**PES University, Bangalore**

(Established under Karnataka Act No. 16 of 2013)

MAY 2020: IN SEMESTER ASSESSMENT (ISA) B.TECH. IV SEMESTER

**UE18MA251- LINEAR ALGEBRA**

**MINI PROJECT REPORT**

ON

**PREDICTION OF ANIMAL TRAITS USING DA, ANN, SVM AND PCA**

Submitted by

1. Name: **Sakshi Shetty** SRN: **PES1201800190**

2. Name: **Snigdha S Chenjeri**  SRN: **PES1201800045**

3. Name: **Sruthy S**  SRN: **PES1201801143**

Branch & Section: **CSE, 4C**

**PROJECT EVALUATION**

( For Official Use Only )

|  |  |  |  |
| --- | --- | --- | --- |
| **Sl.No.** | **Parameter** | **Max Marks** | **Marks Awarded** |
| 1 | Background & Framing of the problem | 4 |  |
| 2 | Approach and Solution | 4 |  |
| 3 | References | 4 |  |
| 4 | Clarity of the concepts & Creativity | 4 |  |
| 5 | Choice of examples and understanding of the topic | 4 |  |
| 6 | Presentation of the work | 5 |  |
|  | **Total** | **25** |  |

Name of the Course Instructor : **P RAMA DEVI**

Signature of the Course Instructor :

**I. INTRODUCTION**

Linear Algebra is the foundational branch of mathematics for most products in technology that we see today. Prediction algorithms, neural networks, facial recognition are the base concepts required for making more user-driven and personalised content in technical products.

The applications of linear algebra in computer science that we have laid emphasis on is Artificial Neural Networks,Support Vector Machine and Principal Component Analysis. To demonstrate this, we have chosen a dataset containing the classification of various traits of more than 100 animals. The end result is the prediction of certain traits and features based on the specific models.

Our dataset contains 101 animals and their various respective traits: physical traits such as aquatic/airborne/teeth etc and characteristic traits such as predator/venomous/domestic and class type. Upon analysing our data, our main focus is on the 4 traits mentioned above. In short, our models will predict these traits in each animal.

In the subsequent sections, we will describe in detail, according to our understanding, how linear algebra's core concepts are the foundation for the very existence of machine learning.

**II. REVIEW OF LITERATURE**

Artificial neural network (ANN), usually called Neural Network (NN), is an algorithm that was originally motivated by the goal of having machines that can mimic the brain. A neural network consists of an interconnected group of artificial neurons. They are physical cellular systems capable of obtaining, storing information, and using experiential knowledge.

Presently, ANNs have been in use for economic and financial purposes such as stock markets, and on a large scale for making personalised predictions in search engines and streaming platforms for media and music.

French et al (1992) developed three layered neural networks to forecast rainfall intensity fields in space and time. After training with input patterns, the neural network was used to forecast rainfall using the current fields as input. They showed that ANN was capable of learning the complex relationship describing the spacetime evolution of rainfall.

Many studies reported in the literature on the application of ANN in the field of water resources were in the field of stream flow and rainfall forecasting. Few studies have been concentrated on reservoir operation. In most of the studies, feed forward structure and the back propagation algorithm have been used to design and train the ANN models respectively.

Support Vector Machines (SVMs) as originally proposed by Vladimir Vapnik

within the area of statistical learning theory and structural risk minimization, have

demonstrated to work successfully on various classification and forecasting problems.

SVMs have been used in many pattern recognition and regression estimation problems

and have been applied to the problems of dependency estimation, forecasting and

constructing intelligent machines.

R. Burbidge et. al., have shown that the support vector machine (SVM)

classification algorithm, proves its potential for structure–activity relationship

analysis.Giorgio Valentini have proposed classification methods, based on non-linear

SVM with polynomial and Gaussian kernels, and output coding (OC), ensembles of

learning machines to separate normal from malignant tissues, to classify different

types of lymphoma and to analyze the role of sets of coordinately expressed genes

in carcinogenic processes of lymphoid tissues.

**III. REPORT ON PRESENT INVESTIGATION**

3.1. VECTORS

Every time we use a dataset, each of its rows are split into and represented as vectors—the combination of these vectors make the data matrix and result in our dataset as we see it.

3.2. LABELLING & ONE-HOT/BINARY ENCODING

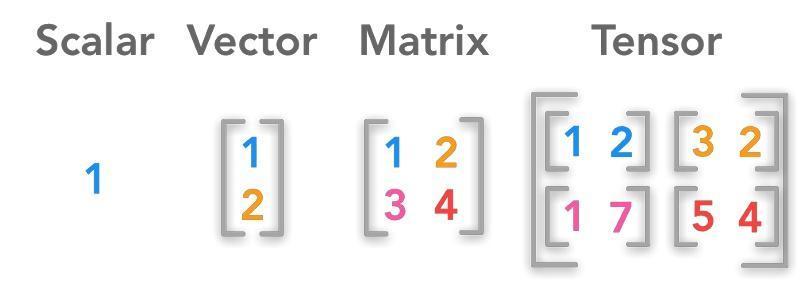
All categorical variables (strings, characters and the like) need to be converted to numerical format to be processed by a neural network. Each row, in such cases, is encoded as a binary vector, i.e. a vector with 0 or 1 value. For instance: in our dataset, if an animal is venomous, it is indicated by '1', and if it's not, it's indicated by '0'.

