**AI Course Work**

**Abstract**

In this study, I was asked to perform an analysis of the penguin dataset using machine learning methods, including unsupervised learning algorithms and classification models. I first applied principal component analysis (PCA) to simplify the data structure and explored how K-means clustering can group the data without prior labels, and then used visualization tools to show the unique characteristics of different penguin species. To predict species, logistic regression and support vector machine (SVM) models were developed. Each model has been tested in detail to confirm its accuracy and robustness. Additionally, techniques such as cross-validation, learning curves, and external validation are used to enhance model stability and prevent overfitting.

**1 Exploration of the Data**

The penguin dataset was evaluated using descriptive statistics, analyzing measurements such as beak length, beak depth, flipper length and body mass, as well as taxonomic details for species and sex. Table I lists the statistics for each of the four features across the 342 data points. The analysis showed that the median values for beak length, beak depth, and body mass were slightly above the mean, indicating a slightly right-skewed distribution of these traits. In contrast, the median flipper length was slightly lower than the mean, indicating a different distribution pattern for this trait. The standard deviation and extreme values for each trait show that body weight has the most variation among these traits, while beak depth has the least variation. In particular, the large difference between the minimum and maximum values of body weight reflects significant variation among individuals in the dataset. Quartiles further reveal details of the data distribution. The interquartile spans for beak length, beak depth and wing length were all small, indicating that most of the data were concentrated and had relatively little variability. The interquartile spans for body weight were much larger, reflecting higher data dispersion.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Statistic | Bill Length (mm) | Bill Depth (mm) | Flipper Length (mm) | Body Mass (g) |
| Count | 342 | 342 | 342 | 342 |
| Mean | 43.92 | 17.15 | 200.92 | 4201.75 |
| Standard Deviation | 5.46 | 1.97 | 14.06 | 801.95 |
| Minimum | 32.1 | 13.1 | 172 | 2700 |
| 25% Percentile | 39.23 | 15.6 | 190 | 3550 |
| 50% Percentile | 44.45 | 17.3 | 197 | 4050 |
| 75% Percentile | 48.5 | 18.7 | 213 | 4750 |
| Maximum | 59.6 | 21.5 | 231 | 6300 |

Overall, this data is very good for us to do some correlation exploration, machine learning algorithm practice. However, there are some anomalies and missing values in the data, so in the data analysis and exploration stage, I filled the missing values of the categorical data according to the multitude, and the missing values of the numerical data according to the median, which is helpful for not disturbing the distributional characteristics of the dataset, and facilitating the next correlation analysis.

Of course, I performed some plotting to show this after the correction. Histograms segmented by species showing differences in features such as finned limb length seem to clearly see the differences between species and facilitate our practice of machine learning algorithms, which are critical for species identification. Scatter plots depicting rostrum length vs. rostrum depth colored by species provide insight into how these dimensions vary between species and may help to differentiate between them.

Next a correlation analysis was performed, as I intended to check for high correlation in case there were consequences of overfitting. The correlation matrix was used to check the relationship between different measurements. It was viewed to see which dimensions were closely related and could predict each other, thus affecting dimensionality reduction or feature selection.

First, penguin beak length showed a strong positive correlation with wing length (0.66), suggesting that penguins with longer beaks also tended to have longer wings. Second, beak depth showed a strong negative correlation with wing length (-0.58), implying that penguins with deeper beaks tend to have shorter wings. In addition, there was a very strong positive correlation between wing length and body weight (0.87), showing that penguins with longer wings also weighed more. Finally, there was a moderately negative correlation between beak depth and body weight (-0.47), pointing out that penguins with deeper beaks tend to weigh less. Because of the slightly higher degree of these correlations, in the next step of the analysis I used PCA downscaling for possible data exploration, visualized by a 3D scatterplot showing how the first three principal components differentiate between species. Based on the 2D principal component analysis, I used the third principal component to capture the missing variance information from the first two principal components and printed the PCA component loadings as shown below.

图表, 散点图

描述已自动生成

The first principal component mainly reflects the overall body size of the penguins, which can be seen from the strong positive loading values of wing length (0.58) and body weight (0.55), both of which contribute significantly and positively to this component. In addition, the loading value of -0.4 for beak depth indicates that beak depth tends to decrease as wing length and body weight increase. This may imply that larger penguins usually have shallower beaks, which may be related to specific ecological adaptations or feeding behaviors. It can be seen that the second principal component has a very strong positive contribution from beak depth (0.8), as does beak length (0.6), but the contributions from wing length and body mass are almost negligible. This suggests that this principal component mainly captures differences in beak morphology, which is strongly affected by beak size, and also suggests that the second principal component may be particularly useful in distinguishing different species of penguins, as penguin beak characteristics are a key factor in distinguishing differences between species. Key Features. On the other hand, the third principal component revealed an interesting inverse correlation pattern, with a strong positive contribution from beak length (0.64) and a strong negative contribution from body mass (-0.6). This may indicate that some penguin species have longer beaks and lighter weights, a pattern that may be related to specific habits or environmental adaptations. Beak depth and wing length also showed inverse loadings, further emphasizing the complex association between body size and beak characteristics in this component. The above analyses provide a deeper understanding of how penguin morphological traits affect their classification, and also allow for further species determination based on these principal components, providing us with an effective way to reduce the dimensionality of the data while maintaining key biological information.

