TAYLOR'S UNIVERSITY
Wisdom · Integrity · Excellence

SCHOOL OF COMPUTER SCIENCE

MID-EVALUATION (Weightage 10%)
JANUARY 2023 SEMESTER

MODULE NAME	**: DATA MINING**
MODULE CODE	**: ITS61504**
DUE DATE	**: 16.06.2023- 25.06.2023, 8.00PM (MYT)**
PLATFORM	**: MyTIMES**

This paper consists of THREE (3) pages, inclusive of this page.

STUDENT DECLARATION

1. I confirm that I am aware of the University's Regulation Governing Cheating in a University Test and Assignment and of the guidance issued by the School of Computing and IT concerning plagiarism and proper academic practice, and that the assessed work now submitted is in accordance with this regulation and guidance.

2. I understand that, unless already agreed with the School of Computing and IT, assessed work may not be submitted that has previously been submitted, either in whole or in part, at this or any other institution.

3. I recognise that should evidence emerge that my work fails to comply with either of the above declarations, then I may be liable to proceedings under Regulation

No	Student Name	Student ID	Date	Signature	Score
1	Low Kay Jing	0342375	27 Jun 2023	LKJ	
2	Ng Kai Hong	0344105	27 Jun 2023	NKH	

Task 3: Marking Rubrics MLO3 (lecturer only)					
Question	**Weight**	**Excellent (90-100)**	**Good (75-89)**	**Average (40-74)**	**Poor (0-39)**
Q1	/10	Understanding and evaluating successfully, different data mining techniques as per the real case scenario requirement, to complete the required analysis.	Understanding and evaluating moderately, different data mining techniques as per the real case scenario requirement, to complete the required analysis.	Understanding and evaluating majorly, different data mining techniques as per the real case scenario requirement, to complete the required analysis.	Understanding and evaluating a few, different data mining techniques as per the real case scenario requirement, to complete the required analysis.
Q2	/10				
Q3	/10				
Q4	/10				

Declaration

- *I pledge to be respectful and supportive of my team member.*
- *I pledge to abide by the deadline set by my lecturer and team member.*

No	Student Name (Student ID)	Work breakdown	Signature
1	Low Kay Jing 0342375	Pre-process dataset & Feature selection and elimination	*LKJ*
2	Ng Kai Hong 0344105	Select prediction model ,determine disease risk factors & Performance metrics and significant risk factors	*NKH*

1.0 Pre-proccessing technique

1.1 Data Cleaning

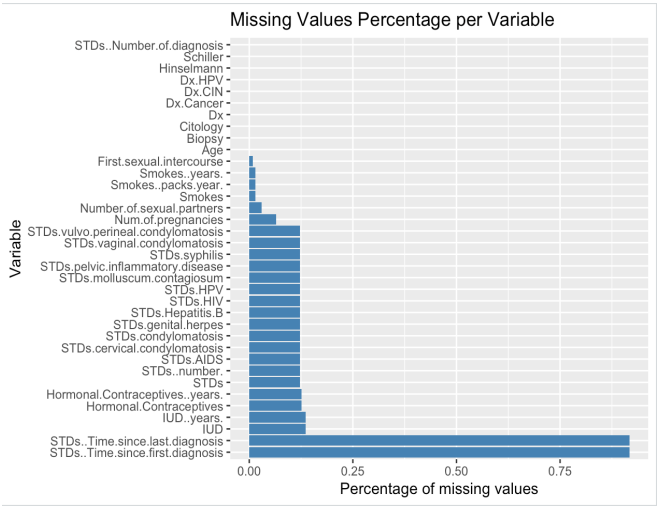
To preprocess the dataset, we took several steps. Firstly, we visualized the distribution of missing values in each attribute using a bar plot and discovered that "STDs..Time.since.first.diagnosis" and "STDs..Time.since.last.diagnosis" had missing values exceeding 70%. Consequently, we decided to eliminate these attributes. The code then calculated the percentage of missing values for each attribute in the cervical_data dataset and created a bar plot to visualize the missing values per variable. Attributes with a missing value percentage higher than 70% were dropped from the dataset.

```
smokes_year_median<- median(cervical_data$Smokes..years., na.rm = TRUE)
# Replace NA values with the median
cervical_data$Smokes..years.[is.na(cervical_data$Smokes..years.)] <- smokes_year_median
```

In the next phase, the provided code focused on replacing missing values with appropriate measures. The above median() function is used to calculated median for those numerical attribute and replace with their missing value.