3.3. NORMALISATION

Normalisation is the process of taking data from a problem and reducing it to a set of relations while ensuring data integrity and eliminating data redundancy. Data integrity—all of the data in the database are consistent, and satisfy all integrity constraints.

3.4. TENSORS

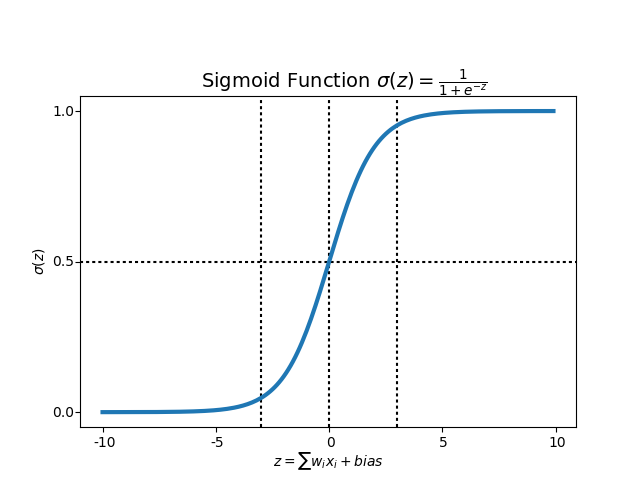
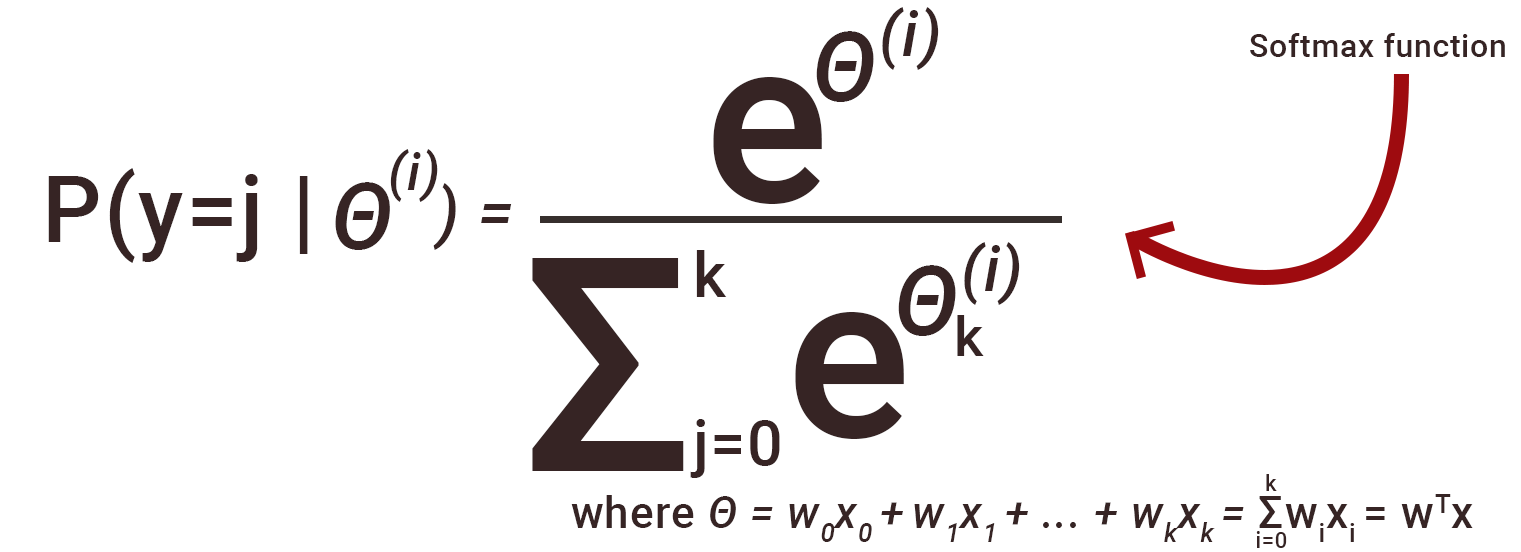
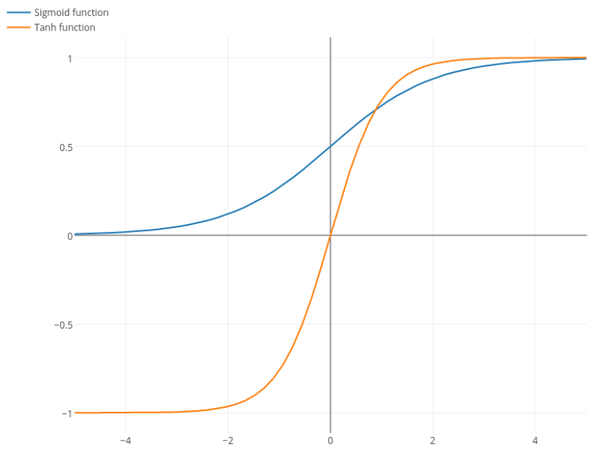
The very core component of any neural network is tensors. To put it simply, tensors house data in N dimensions (matrices are hence 2D tensors). The popular python library tensorflow is aptly named so.



