# Predicting Diabetes Using Supervised Machine Learning Classification

Liam Hampson, Mihir Shrestha

University of Nottingham

**Abstract**

The objective of this paper was to answer three main research questions regarding the dataset of Pima Indian women. Do the independent variables predict diabetes? If so, which variables are the most predictive? Finally, can a good model be built that can classify diabetics vs nondiabetics with strong accuracy. Through data analysis and visualisation, it is determined that all the independent variables are at least weakly associated with diabetes, and the three most predictive are glucose, insulin and BMI. By evaluating and comparing two different data imputation techniques (mean and kNN) scaling techniques (normalisation and standardisation), and three different models (SVM, logistical regression and random forest. It was ascertained that the optimal model that could be built was random forest with a normalised, kNN imputed training set. This model gave an accuracy of 80.7%, with a sensitivity of 84% and a specificity of 74.4%. The trained model is in line with the performance of similar models from literature, and generally performs to a high standard on unseen data.

## I. INTRODUCTION

Type II Diabetes Mellitus is a serious disease that causes the body to become resistant to a chemical, insulin, that is produced by the pancreas. When this occurs, insulin is unable to utilise glucose in the bloodstream as an energy source effectively, leading to high blood sugars [1]. Type II diabetes is a common disease. In 2019, there were approximately 438 million identified cases of type II diabetes globally [2], and 1 in 10 people over 40 in the UK are now living with the disease [3]. The symptoms of type II diabetes often develop very gradually, and as such, can be difficult to notice until many years beyond the onset of the disease. This is problematic as untreated diabetes causes major damage to the body over time and can lead to life-limiting or debilitating complications such as kidney disease, heart failure, blindness and lower limb amputation [4]. Type II diabetes is incurable, however, the risk of developing the disease can be substantially reduced with lifestyle changes, and the risk of serious complications developing in those affected can be mitigated with effective management [5]. Given this, it is vital that individuals at above average risk of having diabetes are accurately identified so that they can reduce their risk or obtain a timely diagnosis. Known major risk factors of the disease include age, BMI and blood pressure.

Predicting whether not an individual has type II diabetes based on risk factors is a major motivation for this paper. Currently, type II diabetes is usually diagnosed when a patient presents with symptoms or when a doctor suspects diabetes and refers the patient for tests, which often leads to late diagnosis or missed cases. Data analysis techniques can be used to train diagnostic models that can automatically identify suspected cases of diabetes based on patient medical data. Achieving this to a reasonable accuracy would aid in the early detection or prevention of diabetes, as well as reducing the strain and costs to healthcare systems [6], both by saving healthcare staff time in the diagnosis process as well as reducing the number of comorbidities needing treatment as a result of undetected diabetes.

The dataset that will be analysed and used to train classification models in this paper is the Pima Indian diabetes dataset. This is a subset of a larger dataset collected by the National Institute of Diabetes and Digestive and Kidney Diseases and contains medical information of 768 women of Pima Indian descent aged 21 and above. The dataset contains 8 predictive (independent) variables, age, insulin, blood pressure, glucose, body mass index, number of pregnancies, diabetes pedigree function and skin thickness. Additionally, there is 1 target (dependant) variable, outcome, which describes whether the individual has been diagnosed with type II diabetes. The Pima Indian people are a population of 20,000 Native Americans found primarily in central and southern Arizona. They have the highest rate of type II diabetes in the world [7] and have been under epidemiological study in 2-year intervals since 1965. As type II diabetes is believed to be caused by a combination of genetic and environmental factors [8], the dataset contains independent variables believed to correlate to the onset of diabetes.

The research questions posed for this analysis are: do the independent variables described in the dataset correlate to a diagnosis of type II

diabetes? If so, which variables correlate most strongly to the onset of diabetes? Finally, can a diagnosis of diabetes be predicted using the features of the dataset by training diagnostic classification models?

## II. LITERATURE REVIEW

When looking at literature on analysing similar datasets, researchers have used a wide range of different modelling, pre-processing and analysis techniques. A compilation table showing methods used in select papers is below.