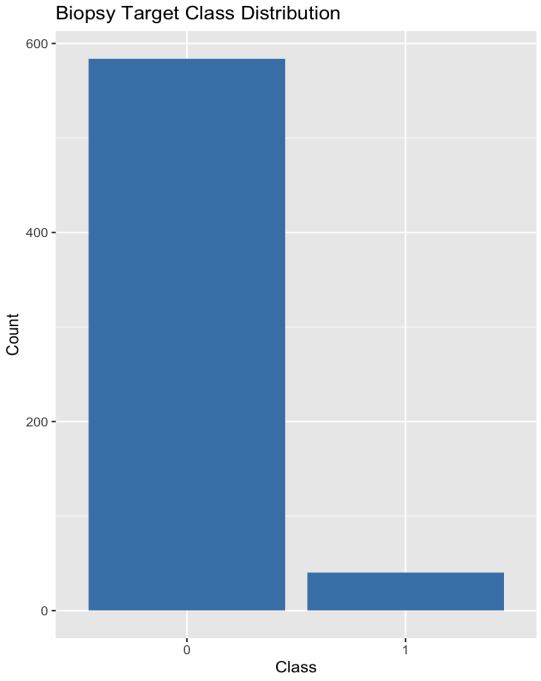
```
cervical_data$First.sexual.intercourse[is.na(cervical_data$First.sexual.intercourse)] <-
  names(which.max(table(cervical_data$First.sexual.intercourse)))
```

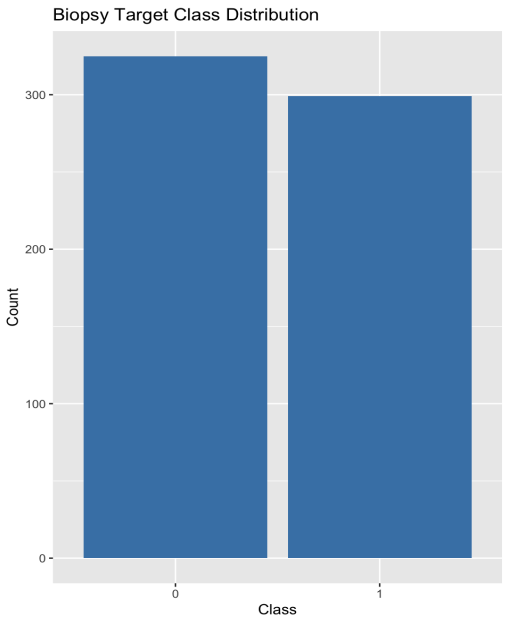
Additionally, those categorical attributes is used above function to replace missing data with rach most frequent value . Finally, three target variables (Citology, Schiller, Hinselmann) were dropped from the dataset, and the remaining attributes were converted to numeric format. The missing values in numerical columns, denoted as y(s)_mis, are substituted with the mean of observed values, y(s)_obs. Similarly, for categorical columns, the missing values y(s)_mis are replaced with the mode of observed values, y(s)_obs, which represents the most frequently occurring value. (Suh & Song, 2023) These comprehensive preprocessing steps ensure that the dataset is ready for further analysis and modeling.




```
> colSums(is.na(cervical_data))
                                  Age                 Number.of.sexual.partners 
                                    0                                         0 
                                  First.sexual.intercourse                                  Num.of.pregnancies 
                                    0                                         0 
                                  Smokes                                  Smokes..years. 
                                    0                                         0 
                                  Smokes..packs.year.                 Hormonal.Contraceptives 
                                    0                                         0 
                                  Hormonal.Contraceptives..years.                                  IUD 
                                    0                                         0 
                                  IUD..years.                                  STDs 
                                    0                                         0 
                                  STDs..number.                 STDs.condylomatosis 
                                    0                                         0 
                                  STDs.cervical.condylomatosis                                  STDs.vaginal.condylomatosis 
                                    0                                         0 
                                  STDs.vulvo.perineal.condylomatosis                                  STDs.syphilis 
                                    0                                         0 
                                  STDs.pelvic.inflammatory.disease                                  STDs.genital.herpes 
                                    0                                         0 
                                  STDs.molluscum.contagiosum                                  STDs.AIDS 
                                    0                                         0 
                                  STDs.HIV                                  STDs.Hepatitis.B 
                                    0                                         0 
                                  STDs.HPV                                  STDs..Number.of.diagnosis 
                                    0                                         0 
                                  Dx.Cancer                                  Dx.CIN 
                                    0                                         0 
                                  Dx.HPV                                  Dx 
                                    0                                         0 
                                  Hinselmann                                  Schiller 
                                    0                                         0 
                                  Citology                                  Biopsy 
                                    0                                         0 
> 
```

