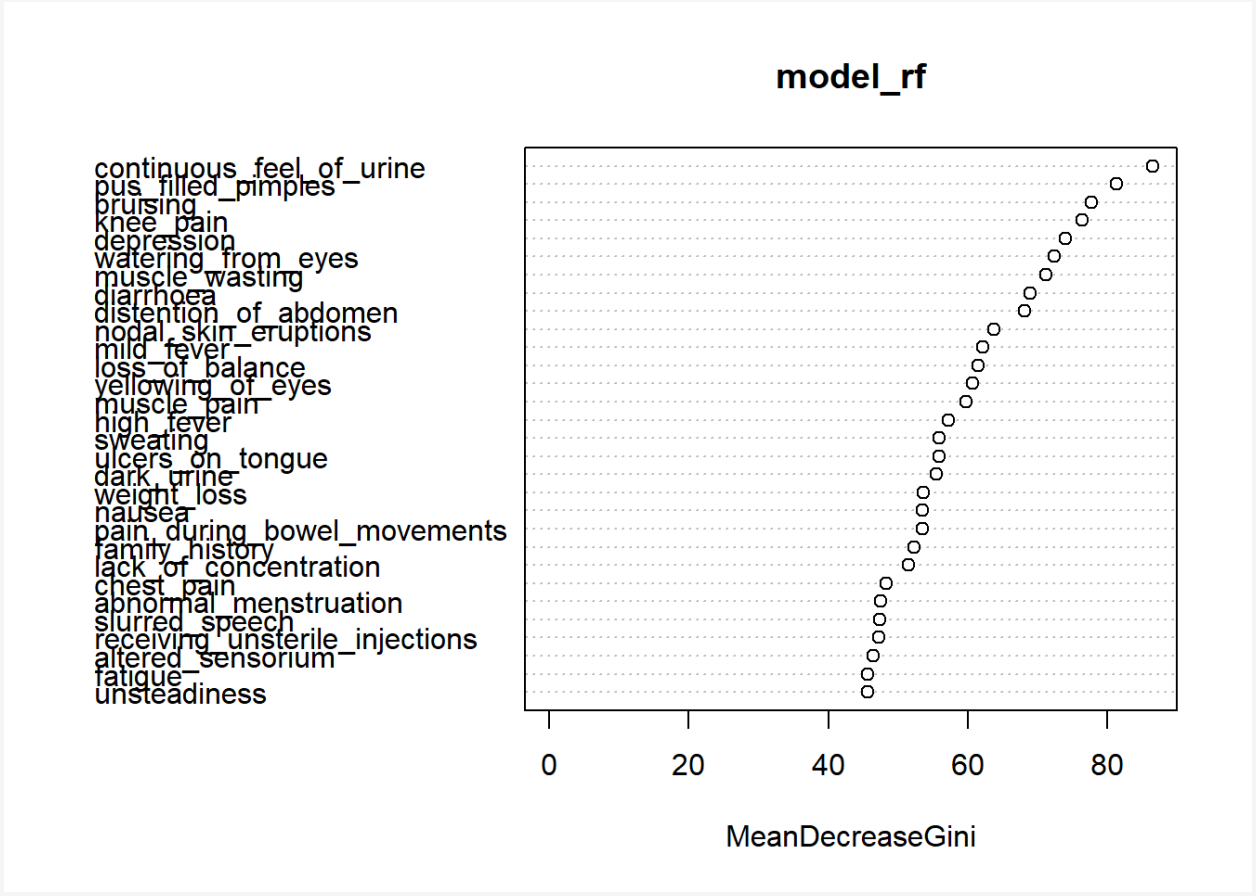
Code

```
# cor_matrix <- cor(features, method = "pearson")
# cor_matrix[is.na(cor_matrix)] <- 0
# dend <- hclust(dist(cor_matrix))
# reordered_matrix <- cor_matrix[rev(order.dendrogram(as.dendrogram(dend))),
#                                rev(order.dendrogram(as.dendrogram(dend)))]
#
# jpeg("heatmap.jpg", width = 1800, height = 1200, res=150)
# heatmap(cor_matrix)
# dev.off()
#
# plot(dend)
#
# jpeg("dendrogram_large.jpg", width = 1800, height = 1200)
# plot(dend, cex = 0.8)
# height_cutoff <- 1
# clusters <- cutree(dend, h=height_cutoff)
# rect.hclust(dend, k =max(clusters), border = clusters)
# dev.off()
#
# # corrplot(cor_matrix, method="color", addClustColors = TRUE)
#
# performClustering <- function(cor_matrix) {
#   methods <- c("euclidean", "manhattan", "maximum", "minkowski", "canberra", "binary", "correlation",
"spearman")
#
#   for (x in methods) {
#     dend <- hclust(dist(cor_matrix, method = x))
#     jpeg(paste0("dendrogram_", x, ".jpg"), width = 1800, height = 1200)
```