In machine learning, we need to encode multidimensional data. For example, images are represented by 3 dimensions—height, width, depth (colour)—but while dealing with huge amounts of data, the sample size becomes the 4th dimension. Hence, 4D tensors can neatly pack images for further processing.

3.5. ACTIVATION FUNCTIONS

Neural network activation functions are a crucial component of deep learning. Activation functions determine the output of a deep learning model, its accuracy, and also the computational efficiency of training a model—which can make or break a large scale neural network. In the subsequent sections, we explain in detail the three functions we have used in our model - sigmoid, softmax and tanh.

* SIGMOID: The sigmoid function is used because it exists between (0 to 1). Therefore, it is especially used for models where we have to predict the probability as an output. Since probability of anything exists only between the range of 0 and 1, sigmoid is the right choice. The formula and graph are as shown. It gives a smooth gradient, preventing “jumps” in output values. The output values are bound between 0 and 1, normalising the output of each neuron.
* SOFTMAX: Softmax is often used in neural networks, to map the non-normalized output of a network to a probability distribution over predicted output classes. It can handle multiple classes—only one class in other activation functions normalizes the outputs for each class between 0 and 1, and divides by their sum, giving the probability of the input value being in a specific class. Softmax is useful for output neurons—typically Softmax is used only for the output layer, for neural networks that need to classify inputs into multiple categories.
* TANH: The logistic sigmoid can cause a neural network to get “stuck” during training. This is due in part to the fact that if a strongly-negative input is provided to the logistic sigmoid, it outputs values very near zero. Since neural networks use the feed-forward activations to calculate parameter gradients. It makes it easier to model inputs that have strongly negative, neutral, and strongly positive values.

3.6. SVM

The main objective of SVM is to find the optimal hyperplane which linearly separates the data points in two component by maximizing the margin . Hyperplane is the line(in 2D) and plane(in 3D) that linearly separates data points into two components.

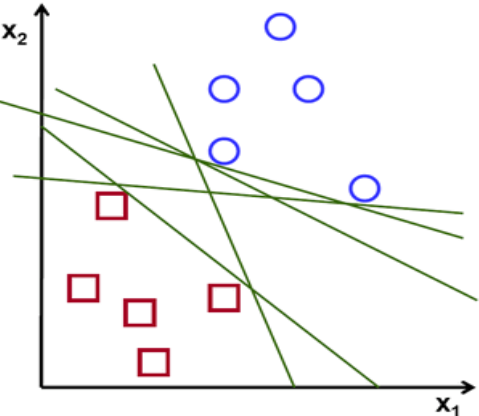
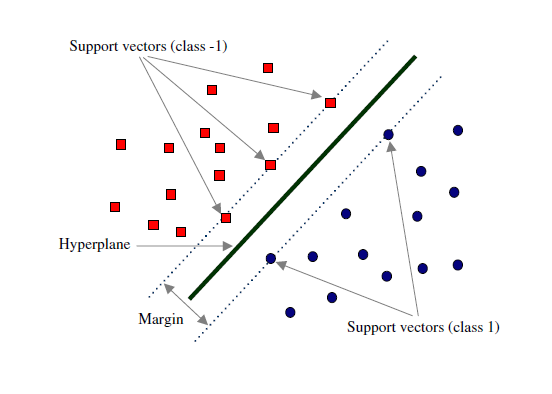
The hyperplane can be written as:

y= ax+b

ax+b-y=0

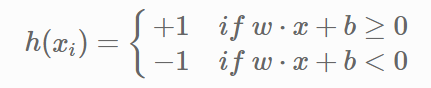
Let X=(x,y) and W=(a,-1) then in vector form,hyperplane is

**W.x+b=0**

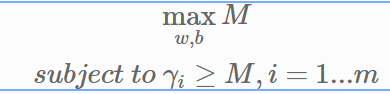
 

As we can notice,there can be any number of hyperplanes to separate the data into two components. The goal is to choose the optimal one. This is done with the help of margins and support vectors.The distance between hyperplane and optimal hyperplane is know as margin, and the closest data-points are known as support vectors.

Once we have the hyperplane,we can then use the hyperplane to make predictions.We define the hypothesis function as

****

To find the values of w and b of the optimal hyperplane, we need to solve the following optimization problem, with the constraint that the geometric margin of each example should be greater than or equal to M:



On further simplification,we get the values of w and b which can be substituted in the respective equation to derive the optimal hyperplane.