1.2 Resampling





(Figure 1)
Before resampling

(Figure 2)
After resampling

Above the figure 1, we can see that the Biopsy target variable is an imbalanced class. A key challenge with datasets that feature a heavily imbalanced target is the higher potential consequences of misclassification errors. (Ruscello, 2020) However, i have utilized the ROSE() function to perform resampling on the 'Biopsy' variable within the training set, while maintaining the original configuration of the test set after data splitting. The choice measures for evaluating prediction models in imbalanced data scenario

2.0 Explanatory Data Analysis

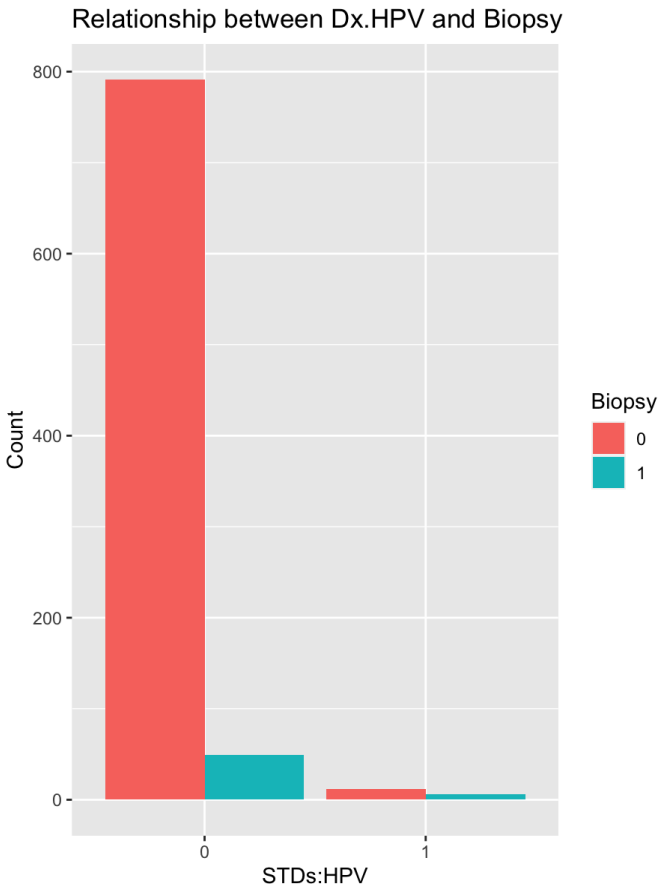


Figure 3

The bar plot visually represents the association between Dx.HPV and Biopsy outcomes in the cervical dataset. The red bar represents Biopsy 0, indicating no biopsy performed, with 800 instances where patients reported no occurrences of Dx.HPV. The green bar represents Biopsy 1, indicating a positive biopsy result, with 50 instances where patients reported no occurrences of Dx.HPV. This plot effectively displays the distribution of Biopsy outcomes based on the presence or absence of Dx.HPV. (World Health Organization (WHO), 2022)