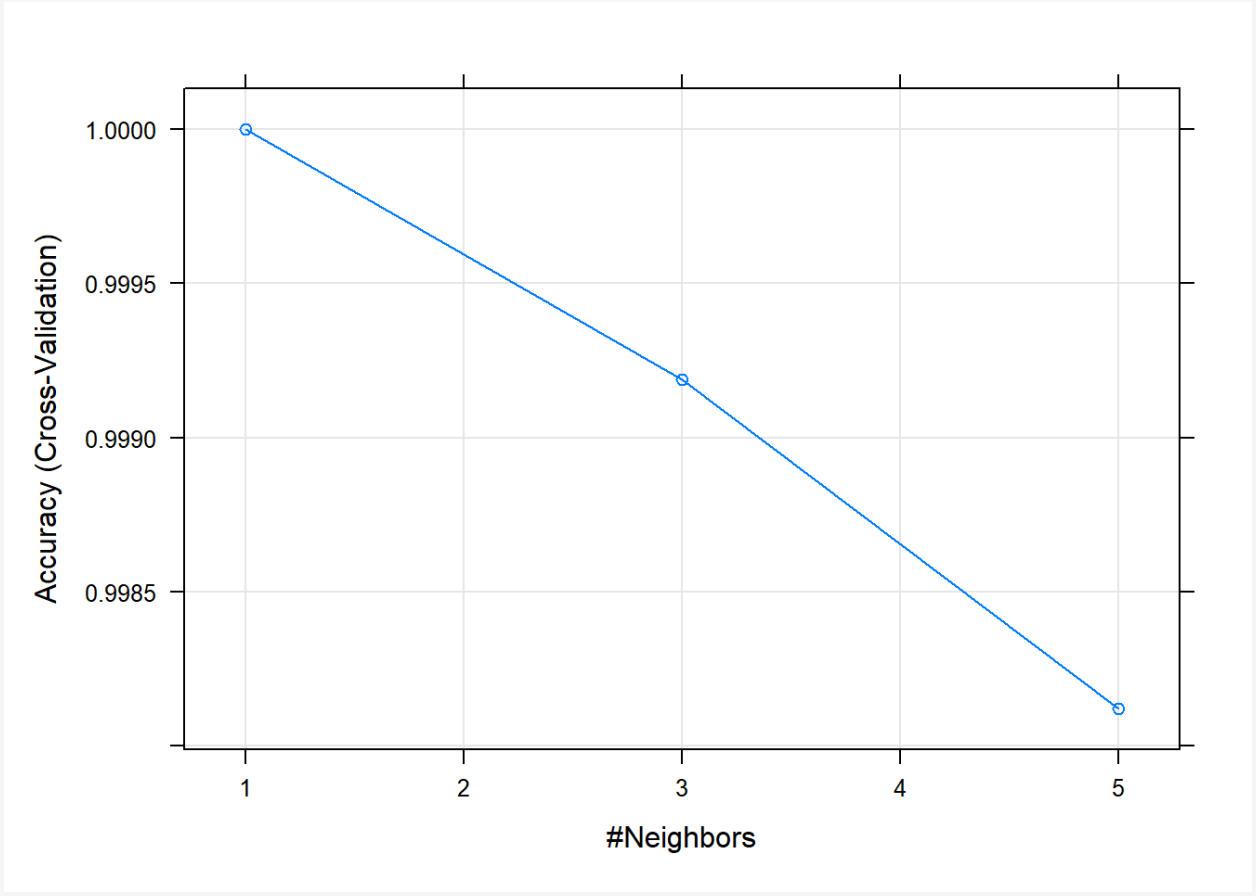
```
#     plot(dend, main = paste("Hierarchical Clustering (Method:", x, ")"))
#     height_cutoff <- .2
#     clusters <- cutree(dend, h=height_cutoff)
#     rect.hclust(dend, k =max(clusters), border = clusters)
#     dev.off()
#   }
# }
# # I'm calling the function
# performClustering(cor_matrix)
#
# # I'm running only the Minkowski method because I couldn't set the p in the function created earlier
# dend <- hclust(dist(cor_matrix, method = "minkowski", p=10))
# jpeg(paste0("dendrogram_", "minkowski", ".jpg"), width = 1800, height = 1200)
# plot(dend, main = paste("Hierarchical Clustering (Method:", "minkowski", ")"))
# height_cutoff <- .2
# clusters <- cutree(dend, h=height_cutoff)
# rect.hclust(dend, k =max(clusters), border = clusters)
# dev.off()

set.seed(123)

cl <- makeCluster(12)

data <- df3

encoding <- as.numeric(data$prognosis)

data$prognosis <- encoding

data$prognosis <- as.factor(data$prognosis)

# Dropping variable "floud_overload" as it causes NAs in the dataset
# This variable only contains 0s

# ------------------------------------------------------------------------

train_index <- sample(nrow(data), .75 * nrow(data))

train <- data[train_index, ]

test <- data[-train_index, ]

# ------------------------------------------------------------------------
model_rpart <- rpart(prognosis~., data=train, control = list(method="cv", k=100))

## Warning in rpart(prognosis ~ ., data = train, control = list(method = "cv", :
## NAs introduced by coercion

predict_rp <- predict(model_rpart, newdata = test, type="class")
```

```
accuracy_rp <- sum(predict_rp == test$prognosis) / nrow(test)

print(paste("The overall accuracy is:", accuracy_rp))

## [1] "The overall accuracy is: 0.742143432715552"

# Selecting important variables to feed to other models

var_imp <- varImp(model_rpart)

var_imp2 <- var_imp %>% arrange(desc(Overall))

var_imp2 <- var_imp2 %>% filter(Overall >0)

var_imp2 <- rownames(var_imp2)

# Top important variables:

var_imp2

##  [1] "sweating"                  "family_history"
##  [3] "ulcers_on_tongue"          "fast_heart_rate"
##  [5] "watering_from_eyes"        "blood_in_sputum"
##  [7] "swelled_lymph_nodes"       "red_sore_around_nose"
##  [9] "receiving_blood_transfusion"   "receiving_unsterile_injections"
## [11] "knee_pain"                 "lack_of_concentration"
## [13] "depression"                "altered_sensorium"
## [15] "rusty_sputum"              "inflammatory_nails"
## [17] "nodal_skin_eruptions"      "muscle_wasting"
## [19] "muscle_pain"               "internal_itching"
## [21] "passage_of_gases"          "bruising"
## [23] "continuous_feel_of_urine"      "yellow_crust_ooze"
## [25] "pus_filled_pimples"        "abnormal_menstruation"
## [27] "increased_appetite"        "polyuria"
## [29] "coma"                      "redness_of_eyes"
## [31] "throat_irritation"         "palpitations"
## [33] "slurred_speech"            "blister"
## [35] "distention_of_abdomen"     "chest_pain"
## [37] "stomach_bleeding"          "pain_behind_the_eyes"
## [39] "unsteadiness"              "brittle_nails"
## [41] "enlarged_thyroid"          "swollen_extremeties"
## [43] "congestion"                "runny_nose"
## [45] "sinus_pressure"            "irritation_in_anus"
## [47] "pain_during_bowel_movements"   "phlegm"
## [49] "red_spots_over_body"       "yellowing_of_eyes"
## [51] "weight_loss"               "weakness_of_one_body_side"
## [53] "silver_like_dusting"       "mild_fever"
## [55] "dark_urine"                "loss_of_balance"
```

```
## [57] "diarrhoea"                "mucoid_sputum"
## [59] "yellowish_skin"           "fatigue"
## [61] "nausea"                   "high_fever"
## [63] "breathlessness"

# ------------------------------------------------------------------------
nb = naiveBayes(prognosis~., data = train[, c(var_imp2, "prognosis")],laplace = 1, na.action=na.pass)

pred_nb = predict(nb, newdata = test, type = "class")

accuracy_nb = mean(pred_nb == test$prognosis)

print(paste("The Naive Bayes overall accuracy is:", accuracy_nb))

## [1] "The Naive Bayes overall accuracy is: 0.89041095890411"

print(paste("The overall accuracy has improved by:",round((accuracy_nb - accuracy_rp) *100, 2),"% !"))

## [1] "The overall accuracy has improved by: 14.83 % !"

# ------------------------------------------------------------------------

train_control_rf <- trainControl(method="cv", number=50)

model_rf <- randomForest(prognosis ~ ., data=train[, c(var_imp2, "prognosis")], mtry=sqrt(63), ntree=100,
trControl = train_control_rf)

predict_rf <- predict(model_rf, test)

accuracy_rf = mean(predict_rf == test$prognosis)

print(paste("The Naive Bayes overall accuracy is:", accuracy_rf))

## [1] "The Naive Bayes overall accuracy is: 0.935535858178888"

varImpPlot(model_rf)
```




```
# ------------------------------------------------------------------------
train_control_knn <- trainControl(method = "cv", number = 100,
search = "grid")

# Contains only zeroes
train <- subset(train, select = -fluid_overload)

model_knn <- train(prognosis ~ ., data = train, method = "knn",
trControl = train_control_knn,
tuneGrid = expand.grid(k = c(1, 3, 5)),
metric = "Accuracy",
preProcess = c("center", "scale"))

predict_knn <- predict(model_knn, test)

accuracy_knn <- mean(predict_knn==test$prognosis)

print(paste("The KNN overall accuracy is:", accuracy_knn))

## [1] "The KNN overall accuracy is: 0.999194198227236"

plot(model_knn)
```


