Decision Tree Vs. Artificial Neural Network Classification in Predicting Diabetes

Octavia Istocescu (Student ID: 4328030) and Celeste Webb (Student ID: 20373245)

Abstract—**Diabetes mellitus affects millions of people, the conditions and its complications cause significant morbidity and mortality, which can be reduced by earlier identification. This study uses the Pima Indian Diabetes Dataset, which comprises a number of variables related to the development of diabetes. The main aim of the study was to predict the presence of diabetes. We used several separate methods of pre-processing and two different classifiers which were trained and tested in the R programming language. The best accuracy achieved by Decision Tree Classification was 77.97%, whereas Artificial Neural Network achieved 83.06%. It can be concluded that DT works well with feature reduction and complete outlier removal, whereas ANN works best with minimal pre-processing.**

Keywords—pima, diabetes, data analysis, classification

I. Introduction

Diabetes Mellitus is a serious, long-term medical condition that has a significant impact on the lives of those who develop it. Diabetes is caused by insulin deficiency or insulin resistance (reduced sensitivity) which results in chronically raised blood glucose levels and metabolic abnormalities[1]. Insulin is a hormone produced by the pancreas which controls the amount of glucose, the body's main energy source, in your bloodstream[1]. There are three main types of diabetes: Type 1 Diabetes Mellitus (T1DM), Type II Diabetes Mellitus (T2DM) and Gestational Diabetes Mellitus (GDM).

In 2019 the global diabetes prevalence was estimated to be 9.3% (463 million people) and this is expected to rise to 10.2% (578 million) by 2030[2]. It is thought that one in two (50.1%) of people with diabetes are unaware they have the condition[2]. Given the life-changing impact such a condition can have on sufferers, it is never more important to be able to identify those with the condition promptly in order to reduce their level of morbidity in later life.

The dataset to be used in the study is the Pima Indian Diabetes dataset from the National Institute of Diabetes and Digestive Kidney Diseases in the United States[3]. This dataset is often used to form models to predict whether or not one has diabetes[3].

The main purpose of this study is to see if it is possible to predict diabetes using the variables within the dataset. The authors will approach the dataset differently with regard to pre-processing and classification methods. Following this the methods used will be compared with one another and then with other research studies.

II. Literature Review

A lot of research has previously been undertaken using the pima indian diabetes dataset, often focusing on utilizing machine learning techniques to predict the diagnosis of diabetes.

AlJaullah et al[4] approaches the issues of missing values by entirely removing the features SkinThickness and Insulin. Further instances that had missing values in other features were removed. The remaining data was discretized to reduce the complexity, as this can lead to greater accuracy for some classifiers. 10 fold cross-validation and the classifier decision tree method was used with glucose as the root, achieving an accuracy of 78.16%.

Barale et al[5] approached missing values differently, they deleted instances that had two, three or four attribute values missing (234 instances deleted). This left 392 with no missing values but 142 cases still had one attribute value missing. Missing values were imputed using k-nearest neighbour (k=3). Outliers were detected with boxplots, as glucose had a large number of outliers these were also eliminated (36 instances deleted). A simple k-means clustering algorithm (k=2) was applied to extract hidden patterns. Misclassified sample cases were found and removed, leaving 349 instances. The dataset was partitioned into two sets (70% training and 30% testing) and 10 fold cross validation used. Several different classifiers were used (logistic regression (LR), artificial neural network (ANN), support vector machine (SVM), decision tree (DT)). By combining k-means algorithm with classifier achieved impressive accuracies, the best being k-means algorithm combined with artificial neural network achieving 99.3%. When the k-mean algorithm was not used the best classifier was SVM, at 79.32%.

Wei et al[6] explored the most popular classification techniques (deep neural network (DNN), Naive Bayes, LR, SVC, DT,) with the pima dataset. As with the previous studies, they also used 10 fold cross validation. They found that the classifiers used achieved their best accuracy when pre-processed data was scaled, hence they scaled their data. They noted there was little benefit of scaling with decision trees and naive bayes, due to the theoretical basis of these techniques. The most important predictive feature of the dataset was found to be Glucose, followed by Pregnancies, then Age. The least important feature was Insulin. The best accuracy achieved was 77.86% with DNN, followed closely by SVC (77.60%).

Huma Naz et al[7] employed four data mining algorithms i.e. DT, NB, ANN, and DL on the PIMA dataset to maximize accuracy in diabetes prediction, with DL being the best performing, at 98.07% accuracy. The dataset was trained using cross-validation and by tuning the following parameters: training cycles(500), learning rate(0.1) and used two hidden layers.

III. Methodology

A. Methodology Structure

This paper will initially explore the pima dataset to gain better understanding of the data contained. Following this both authors will take to different approaches to pre-processing and classifying the data. Pre-processing is an important step to producing tidy data to improve the performance of machine learning algorithms, and may involve dealing with missing values and outliers, feature selection and data normalization[8]. The aim is to predict, based on given attributes, whether diabetes is present or not. Therefore classification techniques will be used which predicts qualitative responses from observations. There are various classification techniques available and each author will use a different approach. The authors will then compare their different approaches in the discussion.