| Ref | Pre-processing | Classification | Accuracy | Feature Selection | Main Features/ Main Contributions | Difficulties and failings/ Minor Drawbacks |
|---|---|---|---|---|---|---|
| (Ramesh, Aburukba and Sagahyroon, 2021) | • Standardisation Scaling<br>• Multivariate imputation by chained equations (MICE) method | KNN | 79.80 | •Chi- squared test,<br>•Extremely randomized trees classifier<br>•Least absolute shrinkage and selection operator | Automates Diabetes detection and alerts medical professionals so they can intervene on time. | The method hasn't yet been tested on large- scale patient datasets.<br>The system requires yet additional testing with hardware integrations. |
| | | SVM | 83.20% | | | |
| | | Logistic Regression | 73.30% | | | |
| | | Naive Bayes | 73.10% | | | |
| (Khanam and Foo, 2021) | • Mean Imputation<br>• Normalization | Logistic Regression | 77% | • Pearson's correlation | Implemented a system utilising Weka; an open source machine learning and software tool; for diabetes dataset's performance analysis.<br>Which can predict diabetes with high accuracy based on 5 input features. | |
| | | SVM | 78% | | | |
| | | Naive Bayes | 78% | | | |
| | | Random Forest | 77.3% | | | |
| (Alaa Khaleel and Al-Bakry, 2021) | • Min Max Scalar Normalization | Logistic Regression | 94% | None | | Only 3 machine learning algorithms were tested |
| | | Naive Bayes | 79% | | | |
| | | K- nearest | 69% | | | |
| (Hrimov et al., 2021) | None | SVM | 85.71% | Backward elimination | Irrelevant attributes were stripped away by backward elimination to identify the most essential features | |
| (Garcia-Ordás et al., 2021) | Normalization augmentation using VAE | SAE+ CNN (convolutional neural network) | 92.31% | None | Convolution neural network was trained to carry out the classification.<br>A new architectural approach combining the Sparse Autoencoder, and Convolution classifier was proposed to obtain a very high accuracy result. | Limited due to the small number of samples in the dataset.<br>Future work could be improved by creating a dataset with more valuable characteristics and more individuals. |
| | | SAE with MLP (multilayer perceptron) | 85.71% | | | |
| (Sisodia and Sisodia, 2018) | None | SVM | 65.1% | None | Three machine learning algorithms are used to detect diabetes. And their performances are evaluated | Only three machine Learning algorithms were used to test the data.<br>No Pre processing or Feature Scaling were applied. |
| | | Naives Bayes | 76.3% | | | |
| | | Decision Tree | 73.82% | | | |
| (Dey, Hossain and Rahman, 2019) | Min-Max Scaling | Artificial Neural Network | 82.35% | None | Created an architecture to predict if a patient is diabetic or not.<br>A web application based on the higher precision of a strong learning algorithm was built. | The Application wasn't fully developed to be launched to the public.<br>No Feature Scaling was utilised. |
| (Faruque, Asaduzzaman and | None | C4.5 Decision Tree | 73% | None | Comparison of 4 different machine learning algorithm to compare the Accuracy. | Only 4 algorithms were used to test the data. |

*Figure 1: Lit Review Table*

## III. METHODOLOGY

### A. Pre-processing

The data pre-processing stage involved preparing the data for analysis and modelling by splitting the data into training and test sets, data cleaning and feature scaling via different approaches. Two approaches were used to split the data. The first approach was simply to split the date into two subsets in an 80/20 ratio, training and test sets, respectively. This is done so that classification models can be trained on the training set and evaluated fairly on the test set, comprised of totally unseen data, without being influenced by overfitting to the training set. Evaluating a model's accuracy on the training set would yield overly optimistic results. Optionally, a third set, validation, is introduced in between the two sets in the ratios 70/15/15. The validation set is used to evaluate a model while its hyper parameters are tuned. Evaluating performance on the test set would result in overfitting, as the model adapts and fits to the test data as it is repeatedly passed through the model. The final model evaluation should be done on totally unseen data, and the model should not be tweaked after the final evaluation. In this paper, both approaches have been used. two of the three models tested use the 70/15/15 split and one uses the 80/20 split, this is because one of the models required no tuning, so a validation set was not required. It should be noted that when directly comparing the accuracy of the models with one another, all the models were trained and tested on same sets (15%) to keep the comparison fair and minimise bias. All pre-processing steps are also trained on the training set and then applied to the test/validation sets. It was essential to keep the seed constant when writing the code so the data split in the same way each time the code was run. Training and validation sets split in different ways will result in slightly different results during the modelling and analysis of the data.

The next step is data cleaning, which involved searching the data set for outliers, incorrect and missing data. The method used to handle missing data was data imputation, which involves replacing the missing or incorrect data with substituted values that are a best guess estimate of the true values. Imputation was chosen as the method because the dataset had a very large amount of missing data, so other methods, such as simply deleting the rows with missing data, would result in an unacceptable reduction in the dataset size, reducing the reliability of any models trained on the data. Two different approaches of data imputation were tried, mean imputation and kNN imputation. Mean imputation is the process of simply substituting all missing values with the mean of all the true values for a given variable. The justification for using mean imputation is that it allowed the full size of the dataset to be preserved, the mean of the data was unchanged in the process, so remained unbiased, and the method is a simple and widely used solution for missing data. The second approach used was K-nearest neighbour (kNN) imputation. The kNN method is a supervised machine learning algorithm that finds the K nearest data points to any given data point in feature space. The missing value for a data point is calculated by taking the average value of the nearest K neighbours. The kNN algorithm uses distance metrics to calculate which data points are closest (most similar) to any given data point. The main justification for using kNN over other types of imputation is that kNN is multivariate, meaning that it takes multiple variables into account when calculating the nearest neighbours, so imputation is likely to be more

accurate when the missing value is dependent on multiple variables.

