**COMP3308 Introduction to Artificial Intelligence**

***Assignment 2***

Comparative Analysis of Classification Algorithms

Achira Tantisuwannakul 530473598

Anabel Geraldine 520360707

Date: 06/05/2024

**Table of Contents**

1. Introduction………………………………………………………………………… 2

1.1 Aim of Study…………………………………………………………………… 2

1.2 Importance of Study……………………………….…………………………… 2

1. Data………………………………………………………………………………… 2

2.1 Dataset Description……………………..……………………………………… 2

2.1.1 Pima Indian Diabetes Dataset..…………..…………………………… 2

2.1.2 Room Occupancy Dataset……………..……………………………… 3

2.2 Dataset Differences and Similarities……………..……………………..……… 4

1. Results and Discussions……...……………..……………………………………… 4

3.1 Accuracy Results………………....……..……………………………………… 4

3.2 Classifier Performance and Comparison..……………………………………… 5

3.3 Implementation and Weka Comparison..……………………….……………… 6

3.4 Dataset Influence on Classifier Prediction..…………………….……………… 7

1. Conclusion………..………………………………………………………………… 8

4.1 Conclusion of findings…………………………………………………….…… 8

4.2 Future Work Suggestions……………………………….……………………… 8

1. Reflection…………………………………………………….…………..………… 9
2. **Introduction**

**1.1 Aim of Study**

The aim of study is to implement and compare different machine learning classifiers for predictive tasks, mostly focusing on K-Nearest Neighbors (KNN) and Naive Bayes algorithms. By evaluating these classifiers along with using 10-fold-cross-validation on Pima Indian Diabetes Dataset and Room Occupancy Dataset, this study aims to *provide insights and understanding to classifiers’ performance across different datasets.* Additionally, Weka, a widely-used machine learning tool, will be used to facilitate comparative analysis among various different classifiers, providing a comprehensive assessment of classifier performance. Understanding the effectiveness of different classifiers can guide future research and application in machine learning.

**1.2 Importance of Study**

As Artificial Intelligence increasingly depends on making efficient and accurate decisions, understanding the analysis of classifiers is *crucial for determining appropriate context in which to apply various decision-making techniques*. This research helps shed light on the performance of different classifiers across different predictive tasks and dataset. Through these evaluations, the study seeks to provide valuable insights into the strengths and limitations of each algorithm in handling real-world data. Moreover, Weka helps enhance the comprehensiveness of the analysis. Understanding the effectiveness of different classifiers can inform practitioners and researchers about the most suitable classifiers or models for specific applications, advancing the development and deployment of machine learning solutions in many different domains.

1. **Data**

**2.1 Dataset Description**

**2.1.1 Pima Indian Diabetes Dataset**

The Pima dataset is used to investigate whether various health related factors can predict the likelihood of an individual having diabetes. This dataset consists of 768 instances, each representing a patent of Pima Indian heritage. Each instance is described by 8 numeric attributes and the target variable denoted as “class”.

| **Column** | **Description** |
| --- | --- |
| tp | Number of times pregnant. |
| gc | Plasma glucose concentration a 2 hours in an oral glucose tolerance test |
| bp | Diastolic blood pressure (mm Hg). |
| sft | Triceps skin fold thickness (mm). |
| si | 2-Hour serum insulin (mu U/ml). |
| bmi | Body mass index (weight in kg/(height in m)^2. |
| dpf | Diabetes pedigree function. |
| age | Age(years). |
| class | Whether the individual tests positive for diabetes (yes) or not (no). |

***Figure 1. Pima Dataset variable description***

This dataset serves as a valuable resource for exploring the relationship between health metrics and presence of diabetes, contributing to the development of predictive models for identifying individuals at risk of the disease.

**2.1.2 Room Occupancy Dataset**

The dataset aims to explore whether sensor readings in a room, such as light, temperature, sound, and carbon dioxide levels, can predict the occupancy status of the room. This dataset comprises 2,025 instances, each representing a point of time between December 2017 and January 2018 when sensor measurements were taken.

| **Column** | **Description** |
| --- | --- |
| temp | Temperature reading (°C). |
| light | Light reading (Lux). |
| sound | Output of an amplifier attached to a microphone (Volts). |
| CO2 | Carbon dioxide reading (PPM). |
| class | Whether or not there were people in the room (no = no, yes = 1-3 people). |