B. Exploration of Dataset

The dataset consists of eight predictor variables and one output variable. There are a total of 768 observations. The constraints placed on this particular dataset is that all the patients are female, at least 21 years of age and of Pima Indian heritage, who are North American Indication who originated from the state of Arizona, United States[9].

Each feature contained in the dataset is summarized in Figure 1 below.

Pima Dataset Description		
Attribute	***Description***	***Data Type***
Pregnancies	total number of times pregnant	Numeric
Glucose (mg/dl)	oral glucose tolerance test (OGTT results) - method to aid with the diagnosis of diabetes [9].	Numeric
BloodPressure (mmHg)	diastolic blood pressure - the blood pressure upon relaxation of the heart between beats[10]. Used in the diagnosis of Hypertension (persistently raised blood pressure[10].	Numeric
SkinThickness (mm)	triceps skin fold thickness - a measurement of subcutaneous fat thought to be indicative of body composition[11]. Triceps skin fold thickness correlates with estimates of total body fat in women[11].	Numeric
Insulin (mu u/ml)	measurement of blood insulin level two hours following ingestion of glucose. No clear diagnostic criteria for this attribute available	Numeric
BMI kg/m2)	Body Mass Index - used to determine whether an individual is overweight or obese, and is an estimate of body fat[12]. For adult BMIs: - Health Weight: 18.5 to 24.9 - Overweight: 25 to 29.9 - Obese: 30 to 39.9 - Severely Obese: 40 or more	Numeric
Diabetes Pedigree Function	score the likelihood of diabetes based on family history	Numeric
Age (years)	age of patient in years	Numeric
Outcome	target variable, indicates absence of presence of a diabetes diagnosis	Nominal

Fig. 1: Summary of Features of Pima Indian Diabetes Dataset

Upon initial inspection of the dataset it is clear there are a number of zeros within many features. For the features Pregnancies and Outcome the zeros are feasible values, as it is feasible a woman has not had any pregnancies and 0 in Outcome represents the absence of diabetes. However for the features BloodPressure, SkinThickness, Glucose, BMI, and Insulin the value zero is impossible, therefore this most likely represents missing values. The percentages of missing data in these features is displayed in Figure 2. There is a large amount of missing data on SkinThickness and Insulin, 29.6% and 48.7% respectively which will need to be tackled in data preprocessing as leaving these values as zeros may negatively impact model performance.

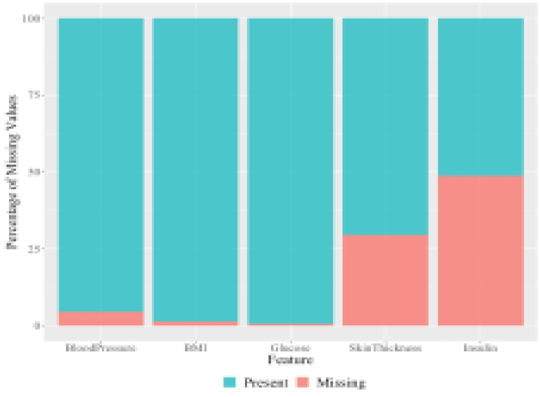


Fig. 2: Features with Missing Values as Percentages

Another potential problem to model performance are outliers that may exist within data. Outliers are observations that exist at extreme limits from other values. Outliers may be an error in data entry or valid data but they can be influential on statistical methods undertaken on data [10].

Boxplots are a good method to display and detect anomalies such as outliers within features. They also help visualize the distribution of data, displaying five descriptive values for the data: minimum, Q1, median, Q3 and maximum.

For this dataset boxplots can also be used to explore how features may influence the presence of diabetes by separating the attributes by their Outcome.

