Training dataset:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Breed Name** | **Size** | **Lifespan (years)** | **Avg Weight (kg)** | **Origin** | **Temperament** |
| Labrador | Large | 10-12 | 30 | Canada | Friendly, Active |
| Beagle | Medium | 12-15 | 10 | England | Curious, Merry |
| German Shepherd | Large | 9-13 | 35 | Germany | Intelligent, Loyal |
| Golden Retriever | Large | 10-12 | 32 | Scotland | Friendly, Reliable |
| Poodle | Medium | 12-15 | 20 | Germany | Intelligent, Active |
| Chihuahua | Small | 14-16 | 2.5 | Mexico | Charming, Lively |

|  |
| --- |
| 1. **Feature:** It is used to represent individual variables that act as input. Here, size, lifespan, avg weight, origin and temperament are the features**.** 2. **Label:** It is the description of what the particular data represent. Breed name is the label here. 3. **Prediction:** It is the process of using the data to foresee the output data. Here prediction is to predict the breed’s name of the dog using features like size, lifespan, origin and temperament. 4. **Outlier:** It is the datapoint that is different from the rest. Here the breed chihuahua with avg weight of 2.5 kg. 5. **Test Data:** After the model is trained, test data is used to evaluate performance and progress of the model. A set of data- size(Medium) , lifespan(16 years) , weight(9 kg) is used to test the model to give the output as beagle. 6. **Training Data:** the data you use to train a machine learning algorithm or model. All the data given in the table(around 4 rows taken) come under training data. 7. **Model**: A program that has been trained to find patterns within new data and make predictions. Example: model is expected to predict the breed, when size, weight and origin is given. 8. **Validation Data:** it is the dataset used to evaluate the model during the training process and helps in tuning the model. Example : a data set with size(small), lifespan(15 years), weight(2.6kg) and origin(Mexico) is expected to be evaluated as Chihuahua. 9. **Hyperparameter:** The parameter whose value is set before the learning process begins. Example: Batch Size( The number of training examples utilized in one iteration) is fixed for a model before its training starts. 10. **Epoch:** It is one complete pass of the training dataset through the algorithm. Example: for a dataset of 6 samples and batch size of 2, each epoch has 3 iterations. 11. **Loss Function**: It is a function consisting difference between the predicted values and the actual values. Example: let the lifespan of labrador be 11 and the prediction be 10.5, then square error is 0.25. Similarly square error of all other breeds are calculated and average is taken to find mean square error value. 12. **Learning Rate:** It is determined by how much step the model takes to minimize the loss function. Example: if learning rate is 0.01, then it takes more iterations to reach optimal solution whereas learning rate of 0.1 takes less iterations. 13. **Overfitting:** It occurs when the model gives perfect result for the training data but poorly for new data. Example: correct result may be given for german shepherd and labrador , that its size is large. But when new breed called Australian shepherd is tested with lifespan of 12-14 years, average weight of 25 kg, and a temperament of energetic, then it may predict wrongly as medium sized. 14. **Underfitting:** It occurs when the model performs poorly in both test and training data. Example:Poodle and beagle are said to be medium sized despite their difference in weights. 15. **Regularization**: It is a method used to reduce error and prevents overfitting. Example: categories like size and origin can be changed to numerical values. 16. **Cross-Validation:** It is a method using to evaluate the model on unseen data. Example: the datasets can be divided into 3 parts, where different combinations can be given as input and the model is evaluated. 17. **Feature Engineering:** New features can be brought from the existing data to improve the performance of the model. Example: Average lifespan can be calculated based on the size of the breed. 18. **Dimensionality Reduction:** It is used to reduce the number of features in a dataset while preserving as much information as possible. Example: size and average weight can be combined together into a single attribute. 19. **Bias:** It is the difference or error occurring between the model’s predicted value and the actual value. Example: let the lifespan of labrador be 11 and the prediction be 10.5, then error is 0.5. 20. **Variance**: Variance measures how much values in a dataset differ from the mean of the dataset. Example: for the above dataset, variance is 143.37 kg². |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |