國立雲林科技大學

資訊管理研究所

機器學習專案作業一

指導教授：　許中川　教授

學生：M11223038 陳品佑

M11223032 張祥恩

目錄

表目錄

圖目錄

摘要

本次專案作業主要目標是透過 Python 語言以及 Keras & Tensorflow 框架時做前饋式神經網路，並進行類別與數值的預測。首先是本課程當中介紹的 MINIST 和 Boston Housing Price 資料集，再者是於 UCI ML Repository 當中取得 Adult 資料集，以這三份資料集來做分類以及迴歸預測。

而實驗結果顯示，在不同超參數的排列組合下，部分資料集對於分類、迴歸模型的績效都有著較明顯的差異，而小部分資料集像是房價預測在各個超參數下其不同績效指標的水準表現都旗鼓相當，由此可以推論出這份資料集的特性可能較少極端離群值，以致於讓績效表現持平。

關鍵字：MINIST、Boston Housing Price、Adult、超參數

Abstract

The main objective of this project assignment is to implement feedforward neural networks using Python language and the Keras & Tensorflow framework for both classification and regression tasks. Initially, the MNIST and Boston Housing Price datasets introduced in the course are utilized, followed by the Adult dataset obtained from the UCI ML Repository, for classification and regression predictions.

The experimental results reveal significant differences in the performance of classification and regression models across various combinations of hyperparameters for some datasets. However, for a small portion of datasets, such as the housing price prediction, the performance metrics remain consistently comparable across different hyperparameter settings. This suggests that these datasets may have fewer extreme outliers, resulting in consistent performance levels.

Keywords: MINIST, Boston Housing Price, Adult, hyperparameters

一、緒論

1.1 研究動機

本研究在資料探勘領域中的分類以及迴歸預測上實現了許多不同演算法像是隨機森林和決策樹，並比較彼此之間的績效。然而本研究此次著重在透過神經網路建立模型，以實驗來練習神經網路的架構並了解如何應用，並說明實驗結果。

1.2 研究目的

本研究在於利用 Python 之 Keras 和 Tenorflow 框架實現神經網路的建模，透過機器學習的方式進行資料的分類以及迴歸預測。由此本組也將有機會學習如何處理不同類型的資料，包含圖像資料、數值型資料或混和型的資料型態，從而擴展本組在資料工程、數據分析方面的應用能力。

本研究以不同的性能評估指標來評估模型的表現能力像是 Precision, Recall, F1- score 等，使本組深入理解其代表之涵義。而在透過調整模型超參數，也讓本研究能夠更有機會深入理解這些參數在模型性能上的影響，從而優化本研究模型以提高分類、預測能力。

二、研究方法

本研究利用課程中提供的 MINIST 和 Boston Housing Price 資料集與程式碼進行資料分析。在資料前置處理部分本組將先做缺失值處理，接著使用 One Hot Encoding 對類別特徵進行編碼，轉換為模型能夠處理的數值型態。再來則是使用 Standard Scaler 對數值型資料進行標準化，使其平均值為 0，標準差為 1。最後做模型的分類、迴歸預測，以及不同超參數調整下之績效比較。在分類的預測上使用 Precision, Recall, F1-score；對於迴歸預測則為平均絕對誤差(MAE)、平均絕對百分比誤差(MAPE)、均方根誤差(RMSE)。每一項實驗的結果以表格的形式呈現。

三、研究實驗

3.1 資料集簡介

首先關於資料集的部分為本研究在課程當中使用教授所提供的(1) MINIST，這是一個被廣泛應用於電腦視覺領域的手寫數字圖像資料集，下圖 MNIST 中包含了上萬張的手寫數字 0 到 9，10 個數字圖像以及每一張圖像正確的標籤(Label)每個數字都有 6000 個訓練樣本和 1000 個測試樣本，共計 7000 個圖像樣本。下表 1"Adult"則是在人口普查收入的數據，也是一份在資料探勘、機器學習領域的實作上經常使用到的資料集，其資料之內容摘要細節由下圖 1 至 4 所示；表 2 是 Boston Housing Price 預測房價資料集與程式碼；下圖 5, 6 為房價分布之可視化圖表。而以下為本次實驗用資料集的屬性以及部份內容。

- 資料集名稱：Adult dataset
- 資料筆數：32560 筆資料
- 屬性數量：14 種屬性

表 1 人口普查資料集 Adult 屬性簡介

屬性名稱	型態	尺度
age	Int	Ordinal
workclass	String	Nominal
fnlgwt	Int	Nominal
education	String	Ordinal
educationnum	Int	Ordinal
marital-status	String	Nominal
occupation	String	Nominal
relationship	String	Nominal
race	String	Nominal
gender	String	Nominal
country	String	Nominal
capital-gain	Int	Ratio
hours-per-week	Int	Ordinal
income	String	Ordinal

