# Comparing Classifier Performances on Pima Indian Diabetes Patients

SIDs: 490402270 and 500485110

## Aim

The aim of this study is to assess the relative performance of self-determined classifiers and those implemented via Weka, in particular the  k-Nearest Neighbour algorithm and Naive Bayes algorithm. In addition to this, these classifiers and some additional classifiers (including ZeroR, 1R, DT, MLP, SVM and RF) will be compared with and without feature selection, to see how this impacts generalisation and predictive performance. Altogether, these tests can depict some factors that influence classifier performance and which classifier is better suited to this dataset.  Knowledge of feature selection will potentially improve the reader's ability to work with larger datasets and can be translated to improve a number of other machine learning techniques that benefit from regularisation. From an academic perspective, this information is important as our findings can provide insight into factors that influence predictive performance and the advantages of certain algorithms. This can then be taken and applied to future projects for better results.

Additionally, it showcases the ability of AI to add value and be applied to a range of industries and scenarios, in this case medicine and the health sector, and how additional applications could be developed, potentially even to the point of being able to diagnose health conditions. This study has importance from a health point of view, as the ability to accurately determine which patients of Pima heritage individuals could have the potential for diabetes within that community. Diabetes is the 7th leading cause of death in the US (CDC, 2021) and is a disease that will affect a person's blood sugar levels which can create long-last health conditions including kidney disease and heart diseases. With no cure available, being able to identify the potential of the onset of diabetes for an individual can be life changing and demonstrates the significance of collecting such data to train an AI algorithm to determine a diagnosis. If expanded and the techniques used are refined and developed further the learning from this study could be applied to diabetics of all heritages and provide life-changing diagnosis for individuals.

# Data

The dataset showcases the medical information of patients with Pima Indian heritage and according to the metadata, was originally sourced from the National Institute of Diabetes and Digestive and Kidney Diseases in May 1990. Pima Indians are a group of Native Americans who traditionally lived along the Gila and Salt rivers in Arizona. There are 768 patients in the datasets, all of whom are at least 21 years of age and female, and 8 numeric attributes that consist of personal characteristics and test measurements. 268 of these patients have diabetes, while 500 do not.

**Attributes of data (with given data name)**

1. Number of times pregnant (pregnant)
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test (plasma_glucose_conc)
3. Diastolic blood pressure (mm Hg) (blood_pressure)
4. Triceps skin fold thickness (mm) (tricep_skin_fold)
5. 2-Hour serum insulin (mu U/ml) (serum_insulin)
6. Body mass index (weight in kg/(height in m)^2) (BMI)
7. Diabetes pedigree function (pedigree_function)
8. Age (years) (Age)

For the response variable, there are two classes, ‘yes’ for if the patient has diabetes and ‘no’ if they do not. Provided with the data is a separate file with the metadata, and descriptors of each variable. Some slight modifications were made on this data for the purposes of this study. All missing values were replaced with averages for that feature, while the class was modified to be nominal values. Furthermore, the values were normalised to make sure they are in the range [0,1], with the class value remaining unchanged. Normalisation of our data ensures a common scale since different attributes are measured on different scales. Figure 1 showcases the key information for each variable (without normalisation).

**Figure 1**: Description and Statistics for Pima Indians Dataset

| **Variable** | **Description** | **Mean** | **Std. Dev.** |
|---|---|---|---|
| pregnant | Number of times pregnant | 3.8 | 3.4 |
| plasma_glucose_conc | Plasma glucose concentration a 2 hours in an oral glucose tolerance test | 121.7 | 30.4 |
| blood_pressure | Diastolic blood pressure (mm Hg) | 72.4 | 12.1 |
| tricep_skin_fold | Triceps skin fold thickness (mm) | 29.1 | 8.8 |
| serum_insulin | 2-Hour serum insulin (mu U/ml) | 155.3 | 85.0 |
| BMI | Body mass index<br>(weight in kg/(height in m)^2) | 32.5 | 6.9 |
| pedigree_function | Diabetes pedigree function | 0.5 | 0.3 |
| Age | Age in years | 33.2 | 11.8 |