3.7. PCA

The central idea of principal component analysis (PCA) is to reduce the dimensionality of a data set consisting of a large number of interrelated variables while retaining as much as possible of the variation present in the data set. This can be achieved using these steps:

1. Take the whole dataset consisting of d+1 dimensions and ignore the labels such that our new dataset becomes d dimensional.
2. Compute the mean of every dimension of the whole dataset.
3. Compute the covariance matrix of the whole dataset.
4. Compute Eigenvectors and corresponding Eigenvalues

* Let A be a square matrix, ν a vector and λ a scalar that satisfies Aν = λν, then λ is called eigenvalue associated with the eigenvector ν of A. Eigenvalues are roots of the characteristic equation:



1. Sort the eigenvectors by decreasing eigenvalues and choose k eigenvectors with the largest eigenvalues to form a d × k dimensional matrix W.
2. Transform the samples onto the new subspace

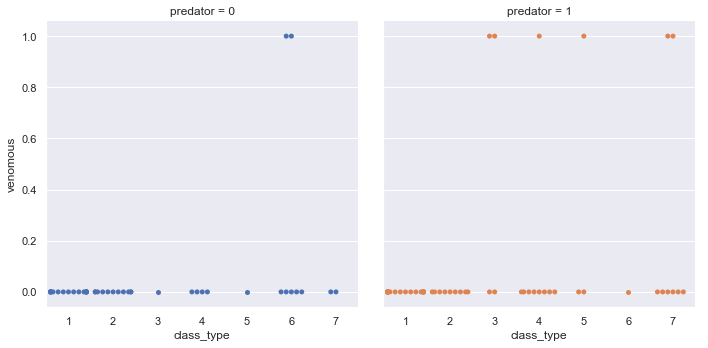
**IV. RESULTS AND DISCUSSIONS**

4.1 DATA ANALYTICS

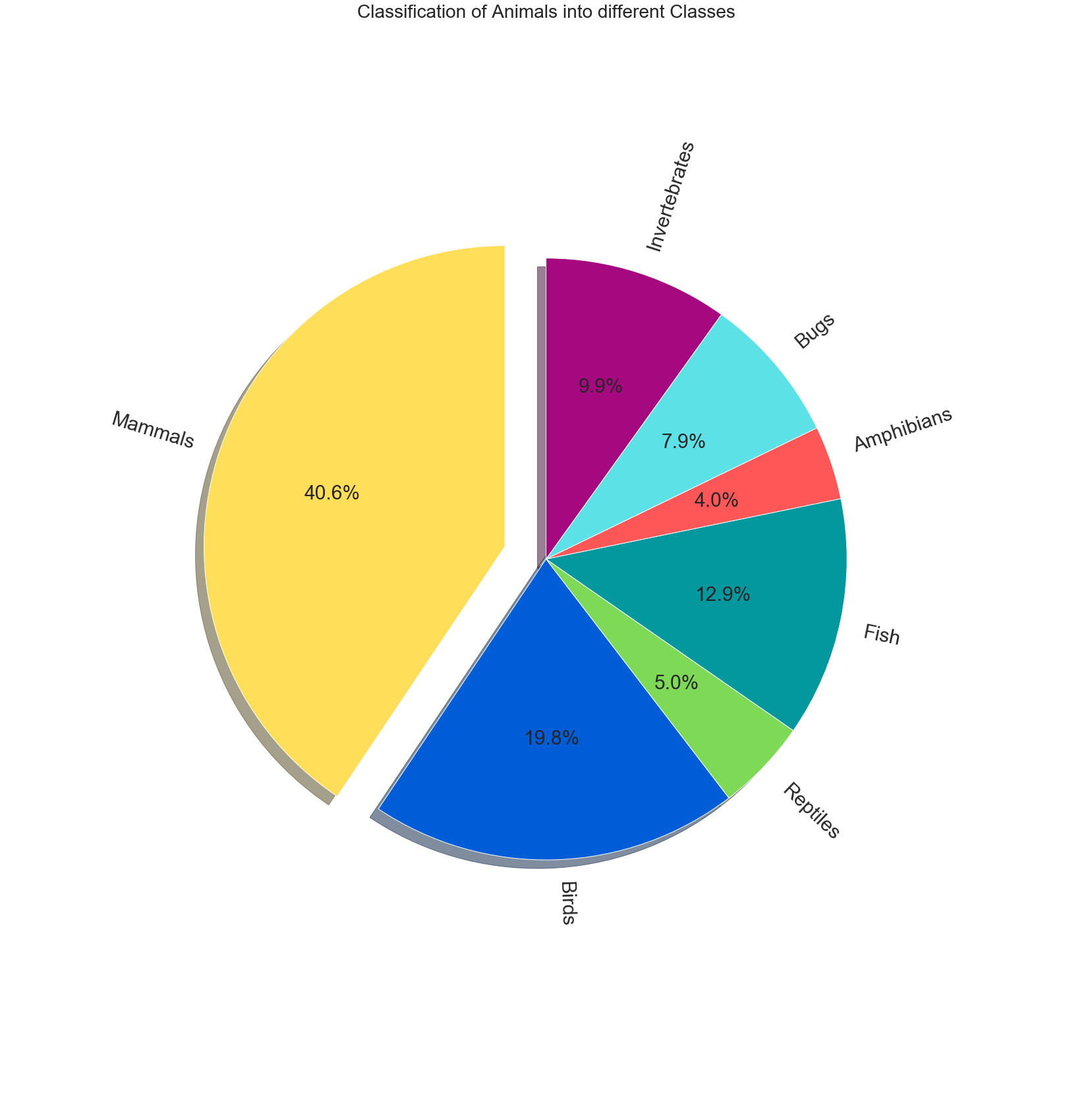
The first step was to clean the dataset by replacing missing and negative values with zero, and further replacing non-numeric characters with zero. After this, we visualised the data to observe interdependencies of traits and their correlations, along with normalisation.

