Machine Learning Project Report

Ayotunde Aribo 2022-1012, Evans Onorieru 2022-06572,Henry Tirla 2022-1016 and Mahmoud Ahmed 2022-1397

December 2022

1 Introduction

This project is a classification problem, as the goal is to predict a categorical outcome (whether or not a patient will suffer from the Smith disease) based on a set of input features.In this project, we set out to build a predictive model to answer the question of who is more likely to suffer from the Smith parasite. To do this, we explored the data on patients infected with the parasite. A fundamental interest to us was sociodemographic, health, and behavioral information. We cleaned and preprocessed the data to prepare it for modeling and then explored and visualized the data better to understand the patterns and relationships present within it. We then built and evaluated a predictive model using machine learning techniques,fine-tuned and optimized it to improve its performance. Ultimately, we were able to build a model that could accurately predict which patients were more likely to suffer from the disease. In this report, we will detail the steps we took to complete this project and discuss the results of our analysis.

2 Exploration

During the data exploration phase, our goal was to understand better the patterns and relationships present within the data on patients infected with the Smith parasite. To do this, we took the following steps:

2.1 Examining the Dataset

After merging our demo, habits and health dataset. We used the **head()** method to see the first five rows of our dataset and used the **info()** method to check the data types of each variable.In this dataset, there are nine variables with the data type int64, which indicates that they are numerical variables that can hold integer values. There are also ten variables with the data type object, which indicates that they are categorical variables that can hold string values.

2.1.1 Statistical Summary

These summary statistics give us a general understanding of the distribution and characteristics of the data in each column.

- `PatientID` column has a range of values from 1001 to 2024.
- `Birth_Year` column has a range of values from 1855 to 1993. The mean value of 1966 and standard deviation of 15 suggest that the majority of the patients were born around 1966, with a relatively even distribution of values around that mean.
- `Region` column has 10 unique values representing different regions. The most common region is East Midlands, with 154 patients.
- `Education` column has 6 unique values representing different education levels. The most common education level is University Complete (3 or more years), with 239 patients.
- `Disease` column represents whether or not each patient has the Smith Disease, and has a range of values from 0 (no disease) to 1 (has disease). The mean value of 0.514 and standard deviation of 0.500 suggest that about half of the patients have the disease, with a relatively even distribution of values around that mean.
- `Height` column has a range of values from 151 to 180. The mean value of 167 and standard deviation of 8 suggest that the majority of the patients are around 167cm tall, with a relatively even distribution of values around that mean.
- `Weight` column has a range of values from 40 to 97. The mean value of 67 and standard deviation of 12 suggest that the majority of the patients weigh around 67kg, with a relatively even distribution of values around that mean.
- `High_Cholesterol` column represents the high cholesterol level of each patient in milligrams per deciliter, and has a range of values from 130 to 568. The mean value of 249 and standard deviation of 52 suggest that the majority of the patients have a high cholesterol level around 249mg/dL, with a relatively even distribution of values around that mean.
- `Blood_Pressure` column represents the blood pressure of each patient in millimeters of mercury, and has a range of values from 94 to 200. The mean value of 131 and standard deviation of 17 suggest that the majority of the patients have a blood pressure around 131mmHg, with a relatively even distribution of values around that mean.
- `Physical_Health` column has a range of values from 0 to 30. The mean value of 5 and standard deviation of 5 suggest that the majority of the patients have a physical health score around 5, with a relatively even distribution of values around that mean.

- Checkup has 4 unique values representing different frequencies. The most common frequency is More than 3 years.

2.1.2 Checking Missing Values & Duplicates

We checked for any missing or invalid values or duplicates in the data. We used isnull & isna methods detects nullable and missing values. duplicated method returns boolean Series denoting duplicate rows. We observed there are 13 missing values in Education column and no duplicates.

2.1.3 Inspecting Categorical Values

The categorical columns in the dataset include Name, Region, Education,Checkup, Diabetes, SmokingHabbit, DrinkingHabit,Exercise, FruitHabit,WaterHabit Each of these columns has a relatively small number of unique values compared to the total number of rows, indicating that the data in these columns is not highly varied. The most frequent value in each of these columns appears a relatively large number of times compared to the other unique values, suggesting that these values are disproportionately represented in our dataset.Overall, these give a general overview of the distribution and frequency of values in the categorical columns of our dataset.