**Correlation-based Feature Selection (CFS)**

CFS is an algorithm that aims to select the best subset of features from a complete dataset in order to reduce complexity and improve model generalisation. The key idea or “central hypothesis is that good feature sets contain features that are highly correlated with the class, yet uncorrelated with each other” (Hall, 1999). To assess these variables a heuristic is developed using the following formula (Figure 2) to find the correlation between components in a test and the variable:

**Figure 2:** Heuristic for CFS

$$r_{zc} = \frac{k\overline{r_{zi}}}{\sqrt{k + k(k-1)\overline{r_{ii}}}},$$

In the formula, known as Pearson’s correlation coefficient, $r_{zc}$ is the correlation between the summed components and the outside variable, k is the number of components, $r_{zi}$ is the average of the correlations between the components and the outside variable, and $r_{ii}$ is the average inter-correlation between components (Hall, 1999).

The following features were selected and dropped by CFS:

**Retained:** plasma_glucose_conc, serum_insulin, BMI, pedigree_function and Age
**Dropped:** pregnant, blood_pressure and tricep_skin_fold

# Results and discussion

**Overall Performance of Classifiers in Weka**

Over the course of the study a range of classifiers were developed and tested with a range of results. For both sets of tests (with and without feature selection) ZeroR, implemented through Weka, was the clear worst performer. This was to be expected as it has no predictive power, simply basing its decisions on the most common class of the target attribute and ignoring the other attributes. It does however provide a useful baseline to judge the respective performance of other classifiers. The worst performing classifier with actual predictive power was the Weka k-Nearest Neighbour algorithm with no CFS, and where k = 1. It outperformed the baseline by just 2.7343%. In comparison, the overall best performer was the Weka Support Vector Machine classifier with CFS applied, which outperformed the baseline by 11.5885%. In both tests, SVM demonstrated the highest accuracy (Figure 3) and is a popular classification method for a reason. The classifier separates data into a higher dimensional space and finds the most optimal linear decision boundary.

**Figure 3:** Summarised Findings

| | Without FS | CFS |
|---|---|---|
| Worst performing classifier | Zero R<br>65.1042 % | Zero R<br>65.1042 % |
| Best performing classifier | SVM<br>76.3021 % | SVM<br>76.6927 % |

**Figure 4**: Effect of CFS on Performance of Weka Classifiers

| | **ZeroR** | **1R** | **1NN** | **5NN** | **NB** |
|---|---|---|---|---|---|
| **No FS** | 65.1042 % | 70.8333 % | 67.8385 % | 74.4792 % | 75.1302 % |
| **CFS** | 65.1042 % | 70.8333 % | 69.0104 % | 74.4792 % | 76.3021 % |

| | **DT** | **MLP** | **SVM** | **RF** |
|---|---|---|---|---|
| **No FS** | 71.7448 % | 75.3906 % | 76.3021 % | 74.8698 % |
| **CFS** | 73.3073 % | 75.7813 % | 76.6927 % | 75.9115 % |

**Figure 5**: Effect of CFS on Performance of Self-Implemented Classifiers

| | **My1NN** | **My5NN** | **MyNB** |
|---|---|---|---|
| **No feature selection** | 0.7512303485987696<br>75.1230% | 0.7512303485987696<br>75.1230% | 0.7526315789473685<br>75.2632% |
| **CFS** | 0.7656869446343132<br>76.5687% | 0.7656869446343132<br>76.5687% | 0.7643198906356801<br>76.4320% |

**Performance of Self Implemented Classifiers Relative to Weka Classifiers**

Altogether, when comparing the 1NN, 5NN and Naive Bayes classifiers, the self-implemented variation with stratified 10-fold cross validation all outperformed the Weka classifiers of the same type. For the 1NN classifier, the performance was significantly higher, improving accuracy +7.2845% and +7.5583% with and without feature selection. In contrast to this, for the 5NN classifier, the difference was +0.6438% without feature selection and +2.0895% with. Finally, for the Naive Bayes it was +0.1330% and +0.1299%. While it is hard to pinpoint the reasons behind this, a few factors can be considered.

