Data Science Project Write-Up

By: Sameed Mubasher and Vanessa Mpofu

Project Overview

Our project focused on exploring a diabetes dataset to uncover potential relationships between various health indicators and the onset of diabetes. Diabetes is a significant health concern worldwide, and early detection is critical for effective management and prevention. Several conditions in the human body can indicate the onset of diabetes for an individual and these can range from varying body mass index (BMI) and blood glucose levels to whether a person smokes or not. We chose this area because it combines a real-world health issue with the opportunity to analyze meaningful data, which could provide insights into factors contributing to diabetes risk. Initially, we aimed to address questions such as:

- Are women more likely to have diabetes after controlling for BMI and other risk factors?
- What are the odds of having diabetes based on BMI, heart disease, and hypertension?

While we initially hoped to answer these questions, as we explored and analyzed our dataset, we came across more interesting questions that caught our attention. We decided to focus on those instead, which is why the questions we ultimately addressed in our project are different from the ones we started with.

Data

year	gender	age	location	race:AfricanAmerican	race:Asian	race:Caucasian	race:Hispanic	race:Other	hypertension	heart_disease	smoking_history	bmi	hbA1c_level	blood_glucose_level	diabetes
2020	Female	32	Alabama	0	0	0	0	1	0	0	never	27.32	5	100	0
2015	Female	29	Alabama	0	1	0	0	0	0	0	never	19.95	5	90	0
2015	Male	18	Alabama	0	0	0	0	1	0	0	never	23.76	4.8	160	0
2015	Male	41	Alabama	0	0	1	0	0	0	0	never	27.32	4	159	0
2016	Female	52	Alabama	1	0	0	0	0	0	0	never	23.75	6.5	90	0
2016	Male	66	Alabama	0	0	1	0	0	0	0	not current	27.32	5.7	159	0
2015	Female	49	Alabama	0	0	1	0	0	0	0	current	24.34	5.7	80	0
2016	Female	15	Alabama	0	0	0	0	1	0	0	No Info	20.98	5	155	0
2016	Male	51	Alabama	1	0	0	0	0	0	0	never	38.14	6	100	0
2015	Male	42	Alabama	0	0	1	0	0	0	0	No Info	27.32	5.7	160	0
2016	Male	15	Alabama	1	0	0	0	0	0	0	No Info	19.15	6.6	200	0
2016	Female	53	Alabama	0	0	1	0	0	0	0	never	34.3	6.6	155	0
2015	Female	3	Alabama	1	0	0	0	0	0	0	No Info	20.28	3.5	159	0
2016	Female	40	Alabama	0	0	0	1	0	0	0	never	27.63	6.5	126	0

Fig.1: Raw data

We got our diabetes dataset from Kaggle. Fig.1 above shows a snippet of the raw dataset. The dataset includes various health indicators such as BMI, blood glucose levels, age, and smoking history, which are relevant to predicting diabetes. Some of the general strengths of this dataset include its large size, with over 100,000 data points and sixteen variables, making it suitable to use for analysis. Additionally, the dataset contains a mix of numerical and categorical variables, such as gender, heart disease, and smoking history, which helped us address specific research

questions. The general weaknesses of the dataset included the need for significant cleaning and preprocessing, such as removing N/A entries and mutating/combining columns. Additionally, the race variable was divided into five separate columns, which not only complicated the analysis but could also introduce bias into the models.