The visualisation of the data is as follows from our code:

a) Swarm Plot:

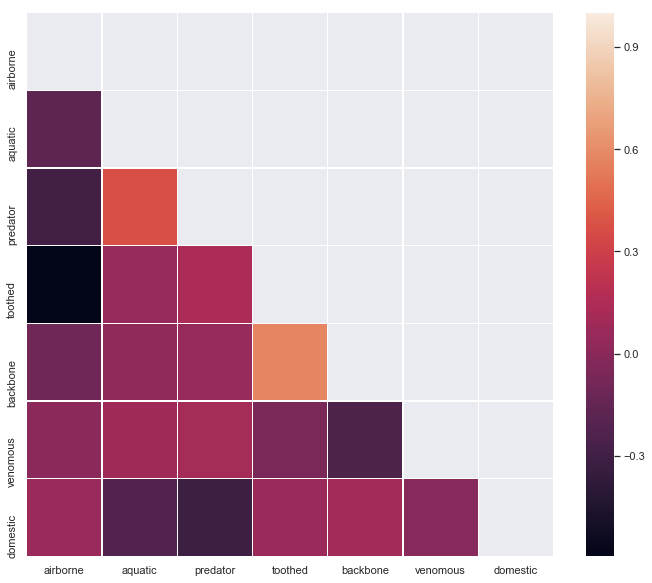
Our dataset has 7 class types of animals: mammals, birds, reptiles, fish, amphibians, bugs and invertebrates (in this order). For all seven types, below we have the left graph that shows the number of non venomous predators in each class (for ex. lions in mammals) and the right graph shows the number of venomous predators in each class (for ex. snakes in reptiles).

b) Pie Chart:

This is a pie chart showing the percentage of animals belonging to each class in our dataset:

c) Heatmap / Graphical Correlation:

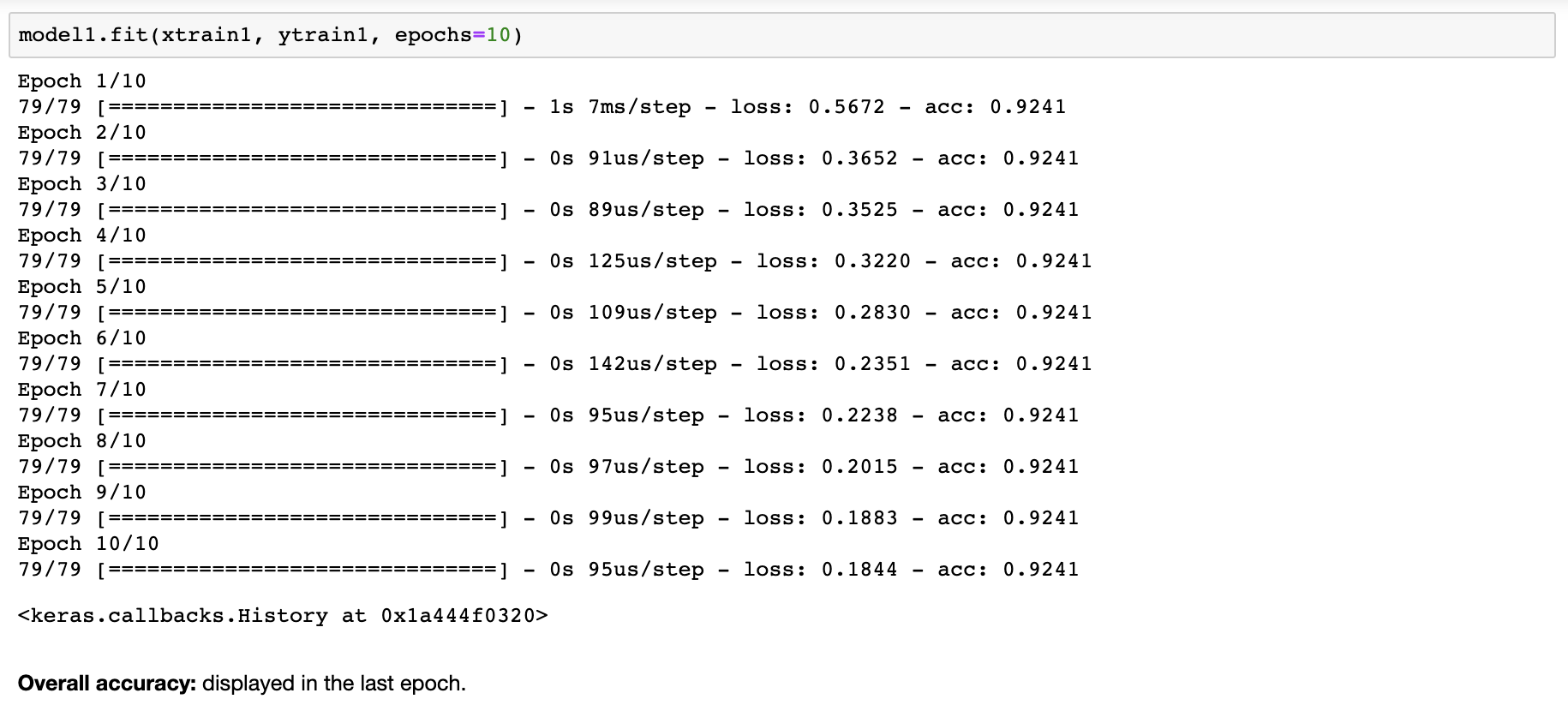
This heat map shows us which traits are most correlated for each other. An inference from this map would be that due to the high correlation between ‘backbone’ and ‘venomous’, invertebrates are less likely to be venomous. Similarly, the low correlation between ‘aquatic’ and ‘predator’ indicates that in this particular dataset, there aren’t many aquatic predators.