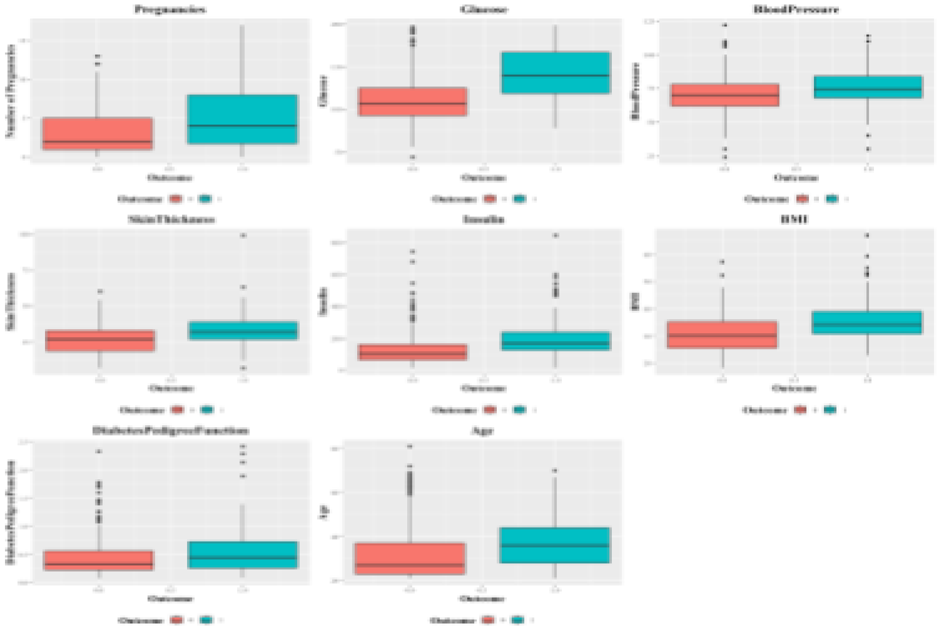


Fig. 3: Boxplots for each feature grouped by Outcome

Regarding outliers, these boxplots show that except for Pregnancies and SkinThickness, each feature has quite a few outliers that will need to be addressed.

The spread of the data compared by Outcomes appears fairly similar for most features, with the spread for Pregnancies, Glucose and DiabetesPedigreeFunction where the Outcome is 1 (diabetes present) appear to have a data which is more spread out than those who do not have diabetes in those features.

The distribution for some features appear skewed. Notably Pregnancies and Age are right (or positively) skewed meaning that most of the data falls into the lower range of values, i.e. most participants are younger and have had fewer babies.

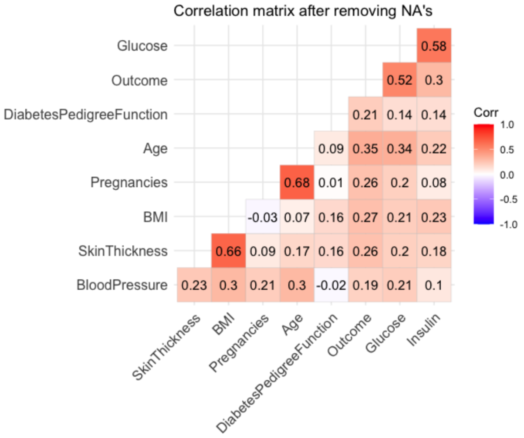


Fig. 4 Correlation matrix after removing missing values

The correlation matrix confirms that even after removing any missing values, the variables Glucose and Insulin are highly correlated, as well as SkinThickness and BMI. What this could mean is that there is another factor that is causing these correlations[11], namely the presence of diabetes in this case.

C. Approach One (Celeste Webb)

Data Pre-Processing

Following exploration the data is pre-processed in preparation for modeling. Data exploration highlighted a number of missing values, the most concerning being the large number missing from SkinThickness and Insulin.

Data can be missing for a number of reasons, commonly classified as MCAR (mining completely at random), MAR (missing at random) and MNAR (missing not at random). Given that we know little about data collection and therefore why some data values are missing, it is difficult to determine why this data may be missing[12].

One approach others have taken to deal with these missing values[5] is to delete all instances that had missing values, but this approach significantly reduces the number of instances. Another option would be to impute these values, which could be reasonable for features missing only a few values, such as BMI but for Insulin, which is missing 48.7%, this would introduce inaccuracy. The approach chosen was to remove the features

SkinThickness and Insulin as AlJaullah[4] chose to do, as this preserved more instances.

BloodPressure, BMI and Glucose are missing much less data (35, 11 and 5 instances respectively), deleting all these instances would leave 724 out of 768 instances (losing 5.73% instances). To minimize any further data loss I initially chose to impute the missing values. There are two main types of imputation available: single imputation, where one value is substituted for the missing value, or multiple imputation, in which several plausible imputable data sets are created and combined to form imputable values[12]. Although multiple imputation programs are built on the MAR assumption, the method can handle MCAR and MNAR[13]. The missing data imputation method of multiple imputation by using chained equations (MICE) has been shown to outperform other methods[14], therefore this method was chosen to replace the missing values in BloodPressure, BMI and Glucose.

