**Question 1 - Visualization and analysis of the Palmer penguin dataset**

|  |  |  |
| --- | --- | --- |
| **Table 1.** Attributes of the Palmer penguin dataset | | |
| **Attribute** | **Type** | **Values in the dataset** |
| species | categorial | Adelie, Chinstrap, Gentoo |
| island | categorial | Torgersen, Biscoe, Dream |
| bill length | numerical | 32.1mm - 59.6mm |
| bill depth | numerical | 13.1mm - 21.5mm |
| flipper length | numerical | 172mm - 231mm |
| body mass | numerical | 2700g - 6300g |
| sex | categorial | Male, Female |

The Palmer penguin dataset consists of 344 records of the physical attributes of three species of penguin living on three islands in Antarctica (Table 1) [ref]. In this report, consideration is given to data cleaning and preparation, the dataset is explored through visualization and analysis is carried out to compare the accuracy performances of a small number of AI approaches.

**Data cleaning and preparation - missing values, standardization and data imbalance**

|  |
| --- |
|  |
| **Figure 1**. All numerical features show a significant statistical difference between the male and female measurements, as seen in the *body mass* boxplot above. Shown are median values, Q1 and Q3 quartiles, as well as outliers that are outside the range Q1-1.5IQR to Q3+1.5IQR, where IQR=Q3-Q1. |
|

In the dataset, 11 records have missing values. Two of these records can be deleted immediately as they are missing values for all of the numerical attributes and the *sex* feature and any imputation is unlikely to be reliable. The remaining nine records have no value only for the *sex* attribute. As can be seen in Figure 1, the physical attributes of the male and female of each species are statistically different and so it is reasonable to consider assigning a *sex* to those records missing this attribute. Following standardization, a Shapiro-Wilk test was performed to confirm each numerical attribute exhibits a normal distribution [ref] and Z-tests were performed to assess separately both the hypothesis that the missing sex value is male and that it is female [ref]. It was found that two of the records could be imputed as male and three as female and these were then retained in the dataset. The remaining four records were removed from the dataset. The cleaned dataset consisted of 338 records made up of 147 Adelie penguins (74 male, 73 female), 68 Chinstrap penguins (34 male, 34 female) and 123 Gentoo penguins (62 male, 61 female).

A number of the methods applied in this work involve distance measures and so may be biased in favour of features with smaller standard deviations [ref]. This bias can be removed by standardizing the four numerical attributes independently (to give zero mean and unity standard deviation). Standardization uses only the statistics of training sets, but standardization is also applied to test sets. If a dataset is imbalanced, AI approaches may be biased in predicting classes that are more commonly found in the training data. The Palmer penguin dataset is somewhat imbalanced, with the number of Chinstrap records being around half of that of either Adelie or Gentoo, which are present in similar numbers. The importance of imbalance depends on the analysis method applied. It is known that all the methods adopted in the current work are generally little affected by imbalanced data [refs] and so no modifications were made to reduce imbalance.

**Visualization of the dataset**

Figure 2 shows the species distribution across the three islands in the study. Chinstrap and Gentoo penguins are found only on one island, so *island* is a potential confounding factor, possibly affecting physical characteristics due to environmental factors (such as predators or food supply). A Shapiro-Wilk test was used to confirm that the numerical features of the Adelie penguins (that are found on

|  |
| --- |
|  |
| **Figure 2**. Adelie penguin samples were from all three islands, but Gentoo and Chinstrap only from one |

all the islands) are normally distributed and an ANOVA test confirmed that their physical characteristics are not significantly influenced by the island inhabited. Consequently, it was considered unlikely that the *island* is a confounding factor in the dataset.

|  |
| --- |
|  |
|
| **Figure 3**. Pairwise distributions of numerical features. Gentoo can be distinguished from the other species, but Adelie and Chinstrap samples may not be completely separable from one another |

Pairwise scatterplots for the numerical features are shown in Figure 3. It can be seen that *bill depth* in combination with either *flipper length* or *body mass* yields a separable cluster of Gentoo penguins (shown in green) allowing them to be identified. No pairwise combination of numerical features completely separates Adelie (orange) from Chinstrap (purple) clusters, but good separation is provided in the distributions involving *bill length*, making this a candidate feature for distinguishing between these species.