The final pre-processing step performed was feature scaling. Feature scaling involves restricting the range of values allowed for the variables of a dataset. This is important when dealing with algorithms that use distance-based metrics, such as kNN, as features with very large ranges will be more heavily weighted in the distance calculation and bias the results. Scaling features ensures that each variable will be weighted approximately equally in any distance-based algorithm. Two methods of feature scaling were considered in this paper, normalisation and standardisation, also known as min-max normalisation and Z-score normalisation, respectively. Normalisation is a simple technique where all dataset features are rescaled to have a maximum value of 1 and a minimum value of 0. The justification for using normalisation is that all features have the same scale after it has been applied, so there will be no bias in distance-based algorithms. Additionally, the shape of the data is unchanged because the relative distance between each data point is preserved. Alternatively, standardisation is a feature scaling method in which the mean and standard deviation of each feature is set to 0 and 1, respectively. Standardisation generally requires that a dataset follows a Gaussian distribution to be implemented effectively. The justification for using standardisation is that it handles outliers very effectively whilst also rescaling features to be closer to one another.

B. Data Analysis

Exploratory data analysis (EDA) is an important stage of any data analysis project. The goal of EDA is broad, but generally involves making interesting, useful or unexpected insights from a dataset using a range of visualisation and statistical techniques. Examples of different approaches of EDA used in this project include correlation heat-maps, boxplots, histograms, bar charts and scatter plots (univariate and multivariate graphs). Univariate non-graphical approaches such as range, mean, root mean squared error and percentage bias calculations are also considered. The justification for using the above EDA techniques in this project is that it allowed pre-processing techniques and classification models to be critically compared and selected. Furthermore, making direct statistical measures on the dataset aided in answering the core research questions and evolving them for future consideration.

C. Classification

The final stage of the project was training various classification models on the cleaned dataset to predict whether an individual has type II diabetes. Three different classification models were trained in the project, support vector machine (SVM), random forest and logistical regression. All three classifiers are supervised learning algorithms that are ideal for binary classification and have been extensively tried and tested with high accuracy results, which is why they have been selected.

In an SVM classifier each sample is plotted in an n-dimensional feature space, where n is the number of features. The SVM then calculates a separating hyperplane that cuts through the data points. The hyperplane acts as a decision boundary for each class, all data points on one side are assigned to class 1 and the data points on the other side are assigned to class 2. Different kernels can be used in SVM which change the shape of the hyperplane. In this instance, a radial kernel was used because it assumes no knowledge about the data and does not require the data to fit to a particular distribution or linearity, so it is a strong default kernel to use when the relationship between each variable in the dataset is not fully known. Radial SVM has two hyper-parameters that can be tuned to optimise the algorithm, cost (C), which is essentially how far the decision boundary is allowed to bend to correctly classify the data, and gamma, which defines how curved the decision boundary is.

Random forest classifiers work by generating an ensemble of individual decision trees, each individual tree casts a vote for which class a particular observation should belong to, and the observation is assigned to the class with the most votes. Each decision tree attempts to optimally separate the data into classes by forming conditions on features that distinguish one class from another. Random forests have two hyper-parameters that can be tuned and optimised, ntrees, which is the number of decision trees that vote in the random forest, and mtry, which is the number of features that are considered at each decision point in a tree.

Logistical regression classifiers are simple classifiers that work by calculating the probability that a sample belongs to a particular class via some linear combination of the independent variables. A logistic function is then applied to this combination to determine class groups. Logistical regression has no tuning parameters and is exclusively a binary classifier.

## IV. RESULTS

A. Pre-processing

The training set contained 445 missing data points, and both methods were successful in estimating all 445 without issue. The same was seen for the validation and test sets. The ability of mean imputation and kNN imputation to accurately estimate the values of missing data was assessed by

manually removing known data from the training set, applying each method, and then comparing the estimates from each method with the known values of the insulin column with root mean square error and percentage bias calculations. The percentage bias of mean imputation was found to be 55.4% with a RMSE of 137.0, the percentage bias of kNN imputation was found to be 34.6% with a RMSE of 122.2. These results can be seen below.