To prepare the dataset for analysis, we addressed several wrangling issues, including preprocessing, cleaning, handling missing values, and resolving inconsistencies. The first step was to remove the ***blood_glucose_level*** variable, as it is a direct predictor of diabetes and would have led to biased models. Next, we focused on the ***smoking_history*** variable, which initially had six different categories: “never”, “ever”, “former”, “current”, “not current”, and “no info”. To clean this variable, we first mutated all “No Info” entries to N/A and then removed the missing values. Afterward, we simplified the remaining categories by grouping them into three meaningful classes: “current”, “never”, and “former”. The race variables posed another challenge, as they were represented in multiple binary columns (0-no, 1-yes) indicating whether a participant belonged to a specific race.

```r
# Replace values using mutate and case_when
diabetes_dataset <- diabetes_dataset %>%
  mutate(smoking_history = case_when(
    smoking_history == "No Info" ~ NA,
    smoking_history == "ever" ~ "current",
    smoking_history == "not current" ~ "former",
    TRUE ~ smoking_history # Keep original value if no match
  ))

View(diabetes_dataset)
dim(diabetes_dataset)
dim(na.omit(diabetes_dataset))



clean_dataset <- (na.omit(diabetes_dataset))
dim(clean_dataset)
# Rename variables for easier handling
colnames(clean_dataset)[colnames(clean_dataset) == "race:AfricanAmerican"] <- "race_AfricanAmerican"
colnames(clean_dataset)[colnames(clean_dataset) == "race:Asian"] <- "race_Asian"
colnames(clean_dataset)[colnames(clean_dataset) == "race:Caucasian"] <- "race_Caucasian"
colnames(clean_dataset)[colnames(clean_dataset) == "race:Hispanic"] <- "race_Hispanic"
colnames(clean_dataset)[colnames(clean_dataset) == "race:Other"] <- "race_Other"
# Clean dataset
attach(clean_dataset)
names(clean_dataset)
```

Fig.2: Data Wrangling, preprocessing code

We renamed the race variables for clarity, as shown in Fig. 2. Then, we combined all the race variables into a single categorical column, after ensuring that each entry was appropriately assigned a race category and only one category. Lastly, to align with the requirements of certain machine learning techniques, we recorded the ***diabetes*** variable into a binary format where 0 represented “yes” and 1 represented “no”. These were the steps we took to ensure that the dataset was clean, consistent, and suitable for further analysis and modeling.

After completing the preprocessing and cleaning steps, we proceeded to analyze our dataset using various machine-learning techniques that we learned in class to answer our research questions. The techniques we applied included Linear Regression, Multiple Linear Regression, Naïve Bayes, and K-means clustering. Additionally, we ran several classification models, including the Null Model, Nearest Neighbor, Support Vector Machines (SVM), Decision Trees, Random Forest, and Logistic Regression. These models allowed us to explore the relationships between the variables, identify key predictors of diabetes, and evaluate the overall performance of each technique. The details of the model's performance, results, and comparisons will be discussed in the following sections.

Linear Regression

Using this technique, we answered the question: Is there a linear relationship between BMI and hbA1c levels? We have included an image of the model summary (Fig.3) and a visual representation in Fig.4.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.1586679  0.0192823  267.53   <2e-16 ***
bmi         0.0142700  0.0006612   21.58   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.092 on 64182 degrees of freedom
Multiple R-squared:  0.007204,  Adjusted R-squared:  0.007189
F-statistic: 465.7 on 1 and 64182 DF,  p-value: < 2.2e-16
```

Fig.3: Linear Regression Model Summary

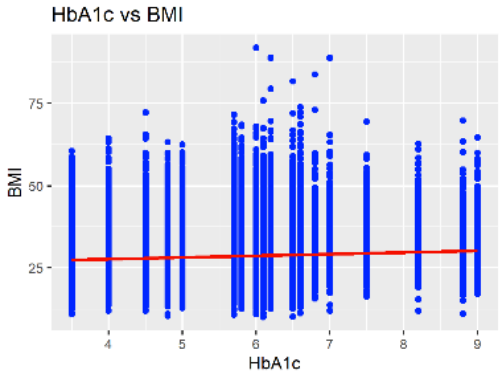

Fig.4: hbA1c vs bmi plot

From the final model, we obtained the equation $hbA1c\_level = 5.15 + 0.014 * bmi$. This suggests a weak but positive relationship between BMI and HbA1c levels, as indicated by the small coefficient of (0.014) for ***bmi***. While the ***bmi*** predictor does have a statistically significant effect on ***hbA1c_level***, the model explains only 0.72% of the variance. This very low percentage indicates that ***bmi*** alone is not a strong predictor of ***hbA1c_level***, and there are likely other factors contributing to variations in ***hbA1c_level***.

Multiple Linear Regression

Using the MLR technique, we answered the question: Which variables provide the best prediction of diabetes risk? For this, we computed 6 models with different predictor variables using the drop-one method. The code used can be seen below.