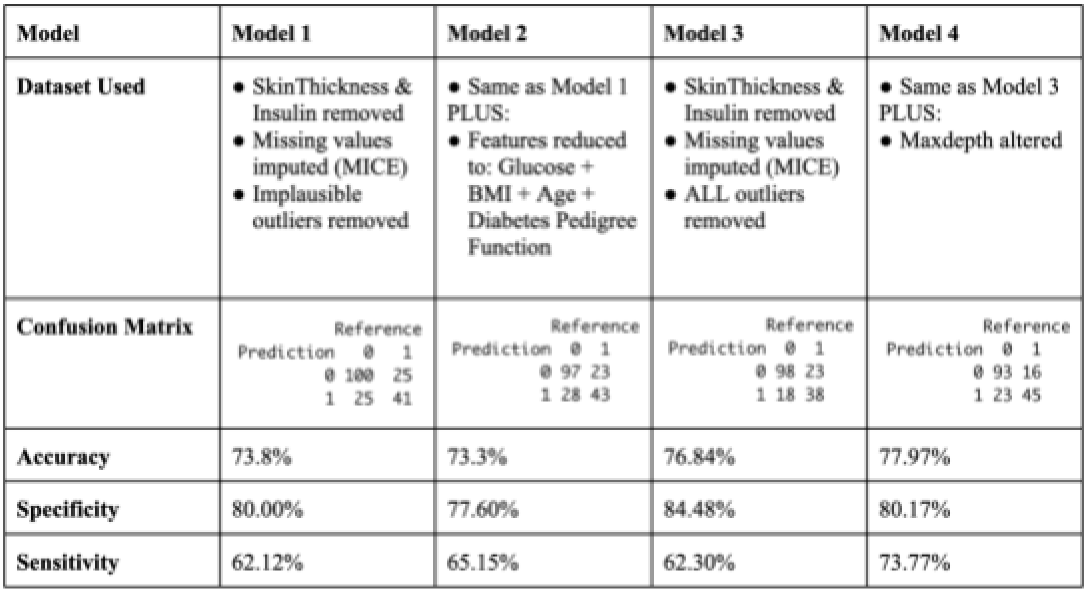
The outliers were then examined in closer detail. Several BloodPressure values were found to be implausible, likely errors, therefore these instances were removed. Regarding DiabetePedigreeFunction it is difficult to comment on whether or not the outlier values are implausible or not as there is little information available on this feature, so no instances were removed.

Classification Model Methodology

To produce a classification model for the data I chose to use the popular supervised learning model Decision Tree Classification. I created four different models which used slightly different methods as summarized in Fig 5, attempting to improve each model's ability to classify the presence of diabetes.

The following methodology was applied to each model:

- The seed is set to ensure replicability (chosen seed 1234)
- The processed dataset was partitioned into a training / testing sets (initial split used 75/25)
- To ensure the two datasets were similar their summary statistics were checked and the data visualized with histograms, box plots and scattergrams to ensure their features had similar distributions, outliers, ranges, and correlations.
- Using the package 'rpart' in RStudio a classification decision tree model was created using the training dataset
- The summary and visual representation of the decision tree was inspected
- Model used to make predictions on the testing dataset
- Performance of the model assessed by using confusion table metrics (accuracy, sensitivity, specificity).



Model	Model 1	Model 2	Model 3	Model 4
Dataset Used	• SkinThickness & Insulin removed • Missing values imputed (MICE) • Implausible outliers removed	• Same as Model 1 PLUS: • Features reduced to: Glucose + BMI + Age + Diabetes Pedigree Function	• SkinThickness & Insulin removed • Missing values imputed (MICE) • ALL outliers removed	• Same as Model 3 PLUS: • Maxdepth altered
Confusion Matrix	Reference Prediction 0 1 0 100 25 1 25 41	Reference Prediction 0 1 0 97 23 1 28 43	Reference Prediction 0 1 0 98 23 1 18 38	Reference Prediction 0 1 0 93 16 1 23 45
Accuracy	73.8%	73.3%	76.84%	77.97%
Specificity	80.00%	77.60%	84.48%	80.17%
Sensitivity	62.12%	65.15%	62.30%	73.77%

Fig. 5 Summary of Models

Model Results

Model One: This used the first dataset created by pre-processing. The model's most important feature was shown to be Glucose (60), followed by Age (13), BMI (11), BloodPressure (6), Pregnancies (5), and DiabetePedigreeFunction (4). This model achieves an accuracy of 73.82%. Interestingly it misclassified the two groups equally.

Model Two: Having learnt from Model One that the most important features for classification are Glucose, Age and BMI this was used to see if by reducing the number of features would improve the model's performance. Various combinations of reduced features were used. No combination improved upon Model One's accuracy (73.82%).

Model Three: Returned to the processed dataset and removed all instances that contained outliers as these were potentially affecting the model's performance. Removal of all outliers reduced the dataset to 707 instances. This improved the model's accuracy and specificity but unfortunately the sensitivity reduced.

Model Four: Given the improvement in several metrics the dataset in which all the outliers were removed was used in this model. The maxdepth parameter in 'rpart' sets the *'maximum depth of any node of the final tree, with the root node counted as depth 0'*[15], its default value is 30. The current decision tree (with the default maxdepth set to 30) does not extend further than 7 levels. Maxdepth was altered from 1 through to 7 and performance assessed. The best accuracy (79.1%) was achieved for maxdepth 5, the best sensitivity (73.77%) for maxdepth 4, and the best specificity (96.55%) for maxdepth 3. The decision tree created using a maxdepth 5 can be seen below in Fig 6.