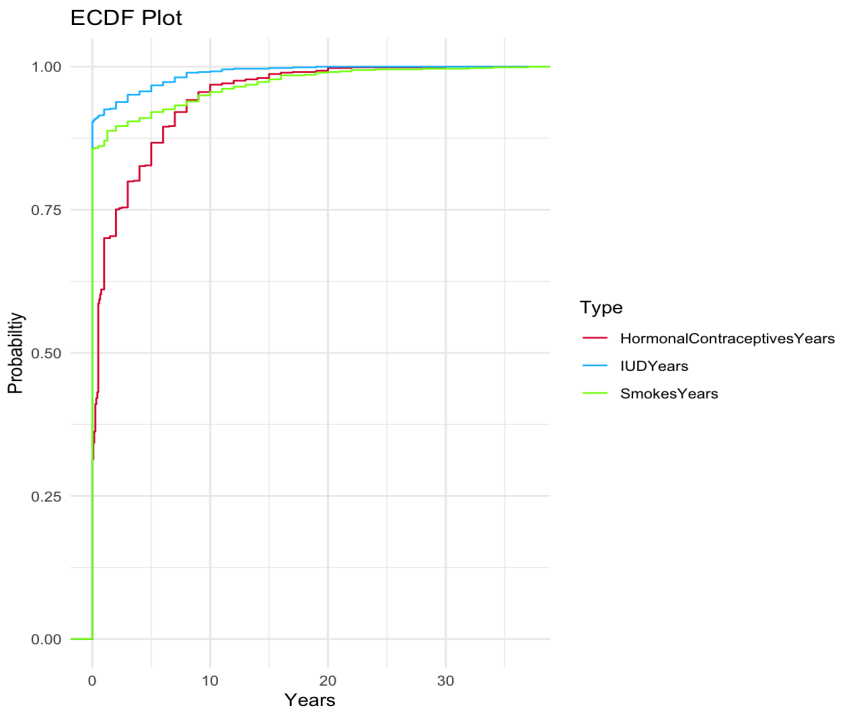


Figure 4

The ECDF plot provides valuable insights into the duration of behaviors and their association with patients. Analysis reveals that there is a high probability (approximately 95%) that patients have smoked, used IUDs, and Hormonal Contraceptives for less than 10 years. These findings highlight the prevalence of relatively shorter durations for these behaviors among the study participants (Allen, 2017).

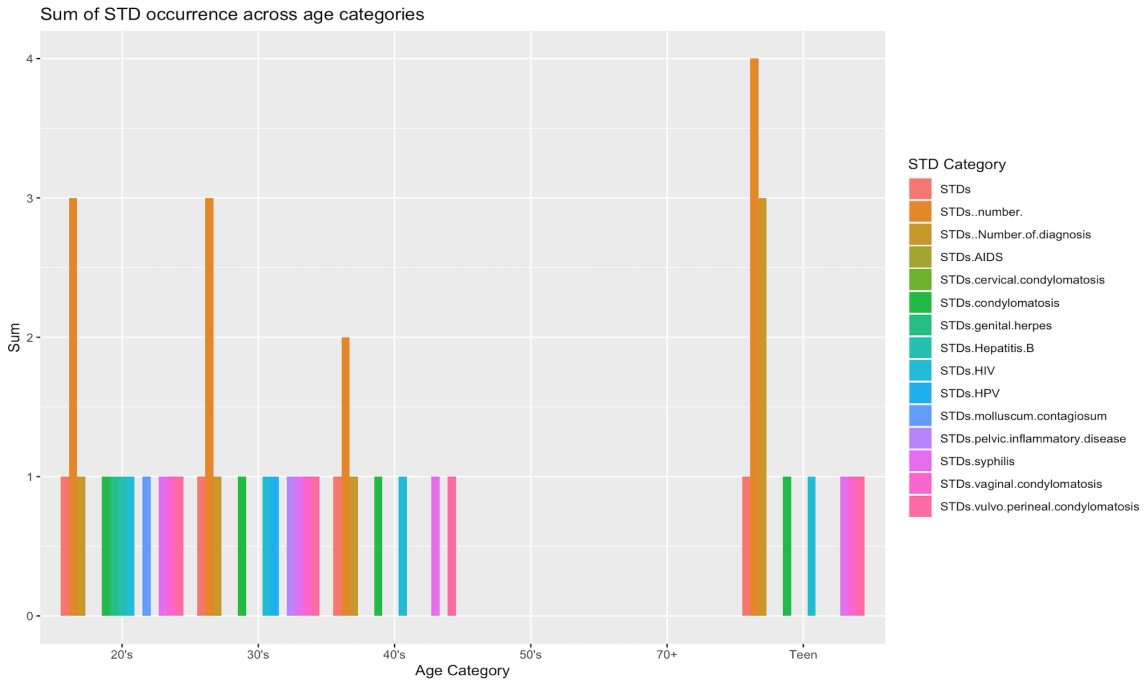


Figure 5

The figure 5 reveals consistent numbers of STD occurrences across different age ranges, except for STDs.number, where there are a total of 4 occurrences in the teenage age group (below 20). This aligns with the well-known fact that STD infections primarily affect individuals aged 15-24, accounting for approximately half of all cases. The high prevalence of STDs among young people can be attributed to various factors such as biological susceptibility, inadequate health screenings, limited communication with healthcare providers, restricted access to testing facilities, and lifestyle behaviors (Centers for Disease Control and Prevention, 2022).