Accuracy

The first of which is the numerous additional variables available in the Weka Classifiers, such as useKernelEstimator and numDecimalPlaces for Naive Bayes and distanceWeighting and batchSize for KNN. These variables allow the model to make several assumptions about the data, or change the model's implementation to improve performance. While they may not have been used in this case, the ability and tendency of automated packages to make these assumptions may impact the training process, and in turn the predictive performance. The numDecimalPlaces variable for Naive Bayes and KNN which was set at 2. Since the data has decimal places of 6 and the self implemented classifiers did not consider rounding the output values in the model, this could significantly have changed the decisions for the Weka models and thus worsened the accuracy.

Computation Time

Another point of difference that can also be considered however is the build time of the respective models. In Weka, the 1NN, 5NN and NB classifiers all had instant run times (0sec output), whereas the self implemented had run times of  72 sec and 69 sec for 1NN, 29 sec and 28 sec for 5NN and 0.41sec and 0.25sec for Naive Bayes (Figure 6 and 7). In the real world, runtime is an important consideration as in many cases models are required to make quick decisions, and in various scenarios stakeholders may be willing to sacrifice predictive accuracy. Furthermore, in the dataset provided there were just 768 patients but in some professional databases there could be many, many times more than this, so if a certain model takes a long time to build and run it may not be viable when the size of data increases and more features are included. As such, for 5NN and NB where the accuracy improvements were sometimes close to just 0.1%, the increased speed of the Weka Classifier may make it the preferred choice.

**Figure 6:** Runtimes of classifiers in Weka

| | ZeroR | OneR | 1NN | 5NN | NB | DT | MLP | SMO | RF |
|---|---|---|---|---|---|---|---|---|---|
| Without FS | 0 | 0.01 | 0 | 0 | 0 | 0.01 | 0.35 | 0.02 | 0.18 |
| CFS | 0 | 0 | 0 | 0 | 0 | 0.03 | 0.29 | 0.03 | 0.27 |

**Figure 7:** Runtime of Self Implemented Classifiers

| | 1NN | 5NN | NB |
|---|---|---|---|
| Without FS | 72.43652510643005 | 29.381059885025024 | 0.408611536026001 |
| CFS | 68.79737544059753 | 28.33295226097107 | 0.2470080852508545 |

**Effect of Feature Selection Analysis**

There are a range of benefits to applying feature selection to a machine learning and AI problem. When applied correctly, it can improve predictive performance, increase the speed of model training and prediction, and improve the interpretability of the model by highlighting key features. As shown above the features that were selected by the CFS were plasma_glucose_conc, serum_insulin, BMI, pedigree_function and Age. The removed variables were pregnant, blood_pressure and tricep_skin_fold. This selection mostly make intuitive sense, as according to a 2020 article by the Mayo Clinic, a non profit  medical organisation, while key risk factors for type 1 diabetes include family history, environmental factors and geography (certain countries and areas have higher rates), and for type 2 diabetes these factors are also relevant, but other prominent risk factors are weight, age, and high blood pressure.

Feature Selection

The features that were kept can make intuitive sense, especially when discussing medical research on diabetes and the factors that increase its risk.