***Figure 2. Occupancy Dataset variable description***

**2.2 Dataset Differences and Similarities**

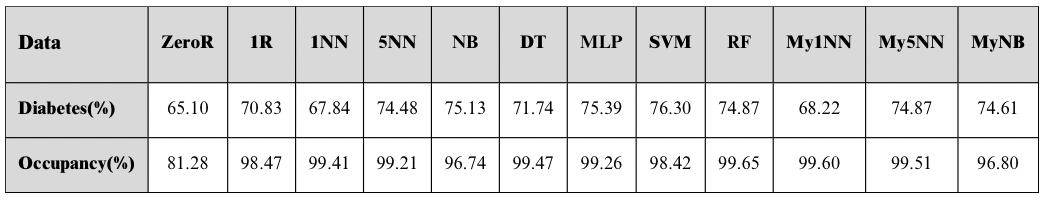
The Pima Indian Diabetes Dataset and the Room Occupancy Dataset are both valuable for predictive modeling but serve distinctly different purposes and each offers unique challenges and insights. Firstly, they have ***different purposes and applications***, while Pima dataset focuses on medical or healthcare research, Occupancy data is usually used in some sort of smart building management. Following that, they also have ***different data types***, Pima data comprises clinical measurements which are continuous numeric attributes and Occupancy data includes reading from physical sensors which are generally continuous and range from environmental to chemical properties. They also have ***different implications of prediction***, where pima dataset predictions have significant health implications and occupancy data’s predictions are generally operational.

However, they also have some similarities including the ***same format and structures***, they both consist of instances described by some attributes leading to a binary classification target (yes or no). They also have the ***same data handling needs*** such as normalization.

1. **Results and Discussions**

**3.1 Accuracy Results**

The table below presents the accuracy of each classifier, expressed as percentages rounded to two decimal places, and obtained through 10-fold cross-validation.



***Figure 3. Table displaying accuracy for each classifiers***

**3.2 Classifier Performance and Comparison**

Classifiers analyzed within this study included the following: zeroR, 1R, 1NN, 5NN, Naive Bayes, Decision Tree, Multi-layer Perceptron, Support Vector Machine and Random Forest.

ZeroR represents baseline performance as prediction is decided by picking the majority. Therefore, the chance of correction prediction is equal to the proportion of yes/no labels (depending on what is labeled as successes) in relation to total observations. KNN and Naive Bayes will also be analyzed, using the WEKA platform but also via native implementations. KNN will be tested using hyperparameters K=1 and K=5. DT, MLP, SVM and RF computations will all be performed via WEKA.

***pima.csv***

From the pima dataset, The ***support vector machine (SVM)*** using SMO optimization demonstrated the highest accuracy at 76.30%, excelling in handling linear separability between classes. It was followed closely by the ***Naive Bayes classifier***, which achieved 75.13% accuracy, showcasing its prowess in managing probabilistic relationships inherent in the dataset’s features. The ***MLP***, notable for its ability to model complex non-linear relationships, also performed well with accuracy of 75.39%, although it required the *longest training time at 0.43 seconds.* ***Random forest*** showed strong accuracy of 74.87%, but had slightly slower training speed compared to simpler models. The ***5NN*** improved on the performance of ***1NN***, indicating that averaging more neighbors helps mitigate the effect of noise, reaching an accuracy of 74.48%. The ***Decision tree*** provided a good balance between accuracy, 71.74%, and model interpretability. Despite its simplicity, ***1R*** has 70.38% accuracy by focusing on a single influential feature, useful for preliminary analysis. As mentioned, ***ZeroR***, served as a baseline with 65.10% accuracy, reflecting major class distribution and highlighting the dataset’s inherent class imbalance.

Overall, SVM and Naive Bayes excel due to their ability to manage the non-linear relationships and conditional dependencies in the data.

***occupancy.csv***

From the analysis of Room Occupancy dataset machine learning classifiers, we observed significant variation in performance. The ***Random Forest*** classifier emerged as the best performer with accuracy 99.65% with a relatively short training time of 0.12 seconds, this is due to their learning technique that effectively reduces overfitting and enhances generalization. Next best classifier for this data would be ***Decision tree*** which has high accuracy and benefits of simplicity and quick training. Nearest neighbor suggested that the dataset likely contains well-defined, compact clusters, with the ***1NN*** model showing high accuracy of 99.41% with an almost instant training. Conversely, the ***Multi-Layer Perceptron***, though accurate at 99.26%, required more computational resources, highlighting the trade-off between complexity and computational cost. The ***Naive Bayes*** model underperformed relative to others, suggesting potential feature dependencies that violate its assumptions. Simpler models like ***Support vector machine*** and ***1R*** also performed admirably, which suggests that they are effective.

The high performance of instance-based methods and random forest suggests that the dataset’s features allow for clear spatial separation and that the ensemble method effectively captures complex patterns without overfitting.

**3.3 Implementation and Weka Comparison**

***pima.csv***

The ***KNN*** classifiers, specifically, showed a notable improvement over WEKA results; Our 1NN implementation achieved accuracy 68.22%, which is 0.38% higher than WEKA’s, and our 5NN achieved 74.87%, compared to WEKA’s 74.48%. These enhancements may be attributed to specific approaches in calculations for tie breaks or distance calculation. Contrarily, our implementation of ***Naive Bayes*** classifier, at 74.61% accuracy, slightly underperformed compared to WEKA’s 75.13%. This might be caused by differences in algorithms for probability estimation, such as Laplace smoothing that WEKA incorporates.