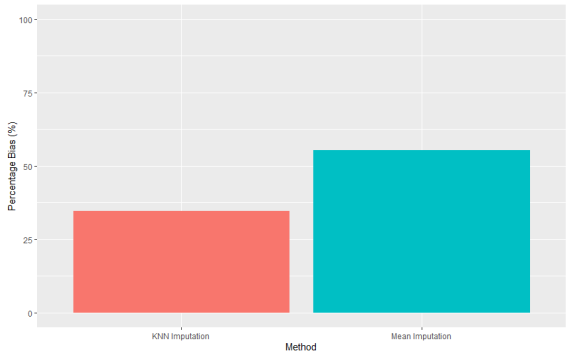


*Figure 2: Percentage Bias for Mean and kNN Imputation*

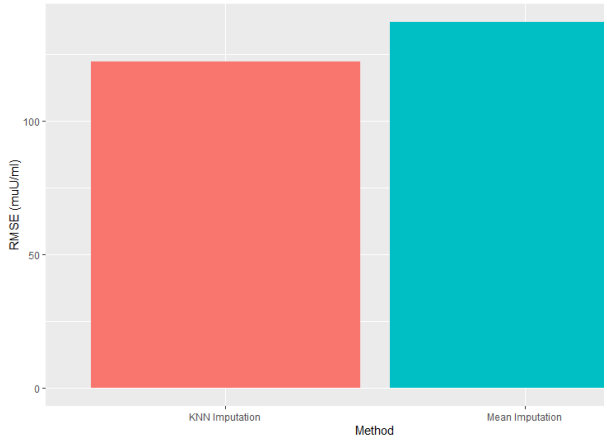


*Figure 3: RMSE (muU/ml) for Mean and kNN Imputation*

A figure showing the observed vs predicted value kNN imputation can be seen below.

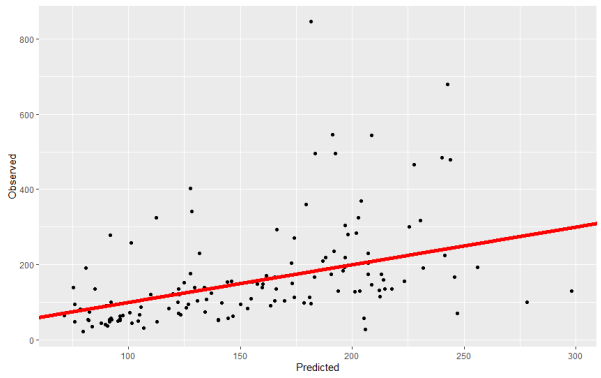


*Figure 4: Predicted vs Observed Values - Insulin*

Each method of scaling bound the data as expected, scatterplots showing how insulin and glucose data values changed with scaling are below.


*Figure 5: Glucose (mg/dL) vs Insulin (mu U/ml) - Raw Scale*


*Figure 6: Glucose (mg/dL) vs Insulin (mu U/ml) - Min-Max Normalisation*

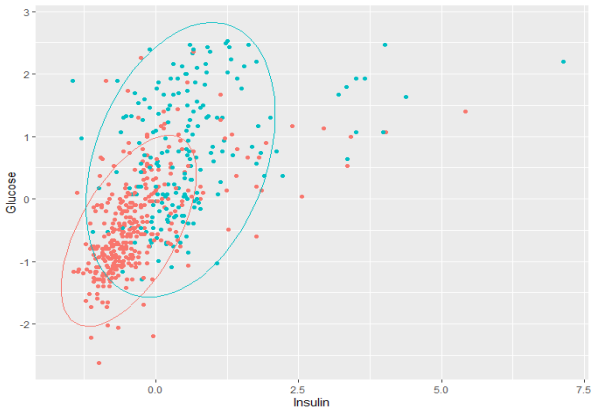


*Figure 7: Glucose (mg/dL) vs Insulin (mu U/ml) - Z-Score Standardisation*

B. Data Analysis

Data analysis techniques were used extensively throughout the project. The main analysis approaches consisted of visualisation and statistical methods that were useful when critically evaluating and selecting different pre-processing techniques and classification models. Exploratory data analysis also allowed insights to be made from the data to help answer the proposed research questions. One of the most useful analysis techniques was a heatmap that showed the correlation between all the dataset features and diabetes.