2.1.4 Detecting Outliers

Here outliers are data points that lie significantly outside the range of the majority of our data. We defined two functions to use different methods for detecting outlier empirical rule / 68-95-99.7 rule & IQR Rule

- empirical rule / 68-95-99.7 rule: The function returns a summary of the number of outliers for each column and a list of the outlier values themselves. The results of the empirical rule show that there are a certain number of values in each column that fall outside of three standard deviations from the mean. For example, in the "Birth_Year" column, there are 12 values that are considered outliers. This means that these values are significantly different from the majority of the values in the column and may warrant further investigation. Similarly, there are 9 values in the "High Cholesterol" column, 6 values in the "Blood_Pressure" column, 4 values in the"Mental_Health" column, and 8 values in the "Physical_Health" column that are considered outliers

- IQR Rule It appears that the two different outlier detection methods produced slightly different results, with this method identifying more outliers in some of the columns.

In conclusion while some values will be dealt with due to inaccuracy, some extreme values will be kept to further investigate their meaning.

2.1.5 Feature Metadata

After the initially analyzing for the columns, we can conclude that the following observations can be removed:

- `PatientID` Randomly generated field for unique identification purposes only.
- `Name` Patient names long text explanation for the loan that we won't need.

2.2 Data Visualization:

We created summary statistics and plots to visualize the distribution of different variables and identify any trends or correlations. This helped us gain insights into the patterns and relationships present within the data.

Based on the statistical summary and the visual data exploration, we can make the following insights:

- The data has relatively even distributions for most of the numeric columns, with means and medians that are close to each other. This suggests that the data is approximately normally distributed, or follows a Gaussian distribution.
- There are potential univariate outliers in the data, as indicated by the results of the empirical rule and IQR rule outlier detection methods. For example, there are 12 values in the `"Birth_Year"` column that are considered outliers, and 11 values in the `"High_Cholesterol"` column that are considered outliers. These unusual observations could potentially impact our analysis.
- There are also potential bivariate outliers in the data, as indicated by the correlation matrix. For example, there is a strong negative correlation between `"Weight"` and `"Physical_Health"` (-0.39), which could potentially impact our analysis.
- Also based on the correlation matrix, we can see that there is a moderate positive correlation between weight and height (0.51). There is also a moderate negative correlation between birth year and high cholesterol (-0.23). All other correlations are relatively weak, with values ranging from -0.12 to 0.30.

3 Data Preprocessing

In this stage, we prepare the data for analysis and modeling by cleaning and transforming it as needed.Our goal in this stage was to get the data into a form that is ready for analysis, while preserving as much relevant information as possible. The data was cleaned and transformed to prepare it for modeling. This included inputting missing values, removing outliers, correcting typos, converting certain categorical variables to boolean data types, creating new features, and encoding and normalizing the data. The data was also split into training, validation, and test sets using the `train_test_split` function. Standardization was also applied to the metric features using the StandardScaler class. These steps were important in ensuring that the data is in a suitable format for modeling and that the results of our modeling are reliable and meaningful.

4 Featured Selection

In this section, our goal was to identify which features in the dataset had a significant relationship with the target variable. To do this, the Chi-Square test of independence was used to test the relationship between the categorical variables and the target variable. Six variables were found to have a significant relationship with the target: Education, Checkup, Diabetes, DrinkingHabit, Exercise, and FruitHabit. These variables were then further analyzed using OneHotEncoding and the Bonferroni-adjusted method to determine the specific classes within each category that were responsible for the relationship with the target. In addition to the categorical variables, several numerical variables (Age, MentalHealth, and PhysicalHealth) were also selected for use in the model based on their correlation with the target. All of these selected features were then used to build and evaluate several machine learning models, including Logistic Regression, K-Nearest Neighbors, and Random Forest, to predict the likelihood of an individual having the disease.

5 Modeling

In this section , the selected features were used to train and evaluate various machine learning models to predict the target variable. The models that were trained and evaluated included Logistic Regression, K-Nearest Neighbors, Decision Trees, Random Forests, and XGBoost.The models were also evaluated using cross-validation to get an estimate of the model's performance on unseen data.

The performance of various classification models was evaluated on the training and validation datasets. The classification models used included DummyClassifier, GaussianNB, MultinomialNB, DecisionTreeClassifier, LogisticRegression, and XGBClassifier.To evaluate the performance of the models, the following metrics were used: accuracy, precision, recall, and F1 score.

6 Performance Assessment

A baseline prediction algorithm was also run to compare the models' performance.Based on these evaluation metrics, the best performing model was the DecisionTreeClassifier with class weight balanced and maximum depth of 5.

This model achieved the highest.`recall_macro` score of 0.843, as well as a relatively high accuracy score of 0.844 and average precision score of 0.890. It also had the highest `roc_auc` score of 0.901, indicating its ability to distinguish between positive and negative classes.To further improve the performance of all models, we used GridSearchCV to tune the parameters of each model.