**2 Unsupervised Learning Approach**

Next unsupervised learning was applied for clustering study, the method used was clustering using the principal component metrics of PCA obtained in the data preprocessing stage, and the conclusion obtained was a profile coefficient conclusion of 0.55. Using this simple clustering approach is easy to be accepted as it can also be seen before that the various metrics are correlated to a high degree, so I have used PCA clustering method as a base method to be used as a basis for the comparison of the other methods that follow, on which improvements are proposed.

The first method of improvement is KNN clustering analysis, which is more suitable for spherical clustering and for the case of complex cluster distribution. In order to find the most suitable number of clusters for clustering, the comprehensive study can be clearly seen based on the elbow diagram that the slope of decline is gentle when 3 or 4 is used. Here we continue to choose 3 as the clustering marker. Cutting the dataset into a training set and a test set yields a contour coefficient of 0.61, the contour coefficient ranges from -1 to 1. Values close to 1 indicate that the clusters are tightly packed within the clusters and the clusters are well spaced apart, and 0.61 indicates that the clustering results are acceptable.

**3 Classification Algorithm**

During the use of the classification algorithm I used logistic regression and support vector machines.

Logistic regression one of the most important hyperparameter selection, by setting the hyperparameter network of logistic regression to find the best meta-parameter, the conclusion was obtained that the value of 1 was selected, which shows that the linear logistic regression is very effective. But the final conclusion of the logistic regression accuracy as high as 1, which inevitably caused my suspicion, I suspect that is in the pre-processing stage is the data to produce any problem, or model selection has a problem, so I put the PCA clustering after the results of the principal component analysis test, the conclusion of the same as high as 0.99, but no longer the value of 1.

Given the validity of logistic regression with linear decision boundaries, the next step for the SVM implementation starts with a linear kernel. Nonlinear kernels (such as RBF, polynomial, or S-functions) can model more complex patterns, but they also increase the risk of overfitting, especially for high-dimensional data, and they are more computationally expensive.The results of the SVM showed an accuracy as high as 0.98, again a very high value. I therefore repeated the data exploration to look at some of the relevant metrics and found nothing wrong.

Prior to this, I chose the PCA cluster analysis model as one of my baseline models, and since logistic regression and support vector machines were too good, I used another baseline model, the most frequent category predictor. This is a very basic model for selecting and predicting the most frequent categories in a dataset, without looking at the specific features of the data, and making decisions based only on which categories occur most often. The results of this model tell us that the fit was 0.48, which is lower than the 0.55 obtained by our other baseline model, the PCA clustering model, showing that the model is really simple, and also at the same time showing that there is nothing wrong with our data, and that it is the method of selection that is effective.

Finally, cross-validation, learning curves and external validation are used to further validate the stability and reliability of the model to avoid overfitting and ensure the generalisation ability of the model. The model was optimised and the dataset was classified. I gained a very deep learning from this practice and benefited from learning a series of machine learning algorithms from it.

TASK2:

The first is the risk and ethical challenge of "avoiding the amplification of implied biases and regressions in historical datasets". Mathematically speaking, there are two metrics in historical datasets that we are most interested in when building data science models, which we often express as "information": the variance and the mean. Of course, since we are often working with historical data and have no way of knowing what the totality is, they are often expressed as bias and regression values. This is the key to our data science modelling, but as there are correlation situations, or conditions that change, making this dependence perhaps affecting the correctness of our models, but difficult to incorporate a posteriori information to adjust for, there is a risk associated with amplifying the implied bias and regression values in the historical dataset. In social practice, the risk from this amplification can sometimes be devastating for a small group. For example, an AI designed to do recruitment work may have discriminatory, unreasonable behaviour due to lack of information about certain geographical, or niche groups. Assuming the existence of ex-prisoners, a certain degree of probability will be insufficient capacity, but the existence of this probability is put into the overall view is objective, but may be due to the acquisition of too little sample information, so that the recruitment of AI to make some discriminatory behaviours, then for the niche groups or the existence of geographically deficient people is unfair, this is a very typical exaggerated historical data implied by the bias and social problems arising from regression values. However, due to the lack of a posteriori information, we cannot optimise the model well.

In the second case, there are still some risks and ethical challenges in the field of data science and AI, such as the safety of AI systems and possible problems with machines. The future potential for AI in healthcare is huge, but we have to be vigilant about the safety of AI itself and problems caused by machines, such as data contamination of healthcare models that have been deliberately hacked, resulting in limited model safety. But if this is a model for medical diagnosis of patients, then the harm caused will be enormous. Machines are machines after all, there is a risk of paralysis, just as computers need the arithmetic power of cpu and gpu, the data storage devices required for AI in healthcare, the systems that run them, rely on the machines, and may also exhibit unpredictable behaviour due to software errors or hardware failures, which is also a matter of caution.