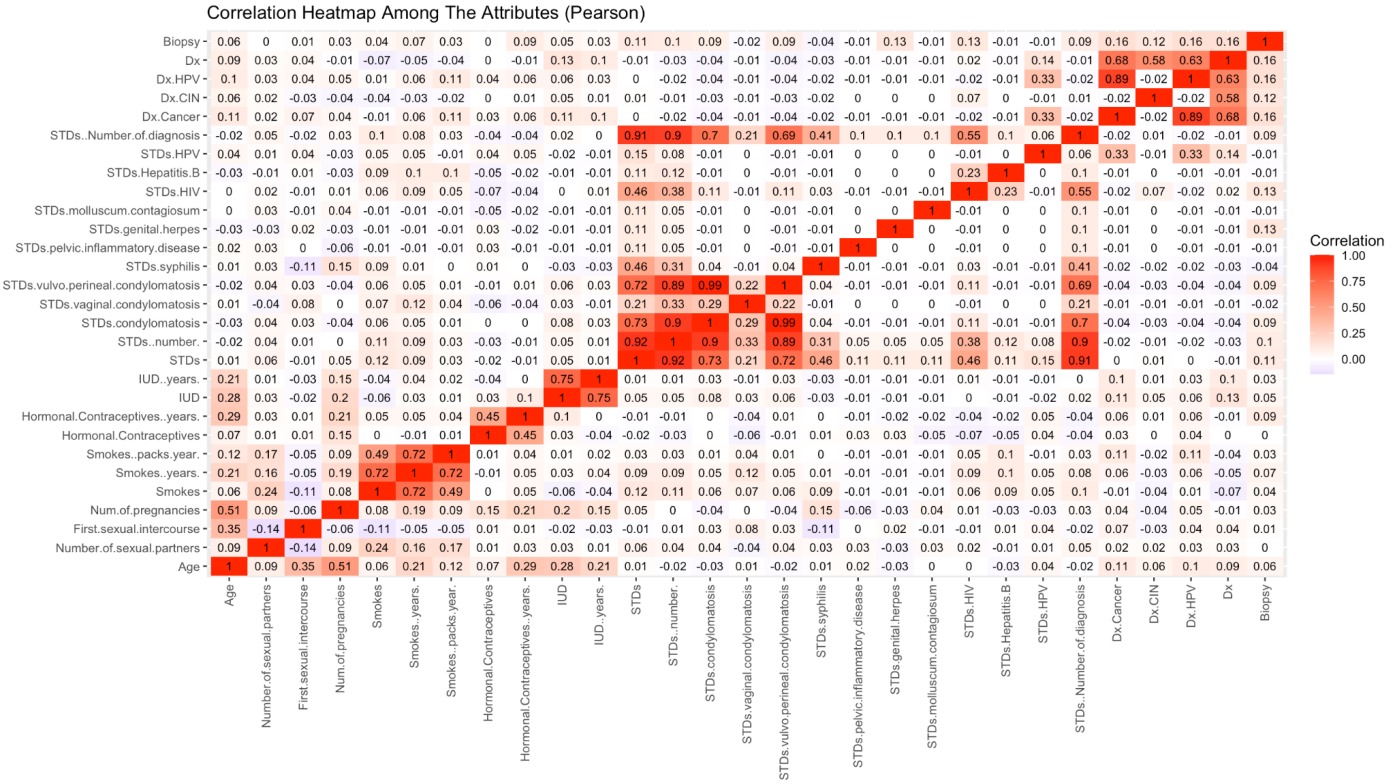


Figure 6

According to the above Pearson correlation heatmap, we can see that the target variable Biopsy has show low correlation to all other variable. In the other hand, there a few set of attribute has show high correlation to each other. Highly correlated variables due to their linear dependence can predict an outcome value almost identically. Simplifying the model by removing one of these variable before training doesn't significantly impact the model's performance This helps in streamlining the learning process and reducing model complexity. (Lannge, 2021).