- **KNN:** The model appears to be slightly better at predicting the positive class than the negative class , as indicated by the higher recall and f1-score scores for class 1 on both sets. However, the precision scores for both classes are similar, indicating that the model is making a similar number of false positive and false negative predictions.

- **Logistic Regression** Based on our results the logistic regression model performed well on the training and validation datasets, with an accuracy of around 84% on the training set and 86% on the validation set. The F1 score, and recall, were also quite high, with a value of around 87% on the validation set. The model also had a high AUC-ROC score, indicating that it was able to distinguish between positive and negative cases well.

  The model also performed well in cross-validation, with an average F1 score of around 90%. This suggests that the model is likely to generalize well to new data.

- **SVM:** Based on our results, the SVM model performed very well on both the training and validation datasets, with an accuracy of around 89% on the training set and 87% on the validation set. The F1 score, and recall, were also relatively high, with a value of around 88% on the validation set. The model also performed well in cross-validation, with an average F1 score of around 92%. This suggests that the model is likely to generalize well to new data.

  We also noted the model's performance improved significantly after training, with the F1 score increasing from 87.65% to 98.37%. This suggests that the model was able to learn effectively from the training data and improve its predictions.

- **Decision Trees:** Based on our results, the decision tree model performed very well on both the training and validation datasets, with an accuracy of around 100% on the training set and 99% on the validation set. The F1 score, and recall, were also relatively high, with a value of around 99% on the validation set.

  The model performed well in cross-validation, with an average F1 score of around 96%. This suggests that the model is likely to generalize well to new data.

  The model's performance improved slightly after training, with the F1 score increasing from 98.78% to 97.96%. This suggests that the model was able to learn effectively from the training data and improve its predictions.

- **Random Forest:** Based on our results the random forest model performed very well on both the training and validation datasets, with an accuracy of around 99% on the training set and 97% on the validation set. The F1 score, and recall, was also quite high, with a value of around 98% on the validation set.

  The model also performed well in cross-validation, with an average F1 score of around 96%. This suggests that the model is likely to generalize well to new data.

  The model's performance improved significantly after training, with the F1 score increasing from 97.54% to 98.79%. This suggests that the model was able to learn effectively from the training data and improve its predictions.

- **XGBoost:** Based on our results the XGBoost model performed very well on both the training and validation datasets, with an accuracy of around 100

  It also performed well in cross-validation, with an average F1 score of around 97%. This suggests that the model is likely to generalize well to new data.

  The model's performance improved slightly after training, with the F1 score increasing from 99.60% to 98.78%.

In summary, the Decision Trees, XGBoost, and Random Forest models all had high accuracy and precision on the validation set. The Support Vector Machines and Logistic Regression models also had high accuracy and precision but performed slightly worse than the top three models.

Here is a summary of the performance of each model based on accuracy and precision:

1. Decision Trees: High accuracy (99.2% on the validation set) and high precision (99.2% on the validation set)

2. XGBoost: High accuracy (99.6% on the validation set) and high precision (99.2% on the validation set)

3. Random Forest: High accuracy (97.9% on the validation set) and high precision (98.4% on the validation set)

4. Support Vector Machines: High accuracy (87.4% on the validation set) and high precision (86.2% on the validation set)

5. Logistic Regression: High accuracy (86.5% on the validation set) and high precision (85.3% on the validation set)

Overall, all of the models performed well on the validation set, with the Decision Trees, XGBoost, and Random Forest models having the highest accuracy and precision.

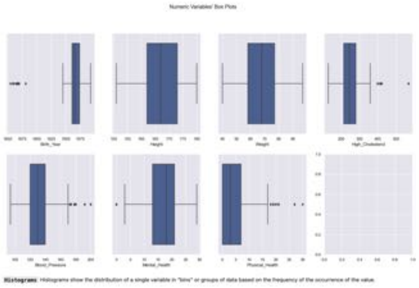
7 References

[1] Alex Galea (2020). "The Applied Datascience Workshop." Packt.Get started with the applications of data science and techniques to explore and assess data effectively.

[2] Geron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow.*This book provides practical, hands-on guidance on using popular machine learning libraries in Python, including scikit-learn, Keras, and TensorFlow.

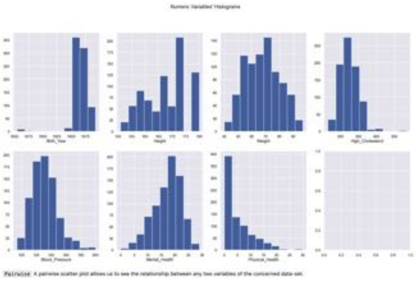
[3] Alpaydin, E. (2010). "Introduction to Machine Learning". This book provides a comprehensive introduction to machine learning, including the basics of supervised and unsupervised learning, decision trees, and ensemble methods.