|  |  |  |
| --- | --- | --- |
| **Table 2.** Metaparameters considered in training the methods. The values shown in bold are those that most consistently produced results of best accuracy during validation and so were selected for generating results. | | |
| **Method** | **Metaparameters** | **Values considered** |
| *k*nn | number of nearest neighbours *k* | **1**, 2, 3, 4, 5, 6, 8, 10 |
| weight function for prediction | **uniform,** distance |
| distance metric for computing neighbours | **Manhattan**, Euclidean |
| random forest | number of trees in the forest | 5, **10**, 15, 20, 25 |
| maximum depth of trees | **no maximum**, 10, 20 |
| minimum number of samples to split node | **2**, 5, 10 |
| minimum number of samples at leaf node  function to measure quality of split | **1**, 2, 4  **gini**, entropy |
| *k*-means | number of clusters *k* | 2, **3**, 4, 5, 6, 7, 8, 9, 10 |
| centroid initialization method | **k-means++**, random |
| number of runs with different centroid seeds | 2, **5**, 10, 20 |
| maximum number of iterations | 5, **10**, 20, 50 |
| CVA | regularization parameter (C) | 0.1, 1, **10**, 100 |
| kernel coefficient (gamma) | **1**, 0.1, 0.01, 0.001 |
| kernel type | rbf, **linear**, polynomial |

Figure 1 above shows there is a difference in the body masses of the male and female samples for each of the three species. Differences between the sexes for the other three numerical physical characteristics in the dataset were also apparent. Since narrower distributions are apparent if the sex of the species is considered rather than just the species itself, including sex is likely to provide a finer grained distinction for species classification and this knowledge can be used to improve performance, as discussed in the analysis section below.

**Implementation**

All code was written in Python 3.11 [ref] using ‘Scikit-Learn’ libraries [ref] running under Ubuntu Linux [ref]. The code is available in a repo [ref]. Predicting the penguin species from the given features is a classification problem. Results are obtained from a baseline method, two convention classification approaches, namely *k*-Nearest Neighbour (*k*nn) [ref] and random forest [ref], unsupervised *k*-means (following cluster labelling) [ref] and a novel combined visualization and analysis (CVA) approach that uses insights from visualizations combined with two-dimensional linear Support Vector Machines (SVMs) classification.

For all the methods implemented, 20% of the dataset was kept for a test set. To reduce the potential for overfitting, the classification methods (all but *k*-means) were trained using ‘holdout validation’, where the remaining 80% of the dataset was used in a five-fold cross-validation configuration [ref]. For all methods, the Scikit-Learn function GridSearchCV was employed to tune metaparameters [ref]. Table 2 shows the values selected for the metaparameter grid for each of the AI methods and those that gave best performance were selected to obtain the accuracy results from the test set. Scikit-Learn also provides a range of pseudo-random procedures for selecting validation and test set values and 100 of these were used both when selecting metaparameters and when generating accuracy results.

|  |  |
| --- | --- |
| **Table 3.** Mean classification accuracy from 100 test sets each generated by a pseudo random approach and using the parameters identified in Table 2 | |
| **Method** | **Accuracy** |
| baseline, most numerous species | 43.49% |
| *k*NN, all features | 99.24% |
| *k*NN, no island | 99.46% |
| random forest, all features | 98.57% |
| random forest, no island, flipper length or body mass | 98.59% |
| *k*-means, all numerical features | 97.06% |
| *k*-means, separate clusters for each sex | 98.23% |
| CVA using bill depth, flipper length, bill length | 98.56% |
| CVA using bill depth, flipper length, bill length, sex | 98.78% |

**Results and analysis**

A baseline is used to demonstrate performance improvements achieved by the AI methods being considered. If the performance cannot be improved significantly compared to the baseline, this may indicate that the method is not suitable or that the problem itself is particularly intractable. In classification, the baseline method is often simply to select the most frequent class in the observations and, in this work, this is the Adelie penguins, giving an accuracy of 43.49% (147/338).

***Classification method 1 -*** ***k*nn** The performance of *k*nn was found to be improved by omitting features from training. An exhaustive search involving omitting all combinations of features in turn determined that the best accuracy was obtained when *island* was omitted and this occurred when *k*=3. It appears that *island* was not providing any additional information and the higher value of *k* implies better generalization may have been achieved.