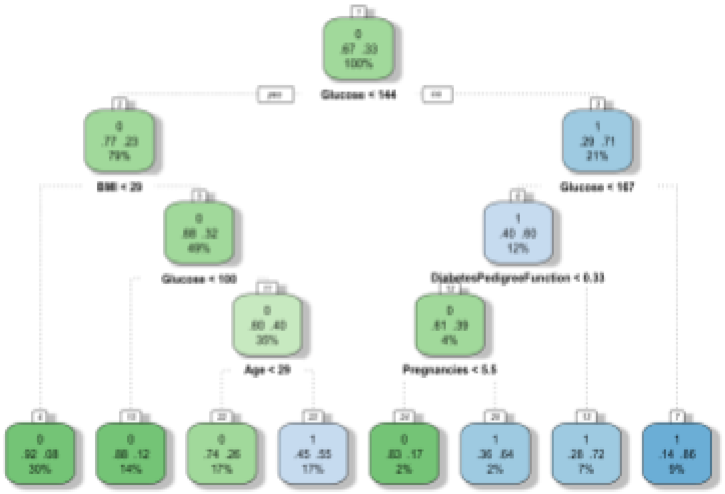


Fig. 6 Decision Tree for Model 4 with maxdepth 5

D. Approach Two - Octavia Istocescu

I. Methodology

Data Analysis and Exploration

Before commencing with pre-processing and building the models, the dataset was explored in order to get myself familiarized with the data and to detect any anomalies, steps which would influence my choice of pre-processing methods. A mix of visualizations and data analysis was used at this stage.

Pre Processing

Pre-processing aims at assessing and improving the quality of data to allow for reliable statistical analysis[16]. After the exploratory analysis of the data, the only issues found were missing values for a few variables and outliers, which need to be addressed. The dataset has to go through several pre-processing steps before we can build our models, otherwise the missing values and outliers can introduce bias and skew our analysis and results. Other than this, the dataset is tidy and doesn't require further cleaning up.

The first step was transforming all the 0 values from the variables Glucose, Blood Pressure, Skin Thickness, Insulin, BMI and Diabetes Pedigree Function into NA's so that the dataset could be used for visualizations and imputed later on.

Outliers

Outliers were found in all variables but Glucose by using boxplots. However, such high measurements of Insulin and BloodPressure are most likely erroneous and because it was found that a classifier degrades in performance with the presence of noise[17], they were removed in Model 1 and Model 2 so that they do not skew the results.

Various techniques have been proposed for dealing with outliers, one of them being oversampling[18]. However, our dataset only has 0.06% outliers, which is not a significant number, thus I opted for removing the ones in Model 1 and Model 2 only for variables SkinThickness and Insulin which were implausible, as the model won't be affected by the loss of instances.

Following the second author's approach, a further two models, namely Model 3 and Model 4, were explored with outliers removed from all variables.

Partitioning the dataset

The dataset was partitioned into training and testing, then both of them checked to see if their distributions were similar using visualizations.

Missing data imputation

The dataset doesn't come with an extensive documentation that details how exactly the data was gathered. Thus, it is unclear to us how the missing data came to be.

These are the three common reasons for missing data:

- Missing Completely at Random (MCAR): the missing value is unrelated to the value of other observed variables
- Missing at Random (MAR): the missing value is related to other observed values in the dataset but not to the variable itself

- Missing Not at Random (MNAR): the missing data depends on both missing and observed values

Little's MCAR test was used to check whether the missing data is MCAR. The test statistic is a chi-squared value that proposes a null hypothesis that the missing data is *Missing Completely At Random*. The p-value we got was 1.28 × 10^-9, less than 0.05, so we interpret it as being that the missing data is not MCAR[19].

The choice of imputation method should be influenced by the reason for the missing data. In our case, values are missing for unidentifiable reasons, so we assume that they are missing because of random and unintended causes. This makes them recoverable using various imputation methods[16].**]**

There are 652 missing values in the dataset, or 9.4% , so removing them is not a viable option,as they would leave us with a reduced dataset that would most likely not generate a lot of insights. Besides that, leaving them in may lead to a loss of predictive power and ability to detect statistically significant differences and it can be a source of bias, affecting the representativeness of the results[16].

Single imputation replaces the missing observations with a single value, which disturbs the relations between variables and introduces bias[20].

Therefore, I opted to use multiple imputation over single imputation in order to replace the missing observations, namely Amelia imputation and MissForest imputation, which was shown to be a highly accurate method in clinical predictive models[21].

Feature Selection/Reduction

In an attempt to try and increase accuracy, the model was tested on a subset of five features that ranked very highly in terms of relevance to make a diagnosis, namely Glucose, BMI, Age, Insulin and Skin Thickness[22]. A further attempt at reducing the dataset was made to improve performance and for comparison purposes by implementing the second author’s way of only removing the features SkinThickness and Insulin, leaving me with 6 features.

Classification

An Artificial Neural Network algorithm was used to produce a classification for the Pima Indians Diabetes dataset and six different models were created and compared.

The classification was done iteratively, the pre-processing was done in stages and each model was tested on the dataset at each of the stages. The stages can be seen in Fig. 7.