1.0000
0.9995
0.9990
0.9985
Accuracy (Cross-Validation)
1
2
3
4
5
#Neighbors

Slides

Disease Prediction Using Machine Learning

Enhancing Diagnostic Accuracy

Problem statement:

Develop a recommendation system tailored for medical practitioners to use as a diagnostic tool given an array of health conditions and symptoms affecting patients.

Goals:

- Leverage clustering and classification methodologies to categorize patients according to their health metrics
- Subsequently, construct a proficient recommendation system capable of proposing potential diagnoses by considering a patient's symptoms and medical history
- Assess the efficacy and utility of the system by conducting user testing

Experimental Design

Dataset Selection

- Selected the “Disease Prediction using Machine Learning" from Kaggle for its diverse and representative records with prognoses and their respective symptoms

Data Processing

- EDA, dataset splitting, data validation, data normalization

Model Selection and Evaluation

- Explored various ML algorithms (decision trees, random forests, clustering)
- Trained multiple models on the training data
- Evaluated performance using confusion matrix accuracy

Feature Importance Analysis

- Identified influential symptoms for prognosis predictions
- Prioritized specific symptoms in future assessments

Dataset Description

- Dataset: "Disease Prediction using Machine Learning" from Kaggle
- Contains health information for patients including symptoms and corresponding prognoses
- Binary matrix structure: each row represents a patient, each column represents a symptom
- 4920 observations (patients) and 132 variables (symptoms)
- Last column represents corresponding prognosis (41 distinct)
- Goal: use machine learning techniques to predict patient prognosis based on symptoms

Prognoses

Infectious Diseases: Influenza, Tuberculosis, Malaria, Hepatitis, Measles, Cholera, Dengue Fever, Ebola Virus Disease, HIV/AIDS, Zika Virus Infection

Autoimmune Diseases: Rheumatoid Arthritis, Lupus, Multiple Sclerosis, Type 1 Diabetes, Graves' Disease, Hashimoto's Thyroiditis, Psoriasis, Crohn's Disease, Ulcerative Colitis

Cardiovascular Diseases: Coronary Artery Disease, Stroke, Heart Failure, Aortic Aneurysm, Peripheral Arterial Disease, Deep Vein Thrombosis, Pulmonary Embolism, Hypertension

Respiratory Diseases: Asthma, Chronic Obstructive Pulmonary Disease (COPD), Pneumonia, Tuberculosis, Pulmonary Fibrosis, Lung Cancer

Neurological Diseases: Alzheimer's Disease, Parkinson's Disease, Multiple Sclerosis, Epilepsy, Amyotrophic Lateral Sclerosis (ALS), Huntington's Disease, Migraine, Brain Tumor

Digestive System Diseases: Irritable Bowel Syndrome (IBS), Crohn's Disease, Ulcerative Colitis,Gastroesophageal Reflux Disease (GERD), Peptic Ulcer Disease, Gallstones, Pancreatitis, Celiac Disease

Symptoms

Dermatological symptoms: itching, skin rash, nodal skin eruptions, patches in throat, redness of eyes, sinus pressure, runny nose, congestion, pimples, blackheads, skin peeling, silver-like dusting, small dents in nails, inflammatory nails, blister, red sore around nose, yellow crust ooze.

Gastrointestinal symptoms: stomach pain, acidity, ulcers on tongue, vomiting, burning micturition, spotting during urination, constipation, abdominal pain, diarrhea, pain during bowel movements, pain in anal region, irritation in anus, foul smell of urine, continuous feel of urine, passage of gases, belly pain, stomach bleeding, distention of abdomen.

Respiratory symptoms: cough, breathlessness, phlegm, throat irritation, chest pain, wheezing, blood in sputum.

Neurological symptoms: headache, dizziness, neck pain, altered sensorium, loss of balance, unsteadiness, weakness of one body side, slurred speech, lack of concentration, visual disturbances, coma.

Metabolic and systemic symptoms: fatigue, weight gain, weight loss, lethargy, irregular sugar level, high fever, malaise, excessive hunger, depression, irritability.

Musculoskeletal symptoms: joint pain, muscle wasting, muscle weakness, stiffness in neck and limbs, knee pain, hip joint pain, painful walking.