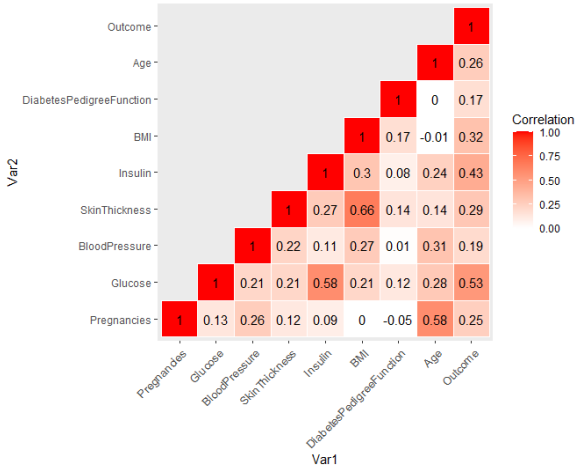


*Figure 8: Correlation Heatmap of All Variables*

Another useful technique was boxplots of each feature, showing the summary of all variables in one image, grouped into diabetic and non-diabetic individuals.

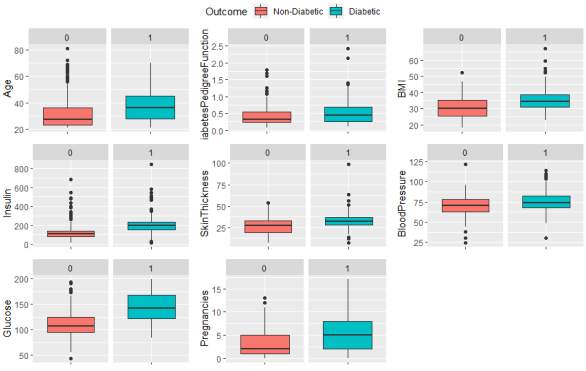


*Figure 9: Boxplots summarising each variable's core characteristics - split by class*

A graphical depiction of the missing data in the dataset was created.

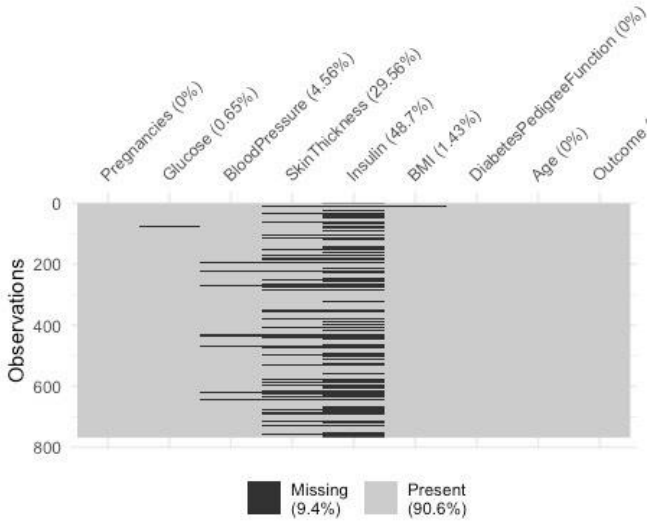


*Figure 10: Visualisation of Missing Data*

C. Classification

The accuracies, sensitivities and specificities of the three models tested can be seen below. These results were obtained by testing the models on the same test set, using kNN imputation, min-max normalisation and using all features.

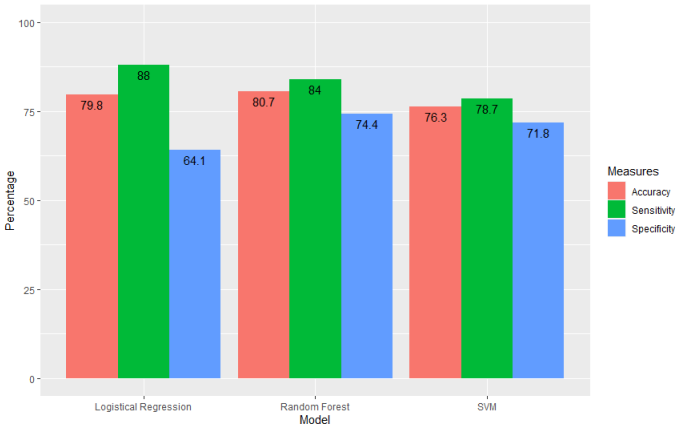


*Figure 11: Accuracy, Sensitivity and Specificity for Three Models*

For comparisons, a logistical regression was also done using mean imputation, Z-score standardisation and using only two features extracted using PCA.