```r
#2. Multiple Regression - Which variables provide the best prediction of diabetes risk?
# Obviously blood_glucose_level does, so we take it out immediately

multi_reg1 <- lm(data=clean_dataset,diabetes~. -(blood_glucose_level) ) #Adjusted R-squared:  0.2733
summary(multi_reg1)
drop1(multi_reg1,test="F")

summary(lm(data=clean_dataset,diabetes~.-(blood_glucose_level+location))) #Adjusted R-squared:  0.2733
summary(lm(data=clean_dataset,diabetes~.-(blood_glucose_level+location+year))) #Adjusted R-squared:  0.2733
summary(lm(data=clean_dataset,diabetes~.-(blood_glucose_level+location+year+`race:AfricanAmerican`+
             `race:Asian` + `race:Caucasian`+ `race:Hispanic` + `race:Other`+smoking_history))) #Adjusted R-squared:  0.2733
summary(lm(data=clean_dataset,diabetes~.-(blood_glucose_level+location+year+`race:AfricanAmerican`+
       `race:Asian` + `race:Caucasian`+ `race:Hispanic` + `race:Other`+gender+smoking_history))) #Adjusted R-squared:  0.2724

summary(lm(data=clean_dataset,diabetes~.-(blood_glucose_level+location+year+`race:AfricanAmerican`+
  `race:Asian` + `race:Caucasian`+ `race:Hispanic` + `race:Other`+gender+smoking_history+age)))#Adjusted R-squared:  0.253
```

Fig.5: Multiple Linear Regression Models

In Fig.5 above we see that the models that best fit our data would be any of the first 4 because they have the same Adjusted R-squared of 0.2733, despite removing variables such as location and year. Because the 4th model removes a lot of insignificant predictors and still retains an Adjusted R-squared of 0.2733, our multiple linear regression model would be:

$$diabetes = -0.8737 + 0.04278 \cdot genderFemale + 0.06210 \cdot genderMale + 0.002389 \cdot age + 0.09691 \cdot hypertension + 0.1280 \cdot heart\_disease + 0.006293 \cdot bmi + 0.1128 \cdot hbA1c\_level$$

```
Call:
lm(formula = diabetes ~ . - (blood_glucose_level + location +
    year + `race:AfricanAmerican` + `race:Asian` + `race:Caucasian` +
    `race:Hispanic` + `race:Other` + smoking_history), data = clean_dataset)

Residuals:
     Min       1Q   Median       3Q      Max
-0.66844 -0.16542 -0.07007  0.07120  1.02079