Training data

132 symptoms

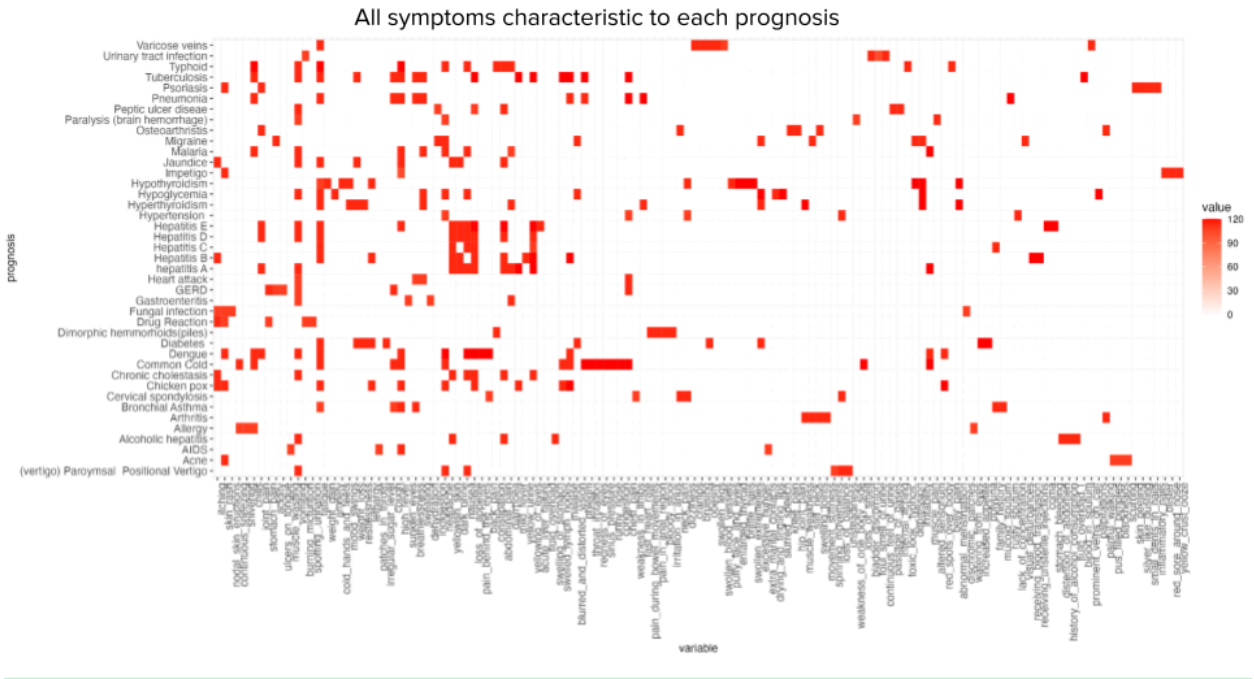
41 distinct prognoses

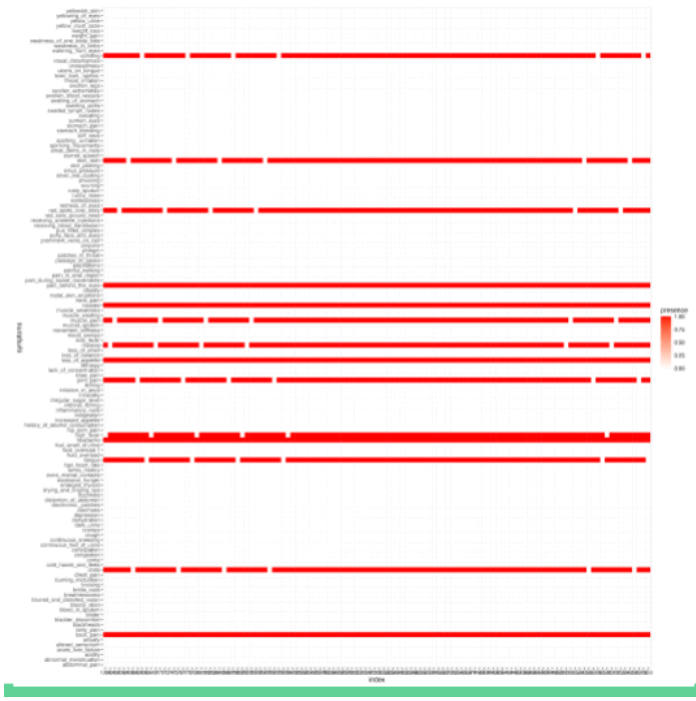
silver_like_dusting	small_dents_in_nails	inflammatory_nails	blister	red_sore_around_nose	yellow_crust_ooze	prognosis
0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	Allergy
0	0	0	0	0	0	Allergy
0	0	0	0	0	0	Allergy
0	0	0	0	0	0	Allergy
0	0	0	0	0	0	Allergy

120 patients are affected by each prognosis

Binary values represent presence or absence of each symptom

4920 patients





All cases of dengue (x) and its associated symptoms (y)