Pearson Heatmap Insight :

Highly Correlated Set of attributes	Top three of attributes that positive correlated to the target attribute	Top three of attributes that negative correlated to the target attribute
Smokes, Smokes..pack.year, Smokes…years	Dx.Cancer Dx.CIN, Dx.HPV	STDs. syphilis STDs.vaginal.condylomatosis STDs.HPV
STDs.number, STDs.condylomatosis, STDs, STDs.vulvo-perinalcondylomatosis, STDs..Number.of.diagnosis		
Dx.Cancer, Dx.CIN, Dx.HPV,Dx		

Random Forest Feature Selection

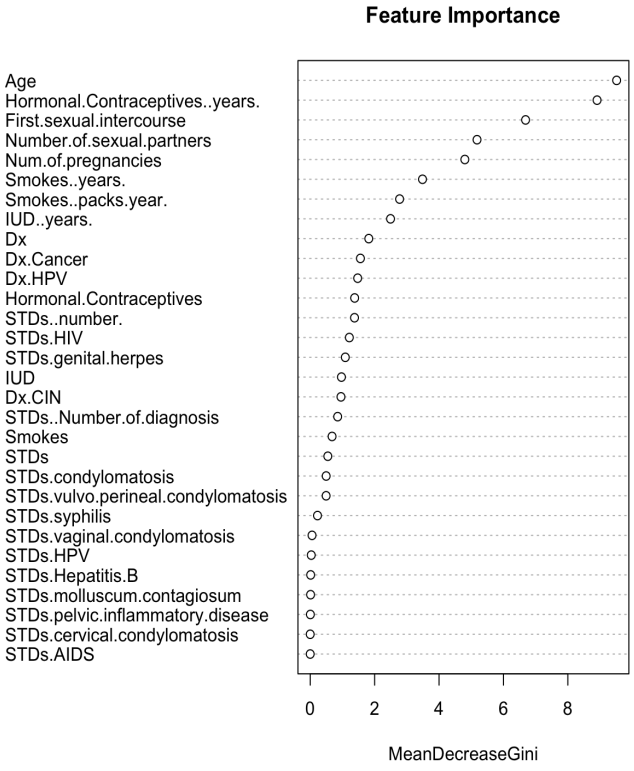


Figure 7

```r
# Train random forest model to find feature importance
random_forest_model <- randomForest(Biopsy ~ ., data = cervical_data, importance = TRUE)

# Get feature importance scores
feature_importance <- importance(random_forest_model)

# Create feature importance plot
varImpPlot(random_forest_model, type=2,main = "Feature Importance")
```

Figure 8

Above Figure 7 is the featueres important score of MeanDecreaseGini which trained by random forest model. It transforms certain variables, trains a Random Forest model to determine feature importance, and generates a plot displaying the importance of each feature. The top 10 features with the highest importance scores are selected, including the "Biopsy" variable. The resulting dataset, "cervical_data_important," contains only the selected variables. The Figure 7 represents the feature importance scores calculated using the MeanDecreaseGini metric, aiding in identifying the most significant features for predicting the "Biopsy" outcome in the cervical dataset.

Feature Selection Decision :

Following extensive research ,statistical analysis and brute force approach, we have finalized our selection of feature for creating a cervical cancer prediction model. The chosen attributes are **'Age'**, '**Smokes..packs.year.**', '**Hormonal.Contraceptives..years**.', '**Number.of.sexual.partners**', '**First.sexual.intercourse**', '**Dx.Cancer**', '**STDs..Number.of.diagnosis**', **'Dx.HPV'**, **'STDs.HIV'** .The target attribute for our classification model will be **'Biopsy'**.