Coefficients: (1 not defined because of singularities)
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -8.737e-01  7.723e-02 -11.312   <2e-16 ***
genderFemale   4.278e-02  7.695e-02   0.556    0.578
genderMale     6.210e-02  7.696e-02   0.807    0.420
genderOther           NA         NA      NA       NA
age            2.389e-03  5.767e-05  41.418   <2e-16 ***
hypertension   9.691e-02  3.687e-03  26.289   <2e-16 ***
heart_disease  1.280e-01  5.150e-03  24.861   <2e-16 ***
bmi            6.293e-03  1.646e-04  38.228   <2e-16 ***
hbA1c_level    1.128e-01  9.715e-04 116.080   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2665 on 64176 degrees of freedom
Multiple R-squared:  0.2733,    Adjusted R-squared:  0.2733
F-statistic:  3449 on 7 and 64176 DF,  p-value: < 2.2e-16
```

Fig.6: Selected Multiple Linear Regression Model Summary

From the summary image (Fig. 6), we can see that genderFemale and genderMale are not statistically significant but do impact the Adjusted R-squared if removed. The predictor variable genderOther shows an output of NA because it is the reference category in the model.

Naïve Bayes

This model was used to answer the question: Given that a patient has a higher BMI, has heart disease, and is older, what is the probability that the patient has diabetes or not? To address this, the dataset was first split into training and testing sets. A model was built using the training dataset and then tested on the testing dataset to evaluate its performance. We calculated the ***age*** and ***bmi*** thresholds by just taking the average of each. The results showed that the overall probability of a person being diabetic was 11%, while the probability of not being diabetic was 89%. When applying specific conditions, ***bmi*** greater than or equal to 28. 4, the presence of ***heart_disease*** (which is 1), and ***age*** greater than or equal to 47, the probability of being diabetic increased to 44%, while the probability of not being diabetic was 56%. The model achieved a high accuracy of 99. 72%, largely due to the class imbalance in the dataset, where the majority class (0, no diabetes) greatly outnumbered the minority class (1, diabetes). This imbalance contributed to the model's accuracy being skewed toward predicting the majority class correctly. Despite this limitation, the model demonstrated its ability to estimate the likelihood of diabetes under specific health conditions, providing useful insights into the relationship between BMI, heart disease, age, and diabetes risk.

K-Means

Next, we ran the K-means clustering model to answer the question: *How do health indicators and demographic factors differ across the identified clusters, and what insights can be drawn about high-risk or unique subgroups?* To address this, we first needed to determine the optimal number of clusters for the model. To identify the best number of clusters, we used the elbow method and obtained the graph in Fig.7. The best number of clusters according to the elbow graph is 4 because the Within-Cluster Sum of Squares (WCSS) decreases rapidly at first but starts to flatten around 4 clusters.

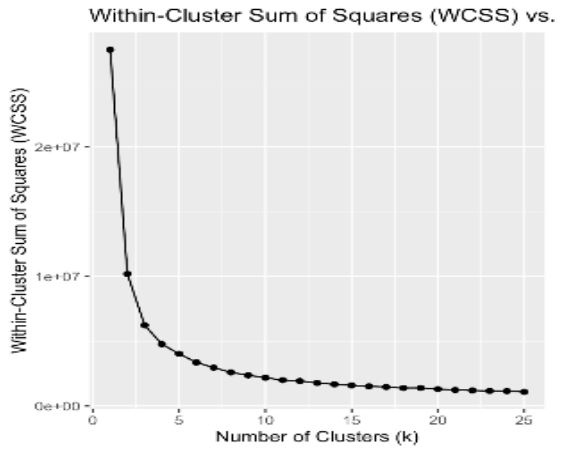

Fig.7: Elbow Graph

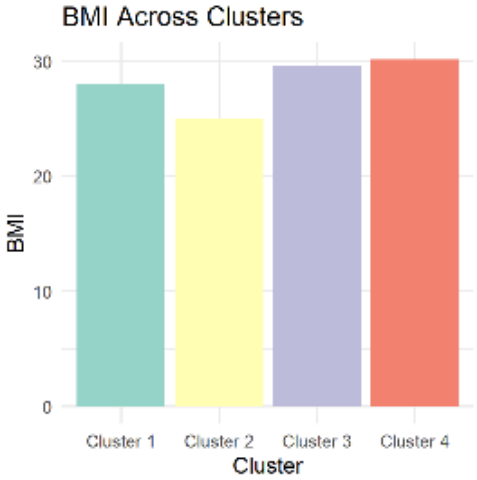

Fig.8.1

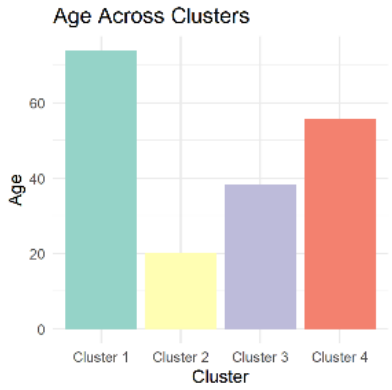

Fig.8.2

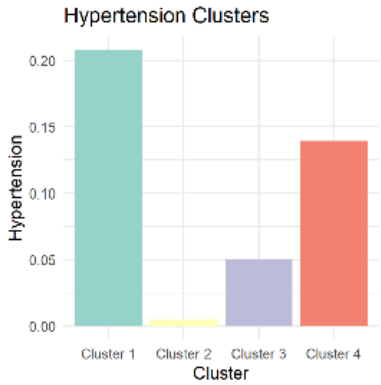

Fig.8.3

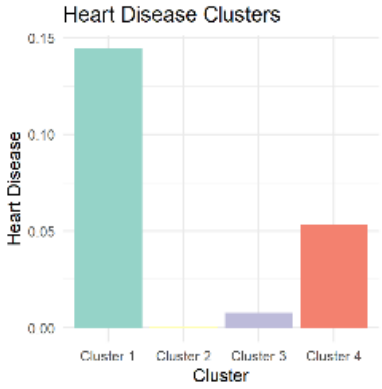

Fig.8.4

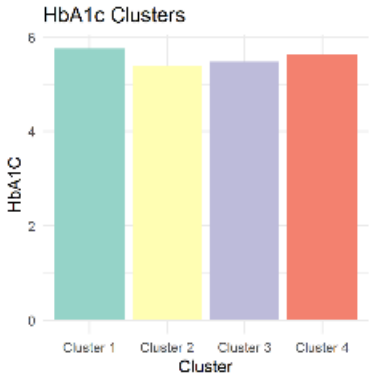

Fig.8.5

The graphs above revealed distinct subgroups within the dataset, offering valuable insights into health risks and opportunities for preventive care. High-risk subgroups were particularly evident in two clusters. In Cluster 1 (Older Group), we observed the highest prevalence of hypertension at 20.8%, heart disease at 14.4%, and hbA1c levels of 5.8, making this group the most at risk for diabetes and cardiovascular complications. This suggests that age is a significant factor in the development of chronic conditions, and targeted healthcare interventions are critical for this group. Similarly, Cluster 4 (Young Middle-Aged Group) showed a combination of obesity (BMI 30.14) and elevated HbA1c levels (5.64), indicating an emerging risk for diabetes and cardiovascular disease.

We also identified Unique Subgroups within the clusters. Cluster 2 (Young, Healthy Group) represented a younger population with a low prevalence of hypertension and heart disease, making them a key opportunity for preventive care. This group highlights the importance of maintaining healthy habits and monitoring health indicators early to avoid future complications. In contrast, Cluster 3 (Young-Middle-Aged Group) presented a unique balance: while it showed a low prevalence of heart disease, it also had a relatively high BMI (29.64). This elevated BMI

suggests a risk for future health complications, particularly if preventive measures are not implemented.

Classification Models

Now, we dive into the classification models, which were designed to answer the question: *Can we classify an individual as diabetic or non-diabetic based on their health metrics and demographic data?* To approach this, we first built a Null Model to serve as a baseline for comparison with more complex classification models. Once the Null Model was established, we compared it to various classification models, including Nearest Neighbor, Support Vector Machines (SVM), Decision Trees, Random Forest, and Logistic Regression.

Null Model

The Null Model assumes that all individuals belong to the majority class, which, in this case, is "non-diabetic." This baseline model allows us to assess the performance of other models and determine whether they provide meaningful improvements over random or majority-class predictions. The accuracy of the Null Model is 89.0% which is driven by the fact that the majority of the cases are No (as in Non-Diabetic). These results and code are shown below