***Classification method 2 - Random forest*** Including all of the features in the analysis provided an accuracy marginally better than could be achieved using *k*nn when its features were carefully selected. No performance improvement was found by using fewer features, indicating that, for the Palmer penguin dataset at least, it requires considerably less implementation effort to achieve good performance using a random forest than it does using *k*nn. A marginal improvement in performance was apparent when *island*, *flipper length* or *body mass* were not included in training.

|  |
| --- |
|  |
| **Figure 4.** To estimate *k* for *k*-means, the elbow method uses the change in slope of ‘inertia’ (here *k*=3) and the silhouette method uses the score closest to 1 (here *k*=2) |

***Unsupervised method - k-means*** Although an unsupervised clustering method, *k*-means can be used for classification by matching clusters with classes. Only numerical features were included in the *k*-means analysis as it is not able to deal with unordered categorical data either directly or by labelling. The number of clusters (*k*) was selected using both the elbow and silhouette methods, giving values of *k*=3 and *k*=2 respectively, as shown in Figure 4. However, in practice it was found that accuracy improved significantly when *k*≥4 and this was probably due to the fact that, for smaller values of *k*, clusters were not always formed for all three species. Figure 5 illustrates the mapping of classes to clusters using two feature dimensions. No improvement in accuracy was obtained by reducing the number of features, but, in an additional experiment, separate sets of *k*-means clusters were created for each *sex* and this led to a small improvement in accuracy.

|  |
| --- |
|  |
| **Figure 5**  *k*-means clusters mapped to species according to majority voting. Assignments to classes are shown by polygon colours (here  *k*=10 and colouring limited to 50 samples). |

***A novel combined visualization and analysis (CVA) approach*** This work introduces the CVA approach that involves using visualizations of pairwise combinations of numerical data to identify a short sequence of two-dimensional linear classifiers based on SVMs. CVA requires greater manual effort to gain a deeper understanding of the nature of the dataset and this is in contrast with ‘black box’ classification approaches that are often applied with limited knowledge of the method adopted and little underlying insight into the nature of the data. The drawbacks of the CVA approach are that it is not generally applicable as it may not always be feasible or possible to extract the necessary insights from visualizations, and that the approach will become more difficult to apply as the number of features is increased. In application to the Penguin data, it was found to be able to produce results of accuracy almost as good as conventional approaches.

An application of CVA to the Penguin dataset is illustrated in Figure 6. Figure 6(a) shows the relationship between *bill depth* and *flipper length* and SVM is used to find a suitable ‘decision boundary’ that separates Gentoo from the other two species. Figure 6(b) then shows a second SVM line that best separates Adelie and Chinstrap using *bill length* and *bill depth*. A small improvement in accuracy was achieved when two separate SVM models were developed, one for each penguin sex.

|  |  |
| --- | --- |
| ` |  |
| **(a)** *bill depth* and *flipper length* with a decision boudary to distinguish Gentoo from the other two species | **(b)** *bill length* and *bill depth* allow the Adelie and Chinstrap species to be distingushed from one another |
| **Figure 6**. Example of two-stages CVA approach applied to the Palmer penguin data. The two-dimensional lines of separation shown in the figures are fitted using SVM and only to training data for the features shown on the axes. | |

**Conclusions**

With careful data preparation, optimization of metaparameters and robust application of training and testing methods, the *k*nn and random forest classification methods produced high-quality results. The *k*-means classification accuracy results were somewhat worse, but this is to be expected as the approach does not take advantage of target data information that is known to the supervised approaches. A classifier that is able to achieve 100% accuracy for the given data is possible, but its performance when applied to new unseen data would likely exhibit poor generalization.

The novel CVA approach is designed to use insights available in visualizations. Although needing to be tailored to each problem and not well-suited to high-dimensionality data, its internal operations are easy to visualize, an advantage not afforded to general-purpose classification methods. For the penguin data, it was able to produce accuracy results similar to those of other classification methods.

**Lost bits**

In addition to a simple baseline method, the analysis consists of k-Nearest Neighbour (*k*nn) and random forest classification, *k*-means unsupervised clustering and a novel and interesting classification approach that uses insights gained from visualizations to define a classifier that is formed of a small number of separate two-dimensional support vector machine (SVM) classifiers.

As the physical features of the male and female of each species differ considerably, the missing sex value made substituting the missing numerical attributes with statistically determined values cannot be justified. Consequently, these two records were deleted.

The remaining nine records with missing values all have no value for the *sex* attribute only. As can be seen in Table 2, the physical attributes of the male and female of each species are different and hence it is reasonable to consider assigning *sex* to those records missing this attribute. The process followed was that each of the four numerical attributes was independently standardized (to zero mean and unity standard deviation). It was then confirmed using a Shapiro-Wilk test that each numerical attribute exhibits a normal distribution [ref]. This confirmation is needed for a robust application of a Z-test, and this was performed with respect to the relevant species population to assess separately both the hypothesis that the missing sex value is Male and that it is Female [ref]. It was found that two of the records could be imputed as Male and three as Female at the 95% confidence level and these were retained in the dataset with an imputed sex attribute value. The remaining four records failed under both hypothesis tests and these were removed from the dataset. The cleaned dataset was used for the subsequent visualization and data analysis presented here and consists of 338 records made up of 147 Adelie penguins (74 male, 73 female), 68 Chinstrap penguins (34 male, 34 female) and 123 Gentoo penguins (62 male, 61 female).