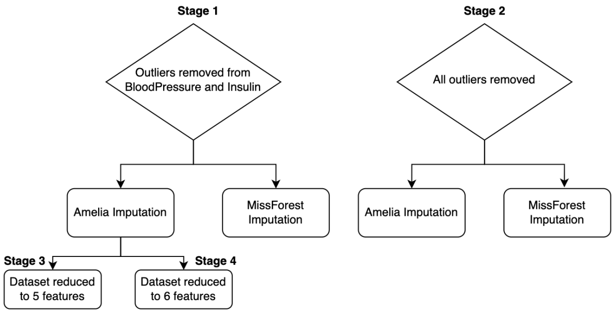


Fig. 7 Approach Two classification stages

II. Results

Data Analysis and Pre-Processing stage

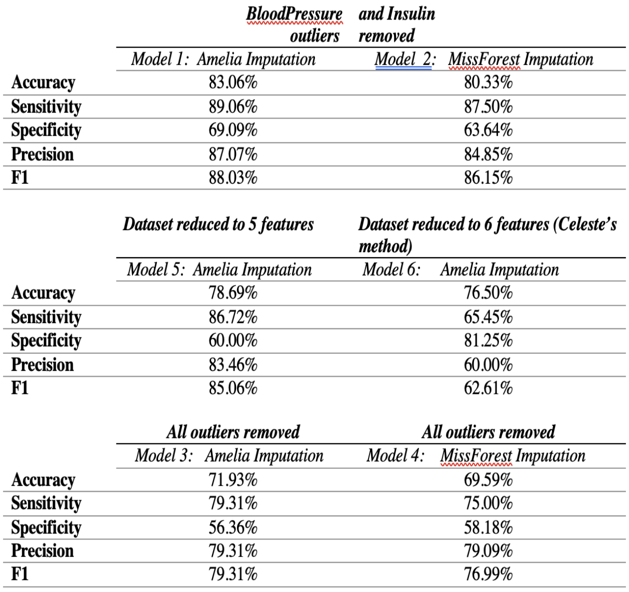
Visualization was an important tool for insight generation and discovering anomalies in the data and was successful in pointing me towards the areas that needed pre-processing, namely discovering missing observations disguised by zero values and boundary violations in the form of outliers.

Visualizing the percentage of missing data for every age group gave another valuable insight. We can see that for the variable Insulin, the percentage of missing values grows as the age gets higher, with over 60% of missing data for patients over 60. However, at closer inspection, it becomes apparent that this is because the number of observations significantly decreases as the patients get older. Removing the entries with missing values entirely will result in excluding observations for patients over 60 and the classification model will not generalize well for older patients, as it will likely introduce bias.

Classification stage

The performance metrics used were Accuracy, Sensitivity, Specificity, Precision, F1-Score. The summary of models can be seen in Fig. 8. Before running each model the seed “1234” was set for reproducible results.

In stage one and stage two, which tested the models with outliers from BloodPressure and Insulin removed and used Amelia and MissForest Imputation, Model 1, which underwent Amelia imputation, with a score of 83.06% Accuracy out-performed Model 2 that used MissForest imputation on all of the metrics, therefore that is the one that was carried forward to the feature reduction phase. So far, this suggested that Amelia imputation was more appropriate and better performing.



	BloodPressure outliers	and Insulin removed
	Model 1: Amelia Imputation	Model 2: MissForest Imputation
Accuracy	83.06%	80.33%
Sensitivity	89.06%	87.50%
Specificity	69.09%	63.64%
Precision	87.07%	84.85%
F1	88.03%	86.15%

	Dataset reduced to 5 features	Dataset reduced to 6 features (Celeste's method)
	Model 5: Amelia Imputation	Model 6: Amelia Imputation
Accuracy	78.69%	76.50%
Sensitivity	86.72%	65.45%
Specificity	60.00%	81.25%
Precision	83.46%	60.00%
F1	85.06%	62.61%

	All outliers removed	All outliers removed
	Model 3: Amelia Imputation	Model 4: MissForest Imputation
Accuracy	71.93%	69.59%
Sensitivity	79.31%	75.00%
Specificity	56.36%	58.18%
Precision	79.31%	79.09%
F1	79.31%	76.99%

Fig. 8 Approach Two model results

However, Stage two results suggest that removing outliers from all variables did not improve performance. In fact, it significantly decreased it, although Model 3 using Amelia imputation once again out-performed Model 4 using MissForest Imputation on four out of the five metrics and by 2.34% on Accuracy.

The best performing model out of the four tested so far was Model 1, so it was carried forward for further testing in the feature reduction stages. It surpassed all of the other models on all of the performance metrics.

At Stage three, Model 1 was tested on a dataset reduced to five features that were deemed the most influential in making a prediction about diabetes. Once again, further pre-processing did not improve performance when compared by the Accuracy score, however it didn't degrade by a lot in the following: Sensitivity (86.72%), Precision (83.46%) and F1 (85.06%).

Since the models were declining in performance with every stage of pre-processing, in Stage four the second author's method of removing SkinThickness and Insulin was tested on Model 1 in an attempt to improve performance. The model scored less than Model 5, with a significant decrease in Sensitivity (65.45%), Precision (60%) and F1 (62.61%)

Finally, the best performing model was Model 1, with an Accuracy of 83.06%, which underwent minimal pre-processing.