BMI, which is a figure that uses height to moderate a healthy body mass level (taller people naturally weigh more) is the variable that informs the model regarding weight and a greater increase in BMI or being obese can increase the risk of developing type 2 diabetes.The plasma glucose concentration and serum insulin features were also retained during feature selection, as these variables are directly related to the symptoms of type 2 diabetes, in which “the body becomes resistant to the normal effects of insulin and gradually loses the capacity to produce enough insulin in the pancreas” (Diabetes Australia, 2021). An oral glucose tolerance test (OGTT) demonstrates how well your body handles sugar from foods and an increased plasma glucose concentration has been associated with increased risks for type 2 diabetes in the future (Abdul-Ghani & DeFronzo, 2009). The function of insulin is to manage blood glucose levels so variables that measure both the presence of insulin and glucose in the blood make sense to include in classifiers that are aiming to predict the diabetes in patients. The pedigree function provides “data on diabetes mellitus history in relatives and the genetic relationship of those relatives to the patient” and is “a measure of genetic influence” (Brownlee, 2019) that would be included in the model for its relationship to diabetes and its history for an individual.

Regarding the features that were dropped, the tricep skinfold thickness test is another factor that assesses nutritional status and fat levels in patients, however it is “prone to large variations, both within and between observers” due to a variety of factors (Ayling, 2014). Despite its correlation to BMI, it is less influential on the classification of diabetes, hence why it can be dropped.

Blood_pressure was a surprising feature to be disregarded, however the presence of BMI, Age and genetic history (via the pedigree function) in the classifier, all factors that can influence high blood pressure, may have resulted in that information being redundant. The final factor that was dropped was the number of times the patient had been pregnant. According to the CDC in America "every year, 2% to 10% of pregnancies in the United States are affected by gestational diabetes". The likelihood is that not enough of the 768 patients in the dataset were impacted by this to make it a significant factor in the classifier and this would only predominantly affect females which wouldn't make up all the individuals.

### Accuracy

Based on the results generated, CFS was able to maintain or increase the classifier accuracy for all self-implemented and Weka algorithms. A possible reason for this is that it reduces overfitting on the training data, which occurs when "a statistical model fits exactly against its training data. When this happens, the algorithm unfortunately cannot perform accurately against unseen data, defeating its purpose" (IBM Cloud Education, 2021). By reducing the number of features and reducing the likelihood of overfitting, we have a more robust algorithm, the model would better deal with random or unexpected values in the test set.  Greater accuracy is beneficial for stakeholders as they can more greatly rely on the decisions created by algorithms. With Pima individuals in mind, the greater certainty is important for medical diagnosis as this will eventually impact the lives of individuals and incorrect diagnosis could have a drastic effect on their physical wellbeing.

### Computational time

Additionally, the training time was reduced for all models, not just between the self-implemented models and Weka models as mentioned earlier. This would be a direct result of the model having less data to train on and thus a quicker training for the algorithms. With the medical usage of the data, by reducing the computational costs and increasing the efficiency, there is a greater potential for earlier diagnosis and more individuals to be diagnosed, creating a beneficial outcome for all Pima individuals.

Feature selection is also important for classifiers such as Naive Bayes since correlated attributes reduce the power of Naive Bayes. By utilising feature selection, we are able to discard correlated attributes and there won't be a violation of the independence assumption.

# Conclusion

In conclusion, the results of the study show that the self-implemented classifiers outperformed the ones implemented via Weka, with the caveat that they were much slower to build and predict. This was particularly evident for the kNN algorithms, as the difference in speed was significant, meaning if the dataset was sufficiently large, the Weka version of the classifier would probably be more useful due to the minimal difference in performance. Furthermore, the study revealed that overall, feature selection via CFS improved predictive accuracy and computational time, as all the classifiers implemented either had the same accuracy or better. Throughout the study, greater insight into the key features affecting diabetes were developed, as well as what influences classifier performance. Hopefully this will lead to further study and development into the ability of AI and machine learning to positively influence the health sector.