**Table 2.** Mean values for the numerical attributes for each sex of the species in the dataset, with means following standardization of the dataset shown in parentheses

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Species** | **Sex** | **bill length** | **bill depth** | **flipper length** | **body mass** |
| Adelie | Female | 37.3mm (-1.22) | 17.6mm (0.24) | 187.8mm (-0.93) | 3368g (-1.04) |
| Adelie | Male | 40.4mm (-0.64) | 19.1mm (0.97) | 192.4mm (-0.61) | 4043g (-0.20) |
| Chinstrap | Female | 46.6mm (0.48) | 17.6mm (0.22) | 191.7mm (-0.65) | 3527g (-0.84) |
| Chinstrap | Male | 51.1mm (1.31) | 19.3mm (1.07) | 199.9mm (-0.07) | 3939g (-0.33) |
| Gentoo | Female | 45.6mm (0.30) | 14.2mm (-1.48) | 212.7mm (0.84) | 4680g (0.60) |
| Gentoo | Male | 49.5mm (1.01) | 15.7mm (-0.73) | 221.5mm (1.47) | 5485g (1.60) |

The imbalance can be reduced either by deleting records of classes that are overrepresented in the dataset, or using ag statistical approaches to generate additional records of classes that are underrepresented.

Unbalanced - knn robust, logistic regression not, random forest robust, svm can be affected, k-means inbalance not relevant

Predicting the penguin species from the given features is a classification problem. A baseline classification method is first implemented, providing a reference for the performance of other methods. This report considers two supervised approaches, namely knn and random forest. An unsupervised k-means approach is also taken, but as the clusters found are unlabelled, they need to be related to the species so that the classifying accuracy can be determined. An interesting and usual approach is also described that uses a combination of the insights found from visualizations. This led to the identification of a short sequence of two-dimensional linear classifications using Support Vector Machines (SVMs) that were able to perform at least as well as the other classification methods. Regression approaches were not considered as although categorical values could be assigned numerical values, performance is likely to poor unless they have recognizable ordinal counterparts, which is not the case for this dataset.

*k*NN is known to be biased in favour of features with smaller standard deviations[ref], so the data were standardized before analysis. The performance of the random forest is known to be largely unaffected by the standard deviation of the data and so no standardization was performed. For *k*-means clustering, standardization or normalization of features is often recommended. This is because *k*-means clustering relies on calculating distances between data points to assign them to clusters. SVM also standardization.

It also gives a basis for the comparison of principal methods being considered. In a classification problem, the baseline used is often to select the class that has the most observations, in regression the mean or the median and in unsupervised learning perhaps a random assignment of data values to clusters is made

Training for the *k*NN algorithm simply involves storing the training dataset and classification requires finding the distances from the test data value to the training examples. The *k* nearest neighbours to the test data value are found and the most common class label among the *k* neighbours is the predicted label.The results obtained for the classification are shown in Table 3. Comments?

A random forest is a collection of decision trees; each tree being trained separately on a subset of the training data using a subset of features. Decision tree training builds a set of nodes each defining a test on selected feature value to determine which output branch to follow. The CART (classification and regression tree) algorithm is often used to determine which feature to use at the node and one of a range of possible methods to assess the quality of the split. Table 3 shows… Comments?

In the first cycle, *k*-means randomly selects *k* data points as cluster centroids in the training data space and assigns the remaining data points to the cluster whose centroid is closest (using Euclidian distance or another method). The means of the data points in each cluster are calculated and then used as the centroids in the next cycle. The cycle repeats until the clusters no longer change or a defined maximum number of iterations is reached.

You should consider how to visualize the data and which algorithms to

try. Nothing you do will be completely successful, this coursework is

not here to judge your final accuracy but the care you bring to your

investigation. Here are some thing you should consider:

\begin{itemize}

\item The kind of algorithm to use, for example whether to classify, regress or cluster.

\item The metric to use to measure the performance of the model.

\item What sort of baseline to compare the model to.

\item How to choose the hyperparameters of your model.

\end{itemize}

For good marks you should include some graphs that illustrate

properties of the data and you should compare two classification

algorithms, both to each other and to a baseline model. The algorithms

you pick do not need to be unusual, for example $k$nn classification

would be perfectly good, though, of course, for full marks this would

include some consideration of how to pick $k$ and how to measure the

distance, though, as you know, no approach to chosing $k$ is every

going to be completely satisfactory. In addition, you should include

either some exploratory regression or unsupervised learning; for

regression you might regress two properties and examine whether the

regression parameters are the same for each penguin type; unsupervised

learning could use $k$-means, for example. You do not need to do both

regression and unsupervised learning.