3. Predictive method or classification model

Comparision of prediction model

Decision Tree Classification	Random Forest Classification	K-Nearest Neigbours Classification
Confusion Matrix and Statistics Reference Prediction 0 1 0 210 22 1 23 3 Accuracy : 0.8256 95% CI : (0.7737, 0.8698) No Information Rate : 0.9031 P-Value [Acc > NIR] : 1 Kappa : 0.0209 Mcnemar's Test P-Value : 1 Sensitivity : 0.9013 Specificity : 0.1200 Pos Pred Value : 0.9052 Neg Pred Value : 0.1154 Precision : 0.9052 Recall : 0.9013 F1 : 0.9032 Prevalence : 0.9031 Detection Rate : 0.8140 Detection Prevalence : 0.8992 Balanced Accuracy : 0.5106 'Positive' Class : 0	Confusion Matrix and Statistics Reference Prediction 0 1 0 206 15 1 27 10 Accuracy : 0.8372 95% CI : (0.7864, 0.8801) No Information Rate : 0.9031 P-Value [Acc > NIR] : 0.99967 Kappa : 0.234 Mcnemar's Test P-Value : 0.08963 Sensitivity : 0.8841 Specificity : 0.4000 Pos Pred Value : 0.9321 Neg Pred Value : 0.2703 Precision : 0.9321 Recall : 0.8841 F1 : 0.9075 Prevalence : 0.9031 Detection Rate : 0.7984 Detection Prevalence : 0.8566 Balanced Accuracy : 0.6421 'Positive' Class : 0	Confusion Matrix and Statistics Reference Prediction 0 1 0 187 11 1 46 14 Accuracy : 0.7791 95% CI : (0.7234, 0.8282) No Information Rate : 0.9031 P-Value [Acc > NIR] : 1 Kappa : 0.2231 Mcnemar's Test P-Value : 6.687e-06 Sensitivity : 0.8026 Specificity : 0.5600 Pos Pred Value : 0.9444 Neg Pred Value : 0.2333 Precision : 0.9444 Recall : 0.8026 F1 : 0.8677 Prevalence : 0.9031 Detection Rate : 0.7248 Detection Prevalence : 0.7674 Balanced Accuracy : 0.6813 'Positive' Class : 0

Figure 11

Above Figure 11 have shown the Confusion Matrix and Statistics with three type classification prediction model. There are different metrics to evaluate the performance of prediction model since we are using the imbalanced dataset. Precision is a valuable metric in imbalanced datasets as it's not swayed by True Negatives. However, precision and recall can be imbalanced - enhancing True Positive might inflate False Positives. To address this, the F-score is used since it equally weight precision and recall. offering a balanced view less prone to bias towards either class.(Canuma, 2023). Based on these assessments, the Random Forest algortihmn emerges as the preferred choice due to its superior F1 score in comparison to other models. Thus, it's chosen for the prediction model to determine attribute that is most reason of the cervical cancer.

Random Forest Classification Prediction Model

```
rf_ctrl <- trainControl(method="cv", number=10)
```

We had sets up the resampling method with 10-fold cross validation for model training by using trainControl() function

```
grid <- expand.grid(mtry = seq(1, 10, by = 1))
```

Define a grid of model tuning parameter using expand.grid() function to control how the Random Forest model is trained.

```
rf_fit <- train(Biopsy ~ .,
                data = balanced_train_data,
                method = "rf",
                metric = "Accuracy",
                trControl = rf_ctrl,
                tuneGrid = grid)
```

Train a Random Forest model on the 'balanced_train_data' to predict 'Biopsy' using the specified cross-validation scheme.

```
rf_predicted_labels <- predict(rf_fit, newdata = data_test)
rf_actual_labels <- data_test$Biopsy
confusionMatrix(rf_predicted_labels, rf_actual_labels,mode = "everything")
```

Uses the trained Random Forest model (rf_fit) to predict ‘Biopsy’ for the test data

Create a object to retrieve the actual ‘Biopsy’ values from the test data

Calculates the confusion matrix for the predicted labels versus the actual labels, providing metrics for evaluating the model’s predictive performance.

4. Performance Evaluation and Most Risk Factor Determination

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 206  15
         1  27  10

               Accuracy : 0.8372
                 95% CI : (0.7864, 0.8801)
    No Information Rate : 0.9031
    P-Value [Acc > NIR] : 0.99967

                  Kappa : 0.234

 Mcnemar's Test P-Value : 0.08963

            Sensitivity : 0.8841
            Specificity : 0.4000
         Pos Pred Value : 0.9321
         Neg Pred Value : 0.2703
              Precision : 0.9321
                 Recall : 0.8841
                     F1 : 0.9075
             Prevalence : 0.9031
         Detection Rate : 0.7984
   Detection Prevalence : 0.8566
      Balanced Accuracy : 0.6421