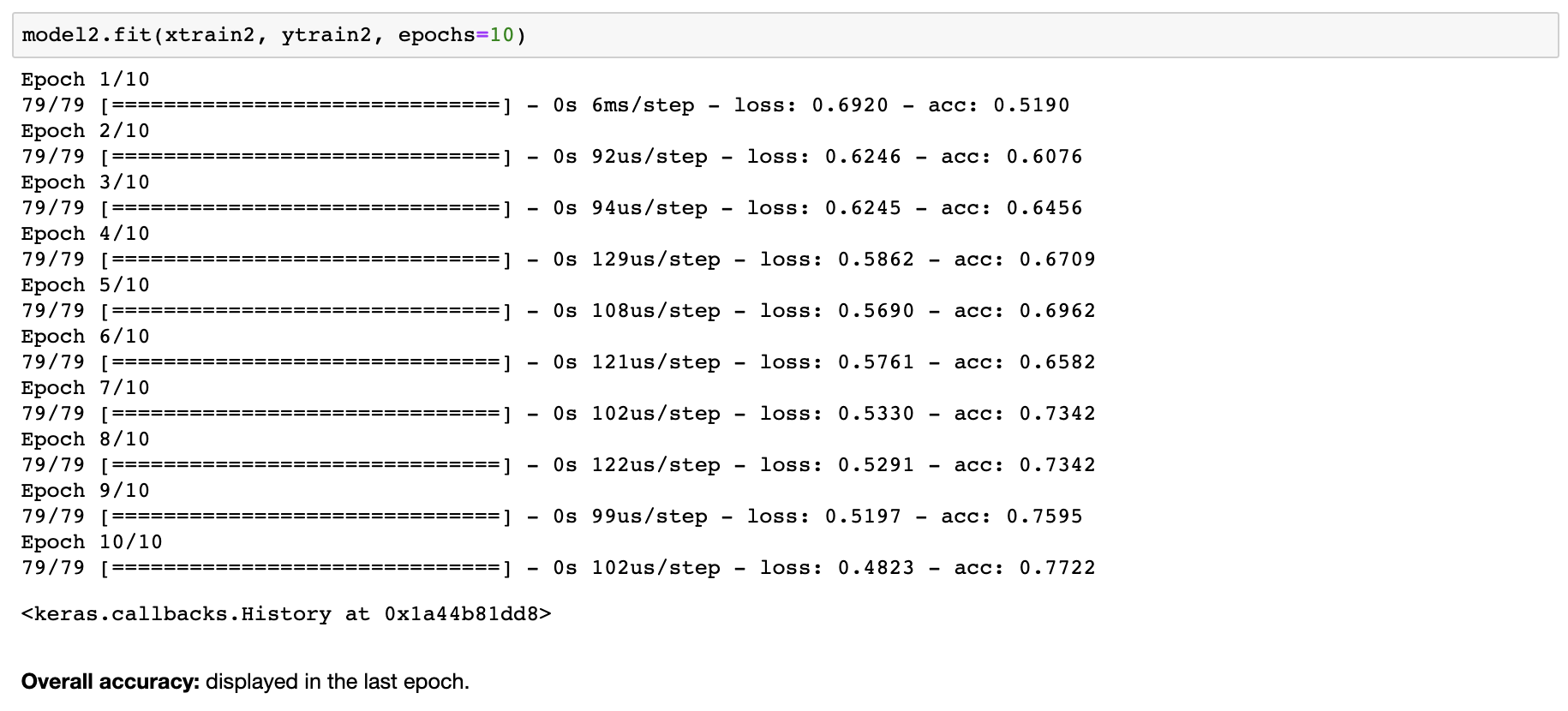
4.2 ARTIFICIAL NEURAL NETWORK

We constructed 4 different models with three activation functions used primarily: sigmoid, softmax and tanh. From the following snapshots, we can conclude that sigmoid resulted in the highest accuracy for our particular model:

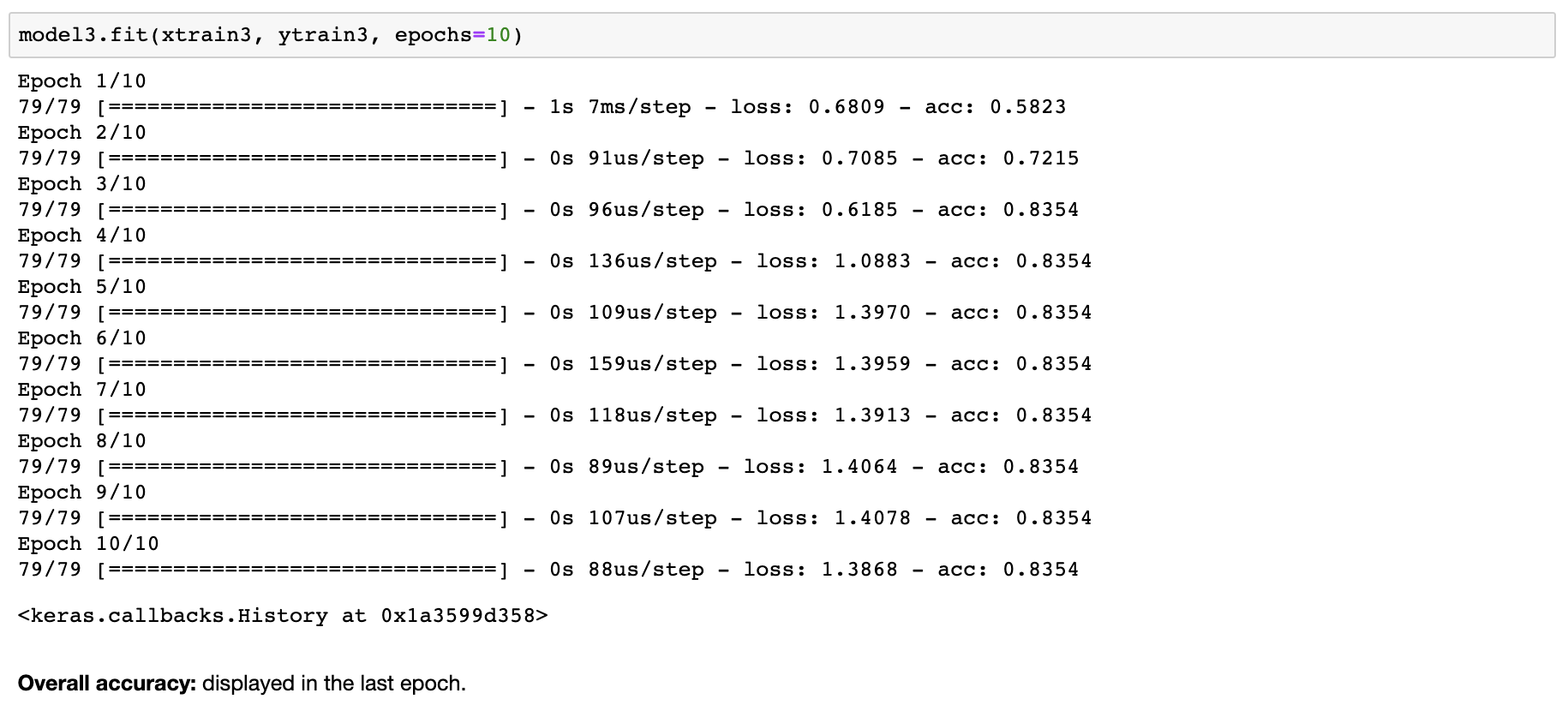
1. Predicting if an animal is venomous: Using the sigmoid activation function, our model uses data from every column except for ‘venomous’ itself to predict if an animal is venomous based on all the other traits. The accuracy turns out to be 92.41%



1. Predicting if an animal is a predator: Using the softmax activation function, our model uses data from every column except for ‘predator’ itself to predict if an animal is a predator based on all the other traits. The accuracy turns out to be 77.22%



1. Predicting if an animal is domestic: Using the tanh activation function, our model uses data from every column except for ‘domestic’ itself to predict if an animal is domestic based on all the other traits. The accuracy turns out to be 83.54%

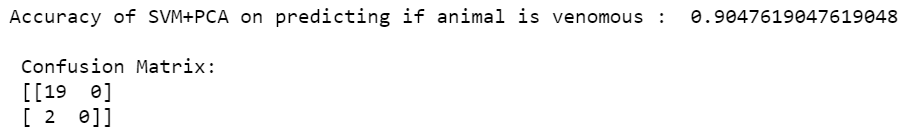


1. Predicting which class an animal belongs to: Refer to our code for more details. Here, we predict which class type an animal belongs to and measure the accuracy of prediction, which turns out to be 92.5% (the accuracy is the only output of this model).