```
> accuracy(pred, truth = diabetes, estimate = diabetes_null)
# A tibble: 1 × 3
  .metric  .estimator .estimate
  <chr>    <chr>          <dbl>
1 accuracy binary         0.890
> confusion_null <- pred |>
+   conf_mat(truth = diabetes, estimate = diabetes_null)
> confusion_null
          Truth
Prediction    No   Yes
       No  11421  1416
       Yes     0     0
```

Fig.9.1: Null model confusion matrix

```
#install.packages("kernlab")
library(kernlab)
mod_null <- svm_linear(mode = "classification") |>
  set_engine("kernlab") |>
  fit(diabetes ~ 1, data = trainingNULL)
```

Fig.9.2: Null Model

Nearest Neighbor

The accuracy of the model is 92.4%, indicating that 92.4% of the predictions made by the model are correct. This suggests that the model performs well in distinguishing between diabetic and non-diabetic individuals. The misclassification rate, which represents the proportion of incorrect predictions, is 7.6%, highlighting the percentage of cases where the model fails to classify correctly.

Below, in Fig. 10, we present the confusion matrix of the k-nearest Neighbors (kknn) model. The confusion matrix provides a detailed breakdown of the model's performance, showing the counts

of true positives, true negatives, false positives, and false negatives, which help evaluate the model's effectiveness in predicting diabetes.