       'Positive' Class : 0
```

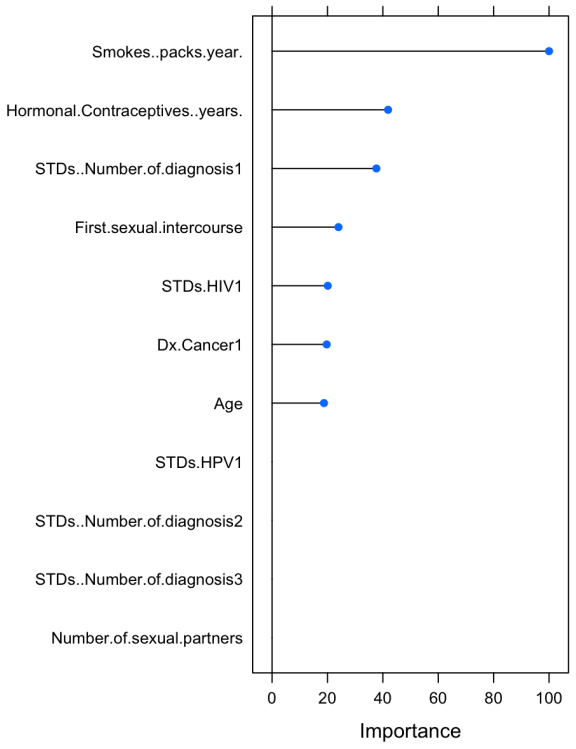


Figure 12

According to the confusion matrix, the random forest model we have built has yielded an accuracy of 83.72%. However, this metric can be misleading when applied to imbalanced datasets as it can be skewed by the larger class. Therefore, we need to consider additional mesures, such as

precision, recall, and F1-score for a more complete understanding of performance.

Our random forest model achieved a precision score of 0.9321, meaning that approximately 93.21% of the instances that were predicted as positive (in class 0) were correctly identified. However, our recall score, sitting at 0.8841 indicates a strong performance since it measures the proportion of actual positive cases that the model is correctly identified. We have achieved the 0.9075 of high F1 score which signifies a balanced model in terms of precision and recall. The Balanced Accuracy, which averages the proportion of correct predictions in each class, is 0.6421. This measure give a better indication of the model's performance on an imbalanced dataset.

The figure 12 displays ranking of feature importance derived from random forest prediction model. We obsser that the attribute 'Smokes..packs.years.' tops the chart with highest MeanDecrease of 62.63. This score signifies that 'Smokes..packs.years.' plays a significant role in the prediction of cervical cancer, according to our model trained on this imbalance datasets. Therefore, it's inferred that this attribute could be a key risk factor contributing to the independence of cervical cancer.

References

1. Allen, A. M. (2017, Nov 18). *Oral Contraceptives and Cigarette Smoking: A Review of the Literature and Future Directions*. NCBI. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6468133/
2. Brownlee, J. (2020, January 8). *Tour of Evaluation Metrics for Imbalanced Classification - MachineLearningMastery.com*. Machine Learning Mastery. https://machinelearningmastery.com/tour-of-evaluation-metrics-for-imbalanced-classification/
3. Cancer Research UK. (2021, May 26). *Risks and causes | Cervical cancer*. Cancer Research UK. https://www.cancerresearchuk.org/about-cancer/cervical-cancer/risks-causes
4. Canuma, P. (2023, April 25). *How to Deal With Imbalanced Classification and Regression Data*. Neptune.ai. https://neptune.ai/blog/how-to-deal-with-imbalanced-classification-and-regression-data
5. Centers for Disease Control and Prevention. (2022, April 12). *Adolescents and STDs | Sexually Transmitted Diseases | CDC*. Centers for Disease Control and Prevention. https://www.cdc.gov/std/life-stages-populations/stdfact-teens.htm
6. Lannge, E. J. (2021). *Does removal of correlated variables affect the classification accuracy of machine learning algorithms?* DiVA portal. https://www.diva-portal.org/smash/get/diva2:1632660/FULLTEXT01.pdf
7. Ruscello, J. (2020, August 22). *Testing recommendations for binary classification with an imbalanced target variable*. Medium. https://medium.com/@mmalinda/testing-recommendations-for-binary-classification-with-an-imbalanced-target-variable-ff8b120ea8c9
8. Suh, H., & Song, J. (2023, May 30). *A comparison of imputation methods using machine learning models*. Communications for Statistical Applications and Methods. Retrieved June 27, 2023, from http://www.csam.or.kr/journal/view.html?uid=2043&pn=lastest&vmd=Full
9. World Health Organization (WHO). (2022, February 22). *Cervical cancer*. World Health Organization (WHO). https://www.who.int/news-room/fact-sheets/detail/cervical-cancer