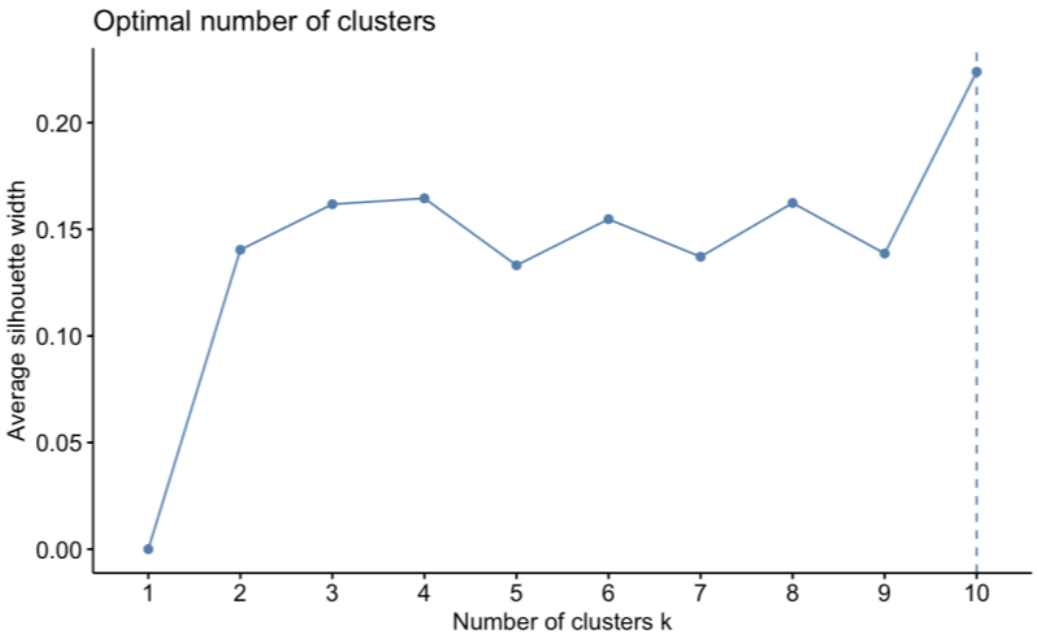
- No false positives
- Few cases with false negatives

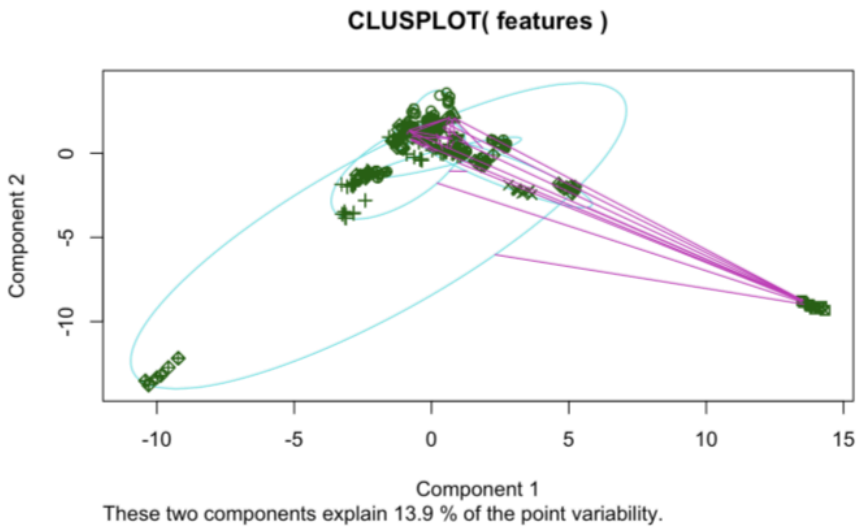
K-means clusters (unsupervised)

	model.cluster	num_prognoses	prognoses
1	1	3	Allergy, GERD, Gastroenteritis
2	2	1	Cervical spondylosis
3	3	1	Peptic ulcer diseae
4	4	1	Osteoarthristis
5	5	1	Alcoholic hepatitis
6	6	1	Arthritis
7	7	1	Hypertension
8	8	1	Impetigo
9	9	1	Dimorphic hemmorhoids(piles)
10	10	2	Chronic cholestasis, Hepatitis C
11	11	1	Urinary tract infection
12	12	1	Varicose veins
13	13	1	Diabetes
14	14	1	Hyperthyroidism
15	15	1	AIDS
16	16	2	Malaria, Dengue
17	17	1	Arthritis

Many clusters with multiple ailments

Ailments falling under multiple clusters





Decision Tree

```
set.seed(123)
cl <- makeCluster(12)
data <- df3
encoding <- as.numeric(data$prognosis)
data$prognosis <- encoding
data$prognosis <- as.factor(data$prognosis)

# ------------------------------------------------------------------------

train_index <- sample(nrow(data), .75 * nrow(data))
train <- data[train_index, ]
test <- data[-train_index, ]

# ------------------------------------------------------------------------
model_rpart <- rpart(prognosis~., data=train, control = list(method="cv", k=100))
predict_rp <- predict(model_rpart, newdata = test, type="class")
accuracy_rp <- sum(predict_rp == test$prognosis) / nrow(test)
print(paste("The overall accuracy is:", accuracy_rp))
# Selecting important variables to feed to other models
var_imp <- varImp(model_rpart)
var_imp2 <- var_imp %>% arrange(desc(Overall))
var_imp2 <- var_imp2 %>% filter(Overall >0)
var_imp2 <- rownames(var_imp2)
# Top important variables:
var_imp2
```

The overall accuracy of the decision tree is 74%

Decision Tree

Using the varImp() function, the decision tree algorithm determined these variables have the most impact on the prediction accuracy.