In terms of future areas of study to further develop classifier performance, there are a range of other types of feature selection that can be used, such as feature importance measurement which is available when using decision tree-based classifiers. Additionally, L1 and L2 regularisation, often used in regression are tools that can reduce overfitting and help with model generalisation. In terms of actuarial classifiers that could be used in future studies, gradient boosted machines, and their derivatives such as XGBoost have been shown to be highly efficient with high levels of performance. The final suggestion for future study is a simple one but a pillar within machine learning and AI in that more data could be used. More data means a greater ability to understand the key factors and patterns within a problem, and such could improve classifier performance and overall understanding of diabetes and potentially key characteristics within the Pima Indian community.

# Reflection

After some discussion and reflection on the assignment, we feel that the key piece of learning from this study was the importance of feature selection when trying to improve model generalisation. Throughout the study of machine learning and classification and regression at university, there is a large focus on more data being better (assuming it is clean). On top of this, feature engineering is highlighted as a step in which more information can be extracted from data and put into the model to improve accuracy. To see that a reduction in complexity improves results is a great reminder that more is not always better, and that overfitting remains a core issue when working with data. This is also prevalent in the workforce, in which some colleagues may not have the same level of expertise when it comes to these processes, so speed of prediction, and greater interpretability are potentially key deliverables as well, not just how well the model performs.

Another key learning was around the breaking down and understanding the factors of success or failure when it comes to model performance and to not just rely on technology to build models for us. Even though Weka was presented as a viable and efficient solution to building classifier models, it came to our attention that our own models were more accurate, but not as efficient. Often projects are results focused aiming to create the best model and we are often under the assumption that technology can build better models, but by comparing model performance based on a range of different factors, in this case self-implemented and using a machine learning software, it provides a greater understanding for future work. In prior assignments, this testing process is limited to which model performs the best and limited effort or understanding is developed as to why. By building this habit and focusing on the background and theory, hopefully future work will also be focused on what we can potentially create ourselves.

# References

Hall, M. A. (1999, April). Correlation-based Feature Selection for Machine Learning. University of Waikato. Retrieved May 2022, from https://www.cs.waikato.ac.nz/~mhall/thesis.pdf

Mayo Clinic. (2020, October 30). Diabetes - Symptoms and causes. Retrieved May 2022, from https://www.mayoclinic.org/diseases-conditions/diabetes/symptoms-causes/syc-20371444#:%7E:text=Abnormal%20cholesterol%20and%20triglyceride%20levels,risk%20of%20type%202%20diabetes.

Diabetes Australia. (n.d.). Type 2 Diabetes. Retrieved May 2022, from https://www.diabetesaustralia.com.au/about-diabetes/type-2-diabetes/

Brownlee, J. (2019, August 22). Case Study: Predicting the Onset of Diabetes Within Five Years (part 1 of 3). Machine Learning Mastery. Retrieved May 2022, from https://machinelearningmastery.com/case-study-predicting-the-onset-of-diabetes-within-five-years-part-1-of-3/

Ayling, R. M. (2014). Skinfold Thickness. Science Direct. Retrieved May 2022, from https://www.sciencedirect.com/topics/agricultural-and-biological-sciences/skinfold-thickness

CDC. (2022, March 2). Gestational Diabetes. Centers for Disease Control and Prevention. Retrieved May 2022, from https://www.cdc.gov/diabetes/basics/gestational.html#:%7E:text=Gestational%20diabetes%20is%20a%20type,pregnancy%20and%20a%20healthy%20baby.

IBM Cloud Education. (2021, March 6). Overfitting. IBM. Retrieved 22–05, from https://www.ibm.com/cloud/learn/overfitting

National Institute of Diabetes and Digestive and Kidney Diseases (1990). *Pima Indians Diabetes Database*. [Data file] Retrieved from https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

Abdul-Ghani, M. A., & DeFronzo, R. A. (2009). Plasma glucose concentration and prediction of future risk of type 2 diabetes. Diabetes care, 32 Suppl 2(Suppl 2), S194–S198. https://doi.org/10.2337/dc09-S309