You should make sure any assessment is not restricted to the data used

in train models or decide on metaparameters. In your report you should

explain your decisions. You code will not be marked for elegance, but

it should run correctly; it is expected you will use Python, but any

of Python, Julia or R is fine. Do not include screenshots of graphs,

they should be imported directly; resize them to the correct size

before importing them, if the labels are tiny the graphs will not be

marked. Make sure figure captions are descriptive, it is better to

have some overlap between figure captions and the main text than to

have figure captions that are not reasonably self-contained.

As a rough guide to marking:

\begin{itemize}

\item Initial description of the data, including some graphs or other approaches to visualisation. 6 marks.

\item Either unsupervised learning or regression. 6 marks.

\item Two algorithms should be tested, if only one algorithm is

included the 28 available marks will be halved.

\item Overall presentation (3 marks), including use of appropriate

sections, plots, diagrams, or tables to make your point. Do not

include code snippets in the report. Instead, describe in words or

equations what you are implementing. Format equations correctly.

\item Suitable choice of algorithms (4 marks).

\item Suitable choice of evaluation for algorithms (3 marks).

\item Comparison with a suitable baseline (3 marks) and a justification for which baseline to use.

\item A description of metaparameter selection (3 marks), if one

algorithm has not metaparameter, then explain that and note why not

and why this do or does not make it a better algorithm for these

data.

\item Describe and compare the results from your two algorithms,

include a description of how you implemented the algorithms. (6 marks)

\item There are some marks (6 marks) for something suprising and unusual.

\end{itemize}

\section\*{Question 2 - Ethical challenge facing us in data science and AI}

For two of these three types of ethical challenge facing us in data science and AI:

\begin{enumerate}

\item The protection of data, of the people whose data they are and participants in any study.

\item Avoiding the amplification of biases and regressive values implicit in historic dataset.

\item The safety of AI systems and the possible of existential threats from machines.

\end{enumerate}

describe what you think is a specific example of a challenge that

could arise or has arisen in the past. Obviously the three broad types

of challenge overlap, do not worry about the boundaries between these

types, but do try to address different types of threat in your

examples. Explain how the ethical problems could be addressed, or at

least made more transparent.

\subsection\*{Report}

Your report should be no longer than five pages, including any

references. It is expected that Question 2 would occupy about a fifth

of this space; use an 11 or 12pt font and do not try tricks like

expanding the margin to fit in more text, shorter is better than

longer.

Your report must be submitted in pdf and should be prepared in LaTeX;

overleaf is a good approach, but not required as long as LaTeX has

been used. As always when using LaTeX, give yourself over to defaults,

our expectation of what a document should look like has been

conditioned on LaTeX, so it is best not to try to override the look of

the document.

Avoid code snippets in the report unless that feels like the best way

to illustrate some subtle aspect of an algorithm; do always though

consider a mathematical description if possible. You will be asked to

submit code and it may be tested to make sure it works and matches

your report. It will not, however, be marked in and of itself.

\subsection\*{knn}

Perhaps use F1-score (there are others!) as the classes are imbalanced in number?

F1-score is a metric that considers both precision and recall. Precision measures the accuracy of positive predictions (TP/(TP+FP)), while recall (also known as sensitivity) measures the fraction of positives that were correctly identified (TP/(TP+FN))

F1-score is the harmonic mean of precision and recall and is calculated as follows:

F1 = 2x(PrecisionxRecall)/(Precision+Recall)

F1-score ranges from 0 to 1, where a higher value indicates better model performance. F1-score is particularly useful when classes are imbalanced because it considers both false positives and false negatives.

\section\*{Report template}

This is a report template, you don't need to use this template, but do

use it if it is helpful.

Here is an example of an equation:

\begin{equation}

\pi=4\left(1-\frac{1}{3}+\frac{1}{5}-\frac{1}{7}\ldots\right)

\end{equation}

or

\begin{equation}

\pi=4\sum\_{n=0}^\infty\frac{(-1)^{n}}{2n+1}

\end{equation}

where $\pi$ can be written in line by using \$'s. Here is a vector:

\begin{equation}

\mathbf{x}=\left(\begin{array}{c}x\_1\\x\_2\end{array}\right)

\end{equation}

<https://machinelearningmastery.com/how-to-get-baseline-results-and-why-they-matter/>