IV. Discussion

With regards to Approach One it was found that reducing the features used failed to improve accuracy, however it was possible to increase the sensitivity to 65.15% (from 62.12%), but at the cost of slightly reducing the specificity (80 to 77.60%) when including only Glucose, BMI, Age, and DiabetePedigreeFunction. Therefore, reducing features can improve some performance metrics but others will suffer as a trade-off.

The model that performed best in Approach One involved removing all of the outliers, regardless or whether or not they were plausible values and setting the maxdepth of the tree to 5. This resulted in accuracy, specificity and sensitivity all improving thus suggesting these outliers were negatively impacting the ability of the model to predict the presence of diabetes.

Whilst trialing different maxdepth values it was found that reducing the value to 3 or less the sensitivity of the model significantly deteriorated. There was a trade-off regarding maxdepth and different performance metrics. The best accuracy (79.1%) was achieved for maxdepth 5, the best sensitivity (73.77%) for maxdepth 4, and the best specificity (96.55%) for maxdepth 3. The sensitivity of a model is arguably one of its most important aspects when used in healthcare as you would not want to miss an individual with a disease, especially one that could be life-limiting or cause significant morbidity. Therefore in order to preserve a good level of sensitivity the best option would be maxdepth of 4. This option provides the best accuracy and sensitivity out of all of my models.

When comparing Approach One's results to other research studies results it performed fairly well. Wei et al[6] achieved their best accuracy (71.10%) using the Decision Tree on the original data prior to any imputation, normalization or scaling. Wei et al did not remove any missing values (instead imputing them all) or outliers which suggests that Approach One's method of removing missing values, imputing as few values as possible and removing all outliers improved accuracy. AlJarullah et al[4] achieved a better accuracy from Approach One with 78.1768%. AlJarullah et al[4] undertook no imputation and instead removed attributes SkinThickness and Insulin and any instance with missing values, they discretized the attributes and used 10-fold cross-validation, all which may explain the improved accuracy compared to Approach One.

Regarding Approach Two, surprisingly, pre-processing the dataset did not improve accuracy, unlike what happened with the second author's approach. In fact, it decreased accuracy as much as 17.6% for Model 4.

However, Approach Two achieved higher performance on three of the models implemented, when compared by Accuracy (83.06%)

In model 6, which followed the second author's method of reducing features to 6 by removing SkinThickness and Insulin, with only BloodPressure and Insulin outliers removed and Amelia imputation done, performance declined significantly. This might be because those two features have a lot of influence and importance on the accuracy of predicting diabetes.

However, removing them in Approach One might be the reason why it's best performing model had significantly worse performance than the best model in Approach Two, with a 6.3% difference.

Neither removing all outliers didn't have a positive effect on performance, which boosted performance in the second author's models, but actually decreased it when tried out by the first author. Since removing SkinThickness and Insulin didn't help, we can only attribute the loss of performance to the choice of imputation method and classification algorithm used.

Compared to other research study results that used Artificial Neural Networks to predict diabetes on the Pima dataset, the best performing model from approach Two performed well, which came 7.28% percent short of Huma Naz et al[7] model. I attribute this difference in performance to their usage of parameter optimization, despite not doing any missing data imputation or outlier removal. Huma Naz et al used two hidden layers, performed 500 iterations, whereas I only used one hidden layer and performed 100 iterations. They also used sampling to tune the learning rate of the model to 0.1.

Inspired by Chang V. et al[23], the pre-processing was done in stages and features were reduced in an attempt to improve accuracy.

V. Conclusion

In this paper we used the Pima Indian Diabetes dataset to propose two different approaches to pre-processing and classifying the data to predict the presence of diabetes.

Approach One's best performing model approached the data by removing the attributes Insulin and SkinThickness that had a significant amount of missing values, imputing leftover missing values via MICE, and removing all outliers. Approach One then used the Decision Tree Classifier achieving an accuracy of 77.97% when the maxdepth of the tree was set to 5.

Approach Two achieved the best performance by using minimal pre-processing with only the extreme outliers

One shortcoming of our work is the choice to partition our dataset into a training/testing set, which can potentially cause overfitting and underfitting our model thus affecting the predictability of our data.

Another option, which was used by some of the researchers within the literature review, is k-fold cross-validation where the dataset is divided randomly into k subsets. k-1 subsets are used to train the model and the remaining subset to test[18]. The average of all the metrics is taken from the subsets giving the best estimate of the "true"performance of the model[12]. This method gives a better estimate of the performance of a model than partitioning into training/testing.

Another shortcoming is that the two authors did not have the same partitions of training and testing sets, which does not make the comparison completely equal and unbiased.

One limitation of the Pima Indian Diabetes Dataset is that it consists only of females which may have introduced unknown bias, making any findings difficult to generalize to males. The variables within the dataset are quite limited. For a more accurate and unbiased model the dataset needs to have a more balanced distribution of ages among the participants. There are other known risk factors for diabetes such as family history, diet, exercise level etc. There was also a significant amount of missing data within the dataset.