## V. DISCUSSION

A. Pre-processing

Based on figures 2 and 3, it is evident that kNN imputation performed significantly better than mean imputation. kNN yielded a result 34.6% bias from the true value (or 137 muU/ml) on average, compared to 55.4% for mean imputation (122.2 muU/ml) based on analysing the variable with the

most missing data (insulin). This is likely because kNN imputation is multivariate and considers the similarity of all the independent features when calculating a replacement value, whereas mean imputation does not consider the effect of other variables on the missing value. Mean imputation will be heavily affected by extreme values, because it has no way of estimating a value significantly different to the mean. Additionally, the kNN imputation package in R required the data to be standardised before it took place, which may have reduced the effect of outliers on the imputations. The bias in both cases was negative, indicating that imputation is systematically underestimating the variance in the true values of the missing data. Other studies have proposed that this is because single imputation methods weight each missing value equally to known values in the analysis. When dealing with large amounts of missing data, this will inevitably result in a misleading dataset and analysis. While kNN was a substantial improvement over mean imputation, both methods are very inaccurate at estimating true values of missing data, so alternate approaches need to be considered in the future.

When examining how other studies have handled missing data in similar circumstances, many researchers have had more success with multiple imputation methods than single ones. Multiple imputation involves performing imputation on thousands of simulated datasets and determining an optimal estimate by combining the results. Other studies looked at during the literature review have achieved biases of under 10% with 40-60% missing data using multiple imputation methods on similar datasets, which is a large improvement over these results. A multiple imputation method such as the popular multiple imputation by chained equations (MICE) method would be a potential way to remove bias from the classification modelling accuracy and is a promising proposal for future work. The only downside to multiple imputation methods is the computational complexity required to carry one out. Alternatively, other researchers decided to simply remove any rows with missing data, however, in this instance removing incomplete rows is not viable as 376/768 rows would have to be deleted, or 49%, which is an unacceptable loss of data and would diminish the validity of the classification models. Other studies adopted an approach of removing entire variables that have a very high percentage of missing data. In cases where a relatively unimportant feature has a large amount of missing data this may improve the reliability of modelling, however, in this dataset the feature with the most missing data is the second highest correlated feature to diabetes. Removing this would likely have a detrimental effect to the classification accuracy.

As shown in figures 5,6,7, the scaling in both methods have been implemented fully and as expected. The data shown in the graphs is insulin against glucose, grouped by diabetes outcome. Insulin and glucose are the most predictive features for a diagnosis of diabetes, so two distinct groups can be clearly seen. The normalised data has been scaled to between 0 and 1 on both axes, whereas the standardised data has been approximately scaled to between -2.5 and +2.5 on the y-axis and -1.5 and +7 on the x-axis. The standardised data has a mean value of 0 and standard deviation of 1 for every variable. As all the features are scaled to the same range in normalisation, algorithms that use distance metrics (SVM, kNN etc) will weight each feature equally, which should improve their performance. The trade off to this is that normalisation is very sensitive to outliers as the relative distance between each point is preserved during scaling. Conversely, standardisation should result in slightly more biased results when using distance-based algorithms because the scales for each feature are not the same, so certain features will be favoured in the modelling algorithm. Although, it will still be a very large improvement over the raw data (which has a scale of 14-846 on the x-axis and 44-199 in the y-axis) as the scales are still all within the same order of magnitude. Additionally, standardisation is slightly less sensitive to extreme values than normalisation and is better for data that follows a Gaussian distribution. As can be seen in the figures, the shape of the data distribution is unchanged during scaling. Overall, because the scales are in the same order or magnitude after scaling with both methods, and standardisation is only slightly more robust against large variance than normalisation, it is likely that both methods will be adequate for modelling and any variance in accuracy due to scaling should be minimal.

Other studies examined during the literature review have taken alternate steps to obtain narrow scales that reflect the majority of a given dataset. Many studies use either Z score standardisation or min max normalisation but take additional measures to remove outliers before scaling is done to give an even more refined scale with smaller variance. This is generally done by setting some sort of cut-off boundary where outliers are removed, for example, all values two standard deviations from the mean may be excluded from the dataset. No attempt was made to eliminate outliers in this study because it was assumed that all logical, non-missing data points were genuine and thus accurately reflected the population of Pima Indians. There was no justification for removing extreme values in this case.

B. Data Analysis

Comparing techniques used in the data analysis process can be difficult as each technique has its own strengths and weaknesses and is generally used to accomplish a different task. For example, the creation of a correlation heatmap was instrumental to answering the research questions posed at the start of the study. Two of the main research questions were, do any of the independent variables in the dataset correlate to a diagnosis of type II diabetes, and if so, which correlate most strongly? The correlation heatmap allowed the answers to these questions to be easily visualised in a single figure. It is clear from the heatmap that all 8 independent variables are at least weakly associated with type II diabetes, and the strongest correlated variables are glucose, insulin and BMI with correlations of 0.53, 0.43 and 0.32 respectively.