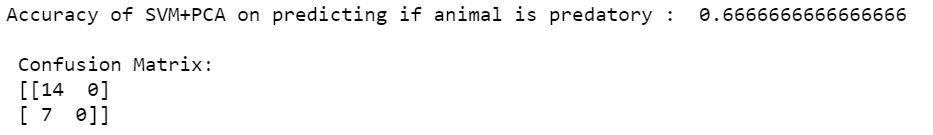
4.3. SVM and PCA

We built 3 different models under this approach for predicting each of the animals being venomous,domestic and predatory. Studies show that SVM yields improved results on being combined with PCA.Since our initial dataset consisted of 18 columns, we used PCA to reduce dimensionality and later predicted using SVM.

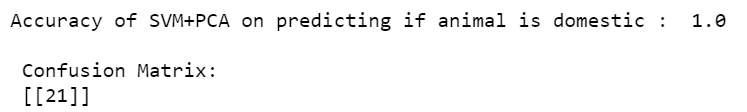
1. Predicting if an animal is venomous: Using the aforementioned approach,we drop the column named ‘venomous’, predict and later verify the prediction against the dropped column.The accuracy shows to be 90.47% for the same.



1. Predicting if an animal is a predator: Using the aforementioned approach,we drop the column named ‘predator’, predict and later verify the prediction against the dropped column.The accuracy shows to be 66.66% for the same.



1. Predicting if an animal is domestic: Using the aforementioned approach,we drop the column named ‘domestic’, predict and later verify the prediction against the dropped column.The accuracy shows to be 100% for the same.



APPLICATIONS OF NEURAL NETWORKS

* With the help of neural networks, we can find the solution of such problems for which an algorithmic method is expensive or does not exist.
* Neural networks can learn by example, hence we do not need to program it to much extent.
* Neural networks have accuracy and significantly faster speed than conventional speed.
* Speech recognition
* Character recognition
* Human face recognition
* Machine learning models for predicting personalised content in Netflix, Spotify etc for each user.

APPLICATIONS OF SVM

* Face Detection
* Text and hypertext classification
* Image Classification
* Bioinformatics
* Protein fold and remote homology detection
* Handwriting Recognition
* Generalized Predictive Control (GPC)

**V. SUMMARY and CONCLUSION**

1. The dataset consists of 100 animals, majority of which belong to the class Mammals(40.6%). The other classes present in order are Bird (19.8%), Fish(12.9%), Invertebrates(9.9%), Bugs(7.9%), Reptiles(5.0%) and Amphibians(4%).
2. The highest correlation is seen between the attributes ‘backbone’ and ‘toothed’. The least correlation (less than 0) is seen between attributes ‘airborne’ and ‘toothed’. It can be inferred that most toothed animals have a backbone and are not airborne.
3. Prediction of certain animal traits such as ‘venomous’, ‘domestic’ and ‘predator’ is done using various artificial neural network models and SVM models. Sigmoid function results in the highest accuracy amongst the three activation functions used, namely sigmoid, softmax and tanh.
4. The ANN models are observed to make better predictions for 2 of the 3 characteristics considered namely ‘venomous’ and ‘predator’
5. Classification of animals into biological classes based on their physical features is done using a multilayer perceptron model and accuracy achieved is 92.5%.

A quick summary of the accuracies obtained from both models (ANN and SVM,PCA) is as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| Characteristic vs Model | **Venomous** | **Predator** | **Domestic** |
| **ANN** | 92.41% | 77.22% | 83.54% |
| **SVM+PCA** | 90.47% | 66.66% | 100% |

Without linear algebra, you cannot represent data as vectors and tensors. Only by using LA’s tools can we utilise such things and make functions and formulae out of which we can further build complex structures for housing data. Everything in today’s era runs on the ability to both efficiently store data as well as intelligently utilise it. Neural networks, machine learning and artificial intelligence (which is the sphere of robotics) exist solely because of linear algebra. It is fascinating, thus, to reflect on how almost every technical, electronic and electrical product that we use in our daily lives is based on linear algebraic concepts.

FUTURE SCOPE

Classification and predictive analysis of data has always been an integral cog of the present day industries. The hunt now is however for a model that provides maximum accuracy and best results. Our project forms the ground basis for all such applications. The extension of the project can include profiling and predicting everything from characteristics of students or employees or even criminals to the nature of stocks,business strategies and weather forecast.

**VI. BIBLIOGRAPHY**

1. “Neural Network Architectures and Learning” - Paper by Bogdan M Wiliamowski
2. “Neural Network Design” - Book by Hagan, Demuth & Beale
3. “Machine Learning” - Book by Thomas G. Dietterich
4. “Searching for Activation Functions” - Paper by Ramachandran, Zoph & Lee
5. “Data Classification using Support Vector Machine”- Paper by Durgesh Srivastava &Lekha Bhambu
6. “Principal Component Analysis” - Book by I.T Jolliffe