The model is also most likely not generalizable to women of other ethnicities. This is because Pima Indians might have a higher genetic predisposition to diabetes[7]. In order to enable the model to be generalizable, it would need to remove DiabetesPedigreeFunction attribute, which is not commonly calculated to help with diabetes diagnosis in a real medical scenario.

Future prediction models could also be improved by the use of more robustly collected data with more variables.

Acknowledgment

Both authors are thankful to our professors at the University of Nottingham for providing their knowledge and guidance to carry out this research.

References

1. O'Neill R, Murphy R, Horton-Szar D. Crash Course Endocrinology. London: Elsevier Health Sciences UK; 2015
2. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition [Internet]. International Diabetes Federation [cited 6 April 2002]. Available from: https://www.diabetesresearchclinicalpractice.com/article/S0168-8227(19)31230-6/fulltext
3. R: PIMA Indians diabetes dataset. [Internet]. Search.r-project.org. [cited 6 April 2022]. Available from: https://search.r-project.org/CRAN/refmans/hhcartr/html/pima.html
4. AlJaullah, A. Decision Tree Discovery for the Diagnosis of Type II Diabetes. 2001 International Conference on INnovations in Information Technology. 2011; 303-7
5. Barale M, Shirke D. Cascaded Modeling for PIMA Indian Diabetes Data. International Journal of Computer Applications. 2016; 139(11):1-4
6. Wei S, Zhao X, Miao C. A comprehensive exploration to the machine learning techniques for diabetes identification. 2018 IEEE 4th World Forum on Internet of Things (WF-IoT). 2018; 291-5
7. Naz, H., Ahuja, S. Deep learning approach for diabetes prediction using PIMA Indian dataset. J Diabetes Metab Disord 19, 391–403 (2020). https://doi.org/10.1007/s40200-020-00520-5
8. Kotsiantis S, Kanellopoulos D, Pintelas P. Data Preprocessing for Supervised Learning, International Journal of Computer Science. 2006; 1(2):111-7
9. Pima [Internet]. Encyclopedia Britannica. [cited 6 April 2022]. Available from: https://www.britannica.com/topic/Pima-people
10. Flynn J. Oxford American handbook of clinical medicine PDA. Oxford: Oxford University Press; 2008.
11. Larose D. Data Mining Methods and Models. New Jersey: John wiley & Sons, Inc.; 2006
12. Cady, Field. The Data Science Handbook, John Wiley & Sons, Incorporated, 2017.ProQuestEbookCentral,https://ebookcentral.proquest.com/lib/nottingham/detail.action?docID=479065
13. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338(7713):79–160
14. Pedersen A, Mikkelsen E, Cronin-Fenton D, Kristensen N, Tra M, Pedersen L, et al. Missing data and multiple imputation in clinical epidemiological research. 2017;
15. Mohammed, M, Zulkafli, H, Adam, M, Baba, I. Comparison of five imputation methods unhanding missing data in a continuous frequency table. AIP Conference Proceedings. 2021; 2355(1)
16. Therneau T, Atkinson B, Ripley B. Recursive Partitioning and Regression Trees [Internet]. Cran.r-project.org. 2022 [cited 26 April 2022]. Available from: https://cran.r-project.org/web/packages/rpart/rpart.pdf
17. Secondary Analysis of Electronic Health Records [Internet]. Cham (CH): Springer; 2016. Available from:https://www.ncbi.nlm.nih.gov/books/NBK543630/ doi: 10.1007/978-3-319-43742-2
18. Napierała, K., Stefanowski, J., Wilk, S. (2010). Learning from Imbalanced Data in Presence of Noisy and Borderline Examples. In: Szczuka, M., Kryszkiewicz, M., Ramanna, S., Jensen, R., Hu, Q. (eds) Rough Sets and Current Trends in Computing. RSCTC 2010. Lecture Notes in Computer Science(), vol 6086. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-13529-3_18
19. Nonso Nnamoko, Ioannis Korkontzelos, Efficient treatment of outliers and class imbalance for diabetes prediction, Artificial Intelligence in Medicine, Volume 104, 2020, 101815, ISSN 0933-3657, https://doi.org/10.1016/j.artmed.2020.101815. (https://www.sciencedirect.com/science/article/piiS093336571830681X)

20. Roderick J. A. Little (1988) A Test of Missing Completely at Random for Multivariate Data with Missing Values, Journal of the American Statistical Association, 83:404, 1198-1202, DOI: 10.1080/01621459.1988.10478722
21. Stef van Buuren, Flexible Imputation of Missing Data, Second Ed., Chapman & Hall/CRC Interdisciplinary Statistics Series, https://stefvanbuuren.name/fimd/
22. Waljee AK, Mukherjee A, Singal AG *, et al*Comparison of imputation methods for missing laboratory data in medicine*BMJ Open* 2013;3:e002847. doi: 10.1136/bmjopen-2013-002847
23. Chang, V., Bailey, J., Xu, Q.A. *et al.* Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. *Neural Comput & Applic* (2022). https://doi.org/10.1007/s00521-022-07049-z