8 Appendix



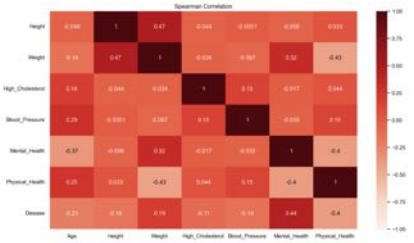
Histograms Histograms show the distribution of a single variable in "bins" or groups of data based on the frequency of the occurrence of the value.

(a) Histograms distribution of a single variable based on the frequency of the occurrence of the value

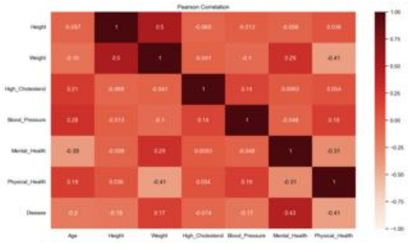


Pairwise A pairwise scatter plot allows us to see the relationship between any two variables of the concerned data-set.

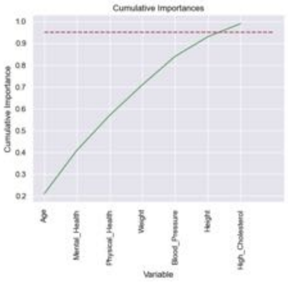
(b) A pairwise scatter plot to see the relationship between two variables



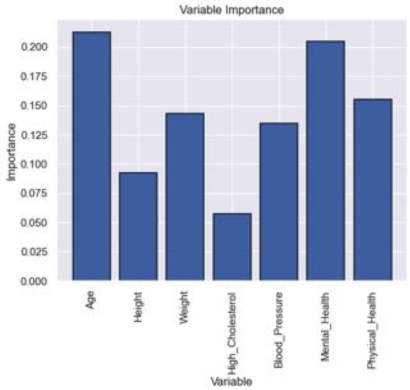
(a) Spearman correlation matrix



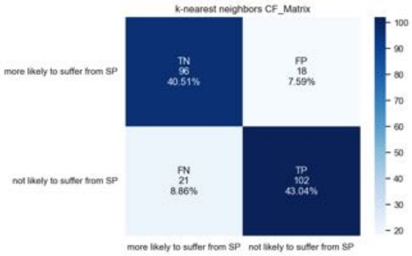
(b) Pearson correlation matrix



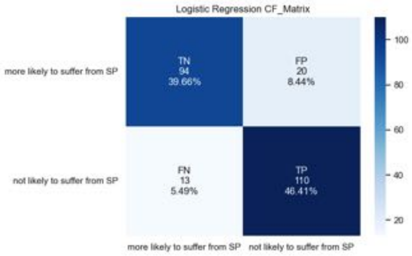
(a) Features cumulative importance



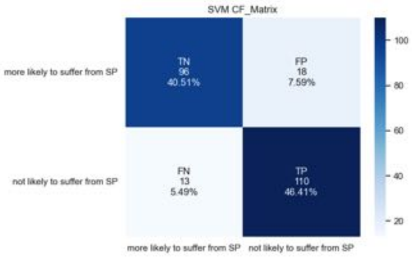
(b) Variable Importance



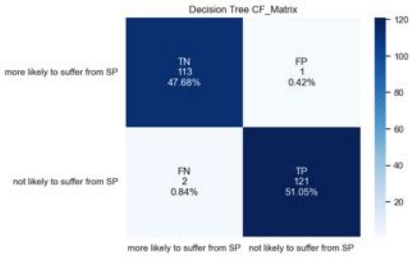
(a) KNN CF Matrix



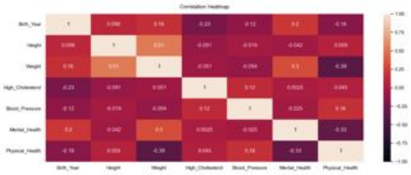
(b) Logistic Regression CFMatrix



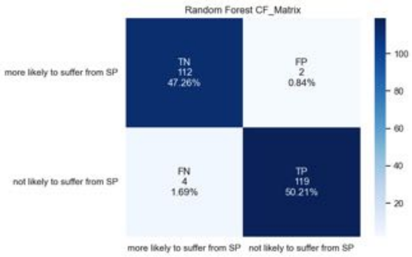
(a) SVM CF Matrix



(b) Decision Trees CFMatrix



(a) Correlation Heatmap



(b) Random Forest CF Matrix