***occupancy.csv***

In this dataset, our implementations of the KNN and Naive Bayes showed great performance, even surpassing WEKA’s result in certain instances. Our ***1NN*** and ***5NN*** classifiers outperformed WEKA’s, with accuracies of 99.60% and 99.51% respectively, compared to WEKA’s 99.41% and 99.21%. This suggests that our implementation may be more tuned for this dataset. Moreover, our ***Naive Bayes*** model slightly outdid WEKA’s, giving 96.80% accuracies, might be due to minor differences in preprocessing and specific numerical precision.

***Overall*,** it reveals that while our models are generally on par with WEKA’s, they occasionally outperform WEKA’s standard algorithms. This success might be linked to our tailored approaches to algorithmic configurations and preprocessing techniques that are specifically optimized for the nuances for each individual dataset. The slightest deviations in performance between or models and WEKA’s, whether as modest improvements or minor underperformance, highlight the impact of methodological variations in machine learning practices. These insights will guide the refinement of our models to enhance their accuracy and reliability.

**3.4 Dataset Influence on Classifier Prediction**

The variability in classifier performance between the Pima and Occupancy dataset can be caused by various factors. One of the factors is ***dataset characteristics***, Pima consist of medical data which are inherently complex due to non-linear relationships and a high degree of interdependence among features such as glucose levels, BMI, and age. This complexity necessitates more sophisticated modeling techniques to effectively capture and utilize the underlying pattern in the data. Occupancy data relies on environmental sensors which typically exhibit more straightforward and linear relationships, such as between light or CO2 levels and room occupancy. This simplicity allows for simpler models to achieve higher accuracy.

The ***size of the dataset*** also had noticeable effects on the average accuracy of the classifiers overall. The Occupancy dataset had less features than the Pima dataset, and the proportion of yes/no, as well as a big disparity between magnitudes of observations; Pima with 768 observations whilst Occupancy with 2,025. Training on more observations improves prediction accuracy, reducing the risk of overfitting and enhancing the model’s ability to generalize.

***Model selection*** ***and tuning*** are critical in adapting specific needs for each dataset. Models that assume feature independence, like Naive Bayes, may perform adequately on the Room Occupancy dataset but less so on the Pima dataset where features are interrelated. Adjusting hyperparameters like depth of decision trees or the k in kNN according to the dataset can significantly enhance model accuracy. Additionally, ***computational resources*** and efficiency of algorithms used can also affect performance. More sophisticated computationally intensive models might be feasible and beneficial for the occupancy dataset due to its simpler relationships and larger size, whereas the same models might overfit on the smaller, more complex Pima dataset.

All in all, for **more complex data** like Pima, they have characteristics such as non-linear relationships among features, high dimensionality or many interdependent variables, and potential class imbalance which requires more sophisticated modeling. According to our findings, classifiers like Random forests, Support Vector Machines, and Multi-layer perceptrons are recommended to be used. This is due to their ability to handle complex feature interactions and reduce overfitting .

For **less complex data** like Occupancy, they have characteristics such as being more straightforward, linear relationship between features, lower dimensionality with less feature independence, and features that directly measure the phenomenon of interest. Some recommended classifiers that are effective for this data type, according to our findings, are Decision tree, Naive bayes, and KNN as it provide easy to understand models and assume independence, they also works well where the output class can be determined by simple distance metrics in a feature space that isn’t overwhelmed by dimensionality.

1. **Conclusion**

**4.1 Conclusion of findings**

In conclusion, both WEKA classifiers and our implementations yielded a range of different accuracies, influenced by factors such as complexity, type of data, and volume of observations, which affects the success of each algorithm respectively. Some classifiers displayed greater performance than others, emphasizing how choosing appropriate classifiers for varying scenarios has a significant effect on prediction accuracy.

This is reflected within the Pima and Occupancy analysis results within the report, with Naive Bayes classifier outperforming the rest within Pima whilst DT and KNN classifiers were better performing within Occupancy. When discussing the performance of these classifiers within their medical contexts, we can assume that provided the appropriate classifier was chosen and applied to a situation wherein a patient was awaiting diagnosis, an application of WEKA implementations or our implementations would be able to provide a prediction with a guaranteed accuracy of 74%.

**4.2 Future Work Suggestions**

These findings could be furthered through the use of more datasets with varying properties operating within varying different contexts. The use of more advanced or nuanced techniques like deep learning neural network models for prediction and nested cross-validation or bootstrapping for assessing performance would allow for an extended study that would analyze in detail the appropriateness of various predictive models within different circumstances.

A future work that would complement these findings could be a domain-specific study diving into specific algorithmic enhancements and feature engineering geared for data problems within medicine; particular data contexts call for specialized data models that excel in their respective domains, outperforming more generalized models.

1. **Reflection**

In the world of Artificial intelligence, classifiers serve as fundamental tools for decision making and prediction tasks across diverse domains. As evidenced by this comparative analysis, the performance of classifiers can vary depending on a lot of different factors. This underscores the importance of understanding the strengths and limitations of different algorithms and selecting the most appropriate ones for specific applications.

Looking ahead, the world of AI holds immense promise but also presents significant challenges. There is still much work to be done to advance the capabilities of classifiers and AI systems more broadly, including the development of more sophisticated algorithms, improved model evaluation techniques, and robust ethical frameworks.