Furthermore, data analysis techniques allow a comprehensive breakdown or summary of the dataset and its core elements to be displayed in a single figure. The boxplots figure shows the mean, median, IQR, minimum value, maximum value and all the outliers for each variable, grouped by diabetic and non-diabetic individuals. Using this, relationships between a variable and diabetes, data dispersion and detailed summaries of each variable can be ascertained visually very quickly, making boxplots an invaluable tool in data analytics. Additionally, the scatter plots show that, in general, the insulin and glucose measurements for non-diabetics are lower and more clustered than those with diabetes.

Analysis techniques can also be used to evaluate and select optimal pre-processing and classification techniques, both visually and statistically. For example, accuracy, sensitivity and specificity are invaluable for comparing the performance of classifiers and can give insight into why certain models have underperformed. Calculating the percentage bias and root mean squared error for each imputation technique clearly demonstrated that kNN imputation was the optimal choice in this case, which is not so easy to demonstrate without these analysis techniques.

In other literature, researchers have used an abundance of different techniques to visualise and critically analyse data. One of the most common techniques is to use histograms to visualise the distribution of data for each variable. This gives insight into which scaling method should be used, as a z-score scaling is a good fit for data that fits a gaussian distribution. Additionally, many researchers use missing data graphs to assess how much data is missing from each variable, which aids in selecting an appropriate method for handling the missing data, depending on the quantity and distribution.

C. Classification

As seen in figure 11, the accuracy, sensitivity and specificity of three classification techniques have been measured. Each model was trained, validated and tested on the same datasets, using the same pre-processing and evaluation methods to ensure a fair comparison. Logistical regression had an accuracy of 79.8%, indicating that it successfully classified 91/114 individuals. The sensitivity and specificity were 88.0% and 64.1%, respectively. A sensitivity of 88.0% indicates that 88% of the individuals in the dataset with diabetes were given the diabetic classification, so logistical regression was highly proficient at detecting people with diabetes. A specificity of 64.1% indicates that 64.1% of people without diabetes were given a non-diabetic classification. This is the lowest specificity for all three models, and it is the most likely model to give false positive classifications. Logistical regression was the best model for identifying genuine cases of diabetes, but also the worst model for identifying non-diabetic individuals, which suggests that logistical regression is over keen to classify an individual as diabetic compared to the other models.

The random forest model had an accuracy of 80.7%, a sensitivity of 84.0% and a specificity of 74.4%. Random forest correctly classified 92/114 cases and was the best performing model of the three, but only by a very small margin of 0.9%, or 1 case, which is not very statistically significant. Random forest was slightly less sensitive than logistical regression but had a significantly higher specificity. This model was proficient at identifying both genuine diabetes and people without diabetes and separating between the two.

SVM resulted in an accuracy of 76.3%, with a sensitivity of 78.7% and a specificity of 71.8%. 87/114 samples were correctly classified. SVM was the worst performing model that was tried, however, the difference between the best and worse performing models was only 5 incorrect classifications. To generate results more statistically significant and increase the reliability of the models, they should be tested on much larger datasets. SVM had a much worse sensitivity than logistical regression but had a better specificity. In contrast, SVM was outclassed in all aspects by random forest.

A reason why random forest may have performed better than the other two models is that random forest is not sensitive to outliers, unlike SVM and logistical regression, and outliers were not handled in the dataset.

Comparing these results to models trained on similar datasets looked at in the literature review, four SVM models yield accuracies of 83.2%, 78.0%, 85.7%, 65.1%. The value of 65.1% was obtained

with no pre-processing applied, which demonstrates how essential pre-processing steps are in the modelling pipeline to achieve reasonable performance. The accuracy of 78.0% was achieved using similar pre-processing steps used in this paper, normalisation and mean imputation. Which is in line with the accuracy achieved in this study and agrees with the results. Interestingly, a higher accuracy has been achieved using mean imputation than with kNN imputation. This could be for a plethora of reasons, including data leakage, optimal tuning parameters and specific model fit for that test dataset. Higher accuracies have been achieved in literature using more advanced pre-processing techniques. A value of 83.2% was obtained using the previously mentioned MICE imputation method and feature selection, which has led to a strong model performance. 85.7% accuracy has been achieved using feature selection to reduce the dataset features to the most correlated ones. This is the highest accuracy seen using SVM, therefore feature selection is a technique that should be studied further to improve modelling accuracy.

In literature, random forest was modelled on a similar dataset using normalisation and mean imputation. This yields an accuracy of 77.3%, slightly less than the result obtained in this paper of 80.7%. This difference may be because kNN imputation is more accurate the mean imputation or could be due to tuning parameters for random forest, but this is just speculation. Overall, the literature reviewed for random forest conform to the results of this paper.