```
> pred |>
+   conf_mat(diabetes, diabetes_knn)
          Truth
Prediction    No   Yes
       No  11181   717
       Yes   261   678
> pred |>
+   accuracy(diabetes, diabetes_knn)
# A tibble: 1 x 3
  .metric  .estimator .estimate
  <chr>    <chr>          <dbl>
1 accuracy binary         0.924
>
```

Logistic Regression

The model achieved an accuracy of 93.15%, meaning it correctly predicted diabetes status in 93.15% of cases. However, the misclassification rate is 6.85%, representing the proportion of cases where the model made incorrect predictions. The confusion matrix below in Fig.11 reveals an imbalance in prediction errors, with significantly more false negatives (770) than false positives (109). This imbalance could pose a problem, especially if the cost of missing a diabetes diagnosis (false negatives) is high, as it could lead to delayed treatment or increased health risks. Additionally, the model appears to favor predicting “No diabetes”, potentially due to the class imbalance in the dataset or the choice of the decision threshold.

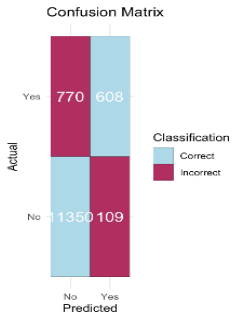

Fig.11: Logistic Regression confusion matrix

Support Vector Machines

The svm model performs well overall, achieving an accuracy of 93.2%. The misclassification rate is 6.8%, indicating that the model makes incorrect predictions 6.8% of the time. The SVM graph in Fig. 12 visually represents the classification regions, where the "Yes" (cream) region corresponds to diabetic predictions and the "No" (red) region corresponds to non-diabetic predictions. The black data points represent non-diabetic individuals, while the red data points represent diabetic individuals.

Misclassified data points are those that fall into the opposite region. For instance, they can be seen as red points located in the "No" region, representing cases where the model incorrectly predicted non-diabetic status. Similarly, misclassified non-diabetic cases would appear in the "Yes" region. Individuals with HbA1c levels greater than 7 are considered diabetic, which highlights the importance of correctly classifying such cases.

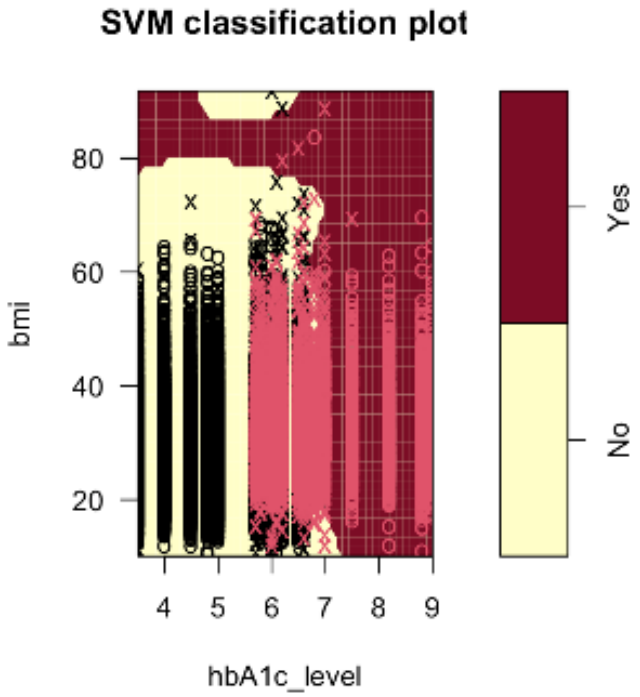

Fig.12: SVM graph

Decision Trees

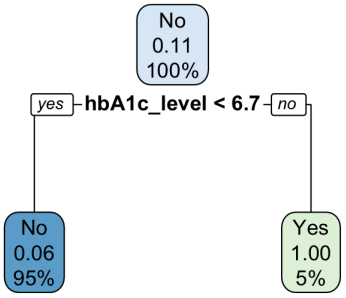

Fig.13.1: Decision Tree Graph

```
> table(guesses,testingDT$diabetes)

guesses    No   Yes
    No  11392   770
    Yes     0   675
> prop.table(table(guesses,testingDT$diabetes))