These variables will be weighted higher in the training phase as the other variables dropped had an impact factor of 0.

```
> var_imp2
 [1] "sweating"                       "family_history"                 "ulcers_on_tongue"
 [4] "fast_heart_rate"                "watering_from_eyes"             "blood_in_sputum"
 [7] "swelled_lymph_nodes"            "red_sore_around_nose"           "receiving_blood_transfusion"
[10] "receiving_unsterile_injections" "knee_pain"                      "lack_of_concentration"
[13] "depression"                     "altered_sensorium"              "rusty_sputum"
[16] "inflammatory_nails"             "nodal_skin_eruptions"           "muscle_wasting"
[19] "muscle_pain"                    "internal_itching"               "passage_of_gases"
[22] "bruising"                       "continuous_feel_of_urine"       "yellow_crust_ooze"
[25] "pus_filled_pimples"             "abnormal_menstruation"          "increased_appetite"
[28] "polyuria"                       "coma"                           "redness_of_eyes"
[31] "throat_irritation"              "palpitations"                   "slurred_speech"
[34] "blister"                        "distention_of_abdomen"          "chest_pain"
[37] "stomach_bleeding"               "pain_behind_the_eyes"           "unsteadiness"
[40] "brittle_nails"                  "enlarged_thyroid"               "swollen_extremeties"
[43] "congestion"                     "runny_nose"                     "sinus_pressure"
[46] "irritation_in_anus"             "pain_during_bowel_movements"    "phlegm"
[49] "red_spots_over_body"            "yellowing_of_eyes"              "weight_loss"
[52] "weakness_of_one_body_side"      "silver_like_dusting"            "mild_fever"
[55] "dark_urine"                     "loss_of_balance"                "diarrhoea"
[58] "mucoid_sputum"                  "yellowish_skin"                 "fatigue"
[61] "nausea"                         "high_fever"                     "breathlessness"
```

Naive Bayes

```
nb = naiveBayes(prognosis~., data = train[, c(var_imp2, "prognosis")],laplace = 1, na.action=na.pass)
pred_nb = predict(nb, newdata = test, type = "class")
accuracy_nb = mean(pred_nb == test$prognosis)
print(paste("The Naive Bayes overall accuracy is:", accuracy_nb))
print(paste("The overall accuracy has improved by:",round(accuracy_nb - accuracy_rp, 2),"% !"))
```

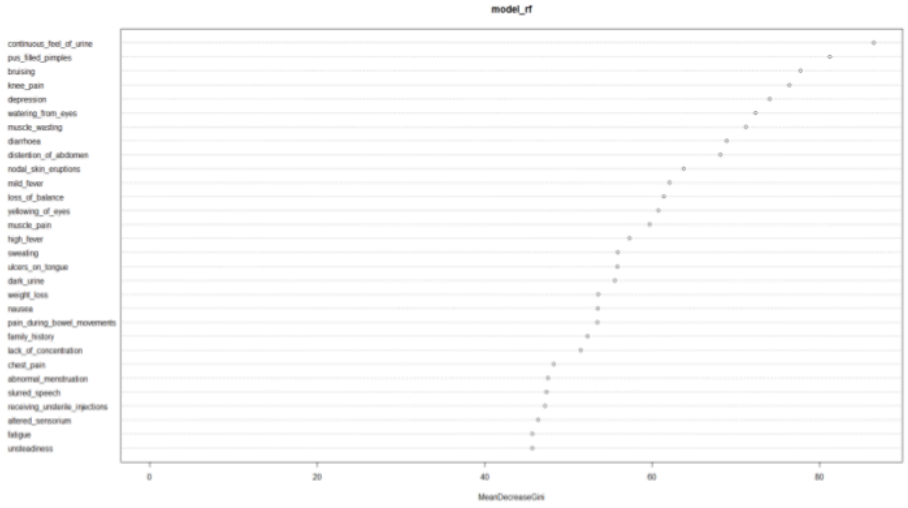
The overall accuracy of the Naive Bayes is 89%. That is a 14.83% increase in the overall accuracy.