Logistical regression in the literature had accuracies of 73.3% and 77%, using MICE imputation/standard scaling/feature selection and mean imputation/normalisation respectively. These approximately conform to the results obtained in this paper, and any difference can be explained in a similar way to the models above. Using min-max scaling, one study reported a logistical regression accuracy of 94.0%, which is substantially superior to other literature results and the results from this study. This may be a case of overfitting, or a heavily bias test set full of extreme values, as no complex pre-processing techniques were applied that can explain this disparity.

Finally, many more complex modelling techniques have been considered in the literature. Small area estimation (SAE) combined with convolutional neural networks seemed to yield the highest average accuracies, with the highest being 92.3% and none being under 85.0%. This would be a promising avenue for feature research.

## VI. CONCLUSION

The aim of this project was to use a variety of pre-processing, data analysis and modelling techniques to answer three key research questions. Do the independent variables predict a diagnosis of diabetes? If so, which variables predict diabetes most strongly? And can models be created that predict if an individual in the Pima Indian dataset has diabetes. The first two research questions were answered using data analysis to create a heatmap of correlations between each variable and the target variable. By doing this, it was clear that all the independent variables in the dataset were at least weakly associated with diabetes, and the strongest predictors were glucose, insulin and BMI with correlations of 0.53,0.43 and 0.32, respectively. To develop models, three different modelling techniques (SVM, logistical regression, random forest) were tested, as well as two methods of data imputation (mean and kNN) and two methods of scaling (normalisation and standardisation). It was determined that the optimal imputation method was kNN imputation (34.6% bias vs 55.4% for mean), the optimal scaling method was normalisation, as it put all variables on the same scale. Using these techniques, the highest accuracy model trained was random forest, with an accuracy of 80.7%, sensitivity of 84.0% and specificity of 74.4%. This was the most accurate model achieved and satisfied the third research question.

A shortcoming of the work is that each model was not trained using mean imputation or standard scaling, while keeping all other metrics constant. It would be useful to see the difference between each model when each pre-processing technique is independently applied, and then cross reference the results to evaluate how much of an impact each technique had on model accuracy. Another shortcoming of the work is the dataset size. As the training set was quite small, percentage differences in accuracy represented only a few data points classified differently. To make the models more statistically significant and reliable, a larger testing set should be used. Additionally, the outliers in the data were not mitigated in pre-processing, which will have biased model accuracy.

Considering the shortcomings, future work would be to repeat each model training process while changing each pre-processing step, while keeping all other metrics constant, which would provide insight into the effect of different techniques on model accuracy. Additionally, more complex pre-processing techniques and models, such as neural networks, SAEs, MICE imputation and robust scaling should be tried to see how much, if at all, they improve model accuracy.

**References**

[1]"Type 2 diabetes", *nhs.uk*, 2022. [Online]. Available: https://www.nhs.uk/conditions/type-2-diabetes/. [Accessed: 09- May- 2022].

[2]"Diabetes", *Who.int*, 2022. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes. [Accessed: 09- May- 2022].

[3]"One in 10 over 40s has type 2 diabetes, charity reveals", *Diabetes*, 2022. [Online]. Available: https://www.diabetes.co.uk/news/2019/feb/one-in-10-over-40s-has-type-2-diabetes,-charity-reveals-96644824.html#:~:text=New%20figures%20from%20charity%20Diabetes,to%20be%20type%202%20diabetes. [Accessed: 09- May- 2022].

[4]"Diabetes - Symptoms and causes", *Mayo Clinic*, 2022. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/diabetes/symptoms-causes/syc-20371444. [Accessed: 09- May- 2022].

[5]"Diet and lifestyle | Information for the public | Type 2 diabetes in adults: management | Guidance | NICE", *Nice.org.uk*, 2022. [Online]. Available: https://www.nice.org.uk/guidance/ng28/ifp/chapter/diet-and-lifestyle. [Accessed: 09- May- 2022].

[6]"Artificial intelligence in diagnostics | Healthcare Transformers", *HEALTHCARE TRANSFORMERS*, 2022. [Online]. Available: https://healthcaretransformers.com/digital-health/artificial-intelligence-diagnostics/#:~:text=AI%20can%20look%20at%20vast,including%20variations%20that%20humans%20cannot.&text=This%20may%20not%20only%20improve,costs%20by%20more%20than%2050%25. [Accessed: 09- May- 2022].

[7] *Core.ac.uk*, 2022. [Online]. Available: https://core.ac.uk/download/pdf/234670023.pdf. [Accessed: 09- May- 2022].

[8]"Genetics of Diabetes | ADA", *Diabetes.org*, 2022. [Online]. Available: https://www.diabetes.org/diabetes/genetics-diabetes#:~:text=Researchers%20are%20learning%20how%20to,your%20child's%20risk%20is%20higher. [Accessed: 09- May- 2022].