guesses         No        Yes
    No  0.88743476 0.05998286
    Yes 0.00000000 0.05258238
> conf_matrix3=table(guesses,testingDT$diabetes)
> accuracy <- sum(diag(conf_matrix3)) / sum(conf_matrix3)
> accuracy
[1] 0.9400171
```

Fig13.2: DTmodel Confusion Matrix

The visualization in Fig.13.1 represents the decision-making process of the Decision Tree model, showcasing how the data is split based on the key predictor ***hbA1c_level***. The Root Node of the model evaluates the condition ***hbA1c_level*** < 6.7. If the condition is true (Left Child), meaning ***hbA1c_level*** < 6.7, the decision leads to a "No" classification (non-diabetic) with a probability of 0.06, indicating a very low likelihood of diabetes. This branch represents 95% of the data. If the condition is false (Right Child), meaning ***hbA1c_level*** >= 6.7, the decision leads to a "Yes" classification (diabetic) with a probability of 1.00, showing full certainty of diabetes, and this branch represents 5% of the data. In conclusion, ***hbA1c_level*** is identified as the key predictor, demonstrating a strong association with diabetes. The model achieves an accuracy of 94.0% (shown in Fig.13.2), correctly predicting diabetes status for the majority of cases, while the misclassification rate is 6%.

Random Forest

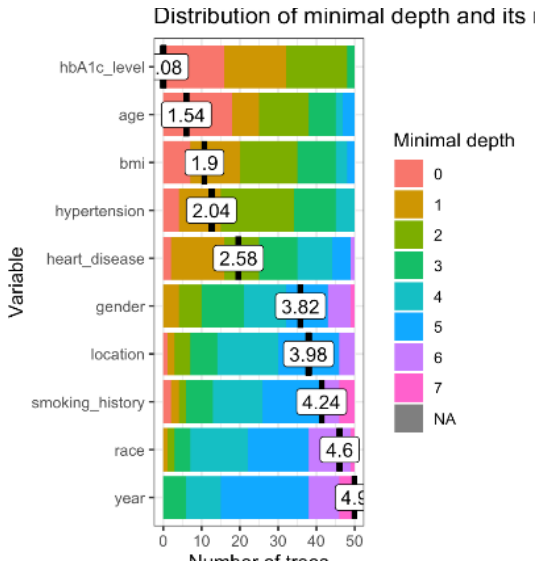

Fig.14.1 Distribution of minimal depth plot

The plot in Fig. 14.1 illustrates how deep each variable typically appears in the decision trees of the random forest model. The colors represent the minimal depth of each variable, and the numbers inside the boxes indicate the average minimal depth. Variables with smaller average depths are more significant for decision-making because they are used for splits earlier in the trees. In contrast, variables with larger average minimal depths are less influential. For example, ***hbA1c_level*** has the smallest average minimal depth (0.08), indicating that it is the most important predictor in the model. On the other hand, year has the largest minimal depth (4.9), making it the least significant variable. This plot highlights the relative importance of the predictors, with ***hbA1c_level*** playing a key role in driving the model's decisions. The model achieved an accuracy of 93.79%, meaning it correctly classified approximately 93.79% of the observations. This includes 56,703 True Negatives (correctly predicted as non-diabetic) and 3,496 True Positives (correctly predicted as diabetic), as shown in Fig. 14.2.

```
> RF1=randomForest(diabetes~.,data=randomF_dataset,ntree=50)
> RF1

Call:
 randomForest(formula = diabetes ~ ., data = randomF_dataset,      ntree = 50)
               Type of random forest: classification
                     Number of trees: 50
No. of variables tried at each split: 3

        OOB estimate of  error rate: 6.21%
Confusion matrix:
      0    1 class.error
0 56703  435 0.007613147
1  3550 3496 0.503831961
```

Fig.14.2: Random Forest Model Confusion matrix

Conclusion

Classification model	Accuracy
Null model	89.0%
Nearest Neighbor	92.4%
Logistic Regression	93.15%
SVM	93.2%
Decison Trees	94.0%
Random Forest	93.79%

In conclusion, all the models outperformed the Null Model, demonstrating their ability to make meaningful predictions. Among them, the Decision Tree Model stands out as the best-performing model, achieving the highest accuracy of 94%. This indicates that the Decision Tree Model provides a strong fit for the data, effectively distinguishing between diabetic and non-diabetic individuals. The analysis from our project demonstrated that health indicators such as bmi, age, hbA1c levels, hypertension, and heart disease are strongly associated, with diabetes and provide valuable insights for future research and healthcare strategies..

Potential Future Work

Our work did not incorporate the Time-Series techniques we learned in class. These techniques could be useful for addressing questions such as analyzing fluctuations in a patient's blood glucose levels over time or predicting hospitalizations due to diabetes-related complications. To apply time-series methods, data such as a patient's periodic blood glucose measurements and history of hospital visits would be required. That said, the machine-learning techniques that we applied in this project have been sufficient to effectively answer the research questions we set out to address.