Random Forest

```
train_control_rf <- trainControl(method="cv", number=50)
model_rf <- randomForest(prognosis ~ ., data=train[, c(var_imp2, "prognosis")], mtry=sqrt(63), ntree=100, trControl=train_control_rf)
predict_rf <- predict(model_rf, test)
accuracy_rf = mean(predict_rf == test$prognosis)
print(paste("The Naive Bayes overall accuracy is:", accuracy_rf))
varImpPlot(model_rf)
```

The overall accuracy of the Naive Bayes is 93%. Not surprising as it is an assemble algorithm.

Random Forest



The higher the Gini mean score means the higher its impact to the overall score.

KNN

```
train_control_knn <- trainControl(method = "cv", number = 10,
search = "grid")
model_knn <- train(prognosis ~ ., data = train, method = "knn",
trControl = train_control_knn,
tuneGrid = expand.grid(k = c(1, 3, 5)),
metric = "Accuracy",
preProcess = c("center", "scale"))
predict_knn <- predict(model_knn, test)
accuracy_knn <- mean(predict_knn==test$prognosis)
print(paste("The KNN overall accuracy is:", accuracy_knn))
print(model_knn)
```

For this model, we incorporated all the variables and applied a grid search to fine-tune and preprocess it.

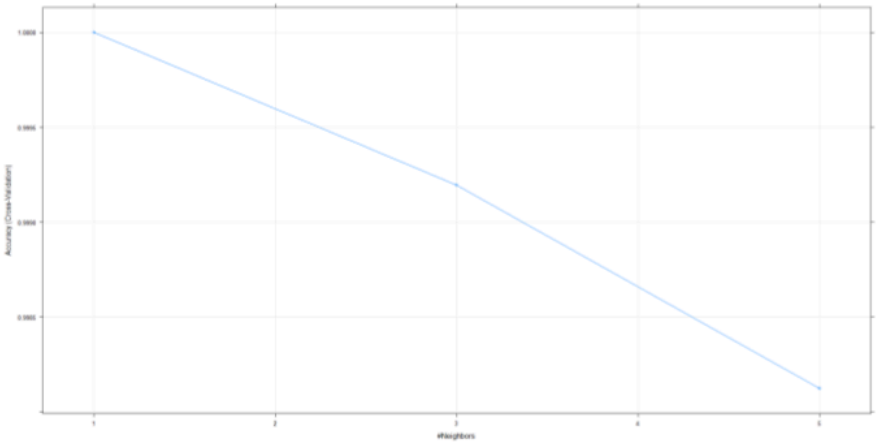
Given the results of the previous test, there is a possibility that the model is overfitting. Overfitting occurs when a model fits the training data extremely well but fails to generalize to new, unseen data. This suggests that our model may be excessively complex or overly tailored to the specific characteristics of the training dataset.

Additionally, while the model demonstrates excellent performance in capturing the patients within the dataset, it is essential to consider its representativeness of the overall population. The model's ability to accurately predict outcomes for the broader population cannot be guaranteed solely based on its performance on the current dataset.

To ensure the model's reliability and generalizability, further evaluation and validation on external data sources are recommended. This will help establish the model's effectiveness and validity beyond the confines of the current dataset.

The overall accuracy of the Naive Bayes is 99%.

KNN



The low number of k neighbors suggests that the data may be overfitting, particularly when considering k values in the range of 1 to 3.

Conclusions

- The model demonstrates proficiency in predicting the provided data. However, its efficacy diminishes when applied to conditions beyond the current patient population. This observation indicates a bias towards predicting specific conditions present in the dataset, highlighting a limitation in generalization. To effectively encompass the broader population, additional investigation is required, considering that the dataset covers only 43 diseases while there are at least 10,000 diseases documented (Washington Post, 2016).

Key Findings

- ML can effectively assist in disease prediction and diagnosis based on symptom data
- Accurate prediction of common diseases can significantly impact patient outcomes and healthcare decision-making

Recommendations for Real-World, Business, and Medical Use

Clinical Decision Support

- Incorporate the disease prediction model into clinical support systems to assist healthcare professionals in making accurate and timely decisions

Early Intervention Planning

- Utilize the model to identify high-risk patients and prioritize early interventions, potentially improving patient outcomes and reducing healthcare costs

Research Insights

Additionally, it is crucial to acknowledge that the training set and the testing set may not necessarily belong to the same population. Therefore, to assess the similarity or dissimilarity between the distributions of the training and testing sets, an ANOVA test can be employed. This statistical test will provide valuable insights into whether there are significant differences in the distributions of the two sets, highlighting the need for careful consideration and potential adjustments in the modeling process.

Reference

nnet: Venables, W. N., & Ripley, B. D. (2002). Modern Applied Statistics with S. Springer.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

L., K. (2020). Disease Prediction using Machine Learning [Dataset]. Kaggle. Retrieved Month Day, Year, from https://www.kaggle.com/datasets/kaushil268/disease-prediction-using-machine-learning

Thanks!