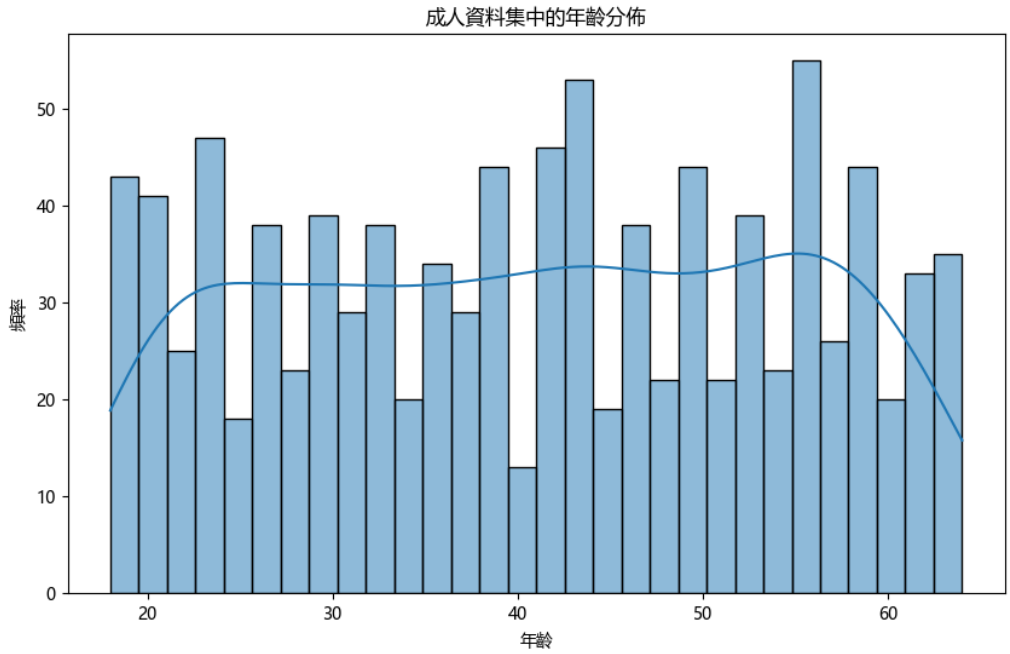

圖 1 年齡分佈圖

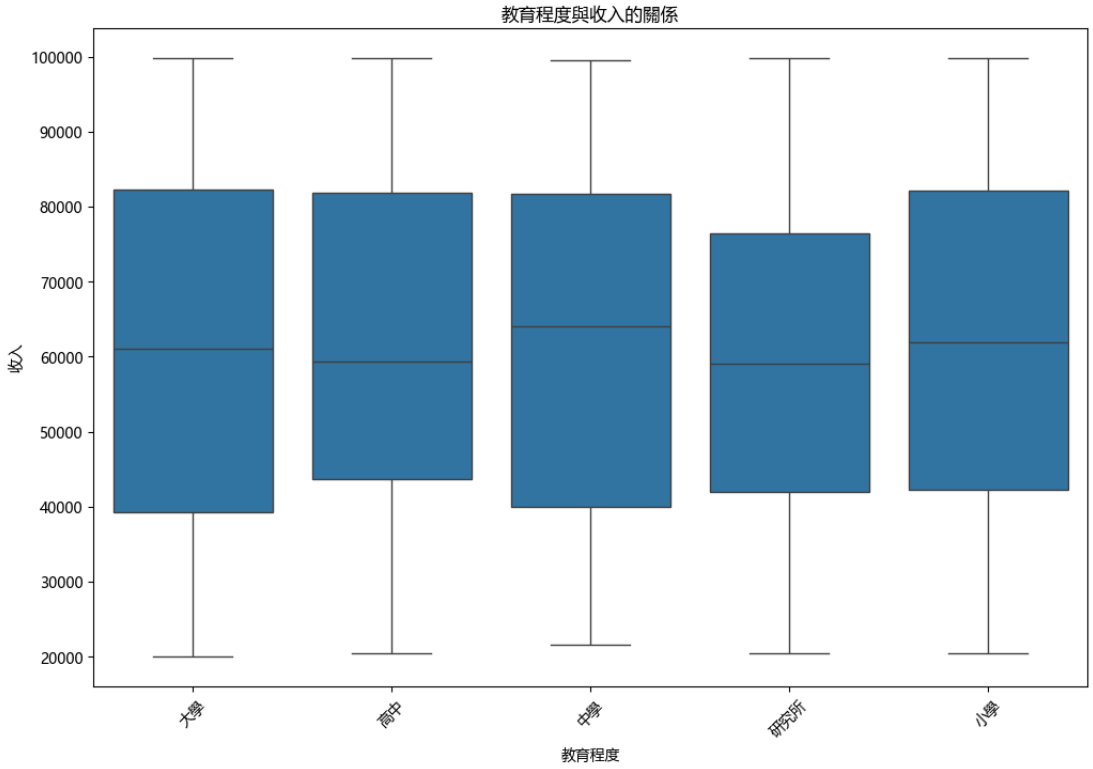

圖 2 教育程度與收入之關係圖

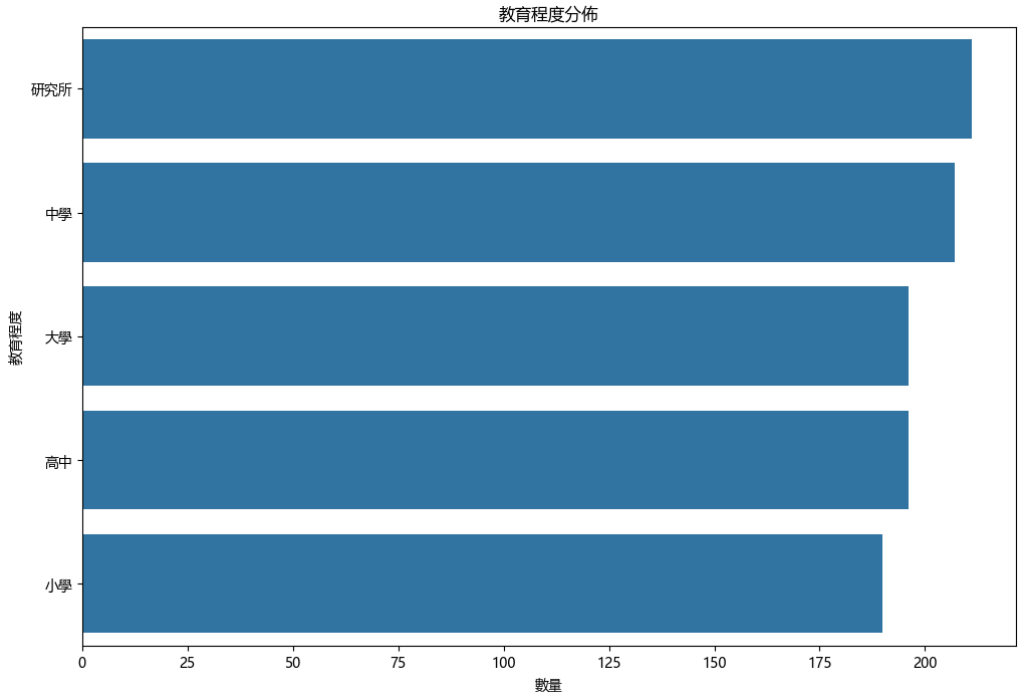

圖 3 教育程度分佈圖

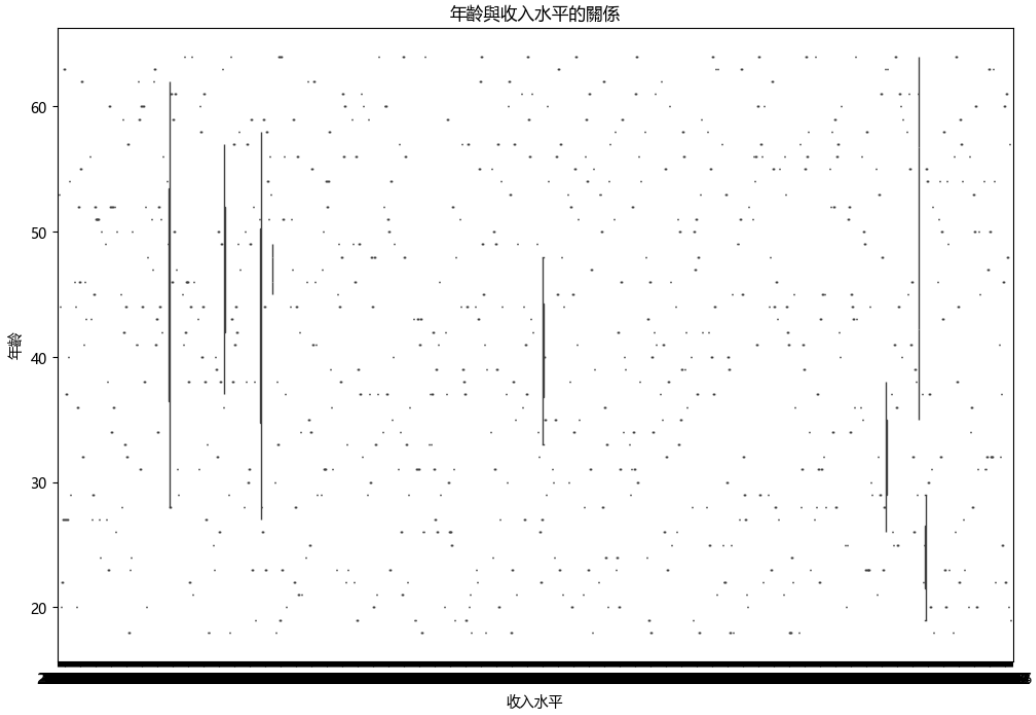

圖 4 年齡與收入水平關係圖

- Boston Housing Dataset
- 資料集名稱：Boston Housing
- 資料筆數：506
- 屬性數量：14

表 2 Boston Housing 資料集屬性簡介

屬性名稱	型態	尺度
crim	Float	Ratio
zn	Float	Ratio
indus	Float	Ratio
chas	Int	Nominal
nox	Float	Ratio
rm	Float	Ratio
age	Float	Ratio
dis	Float	Ratio
rad	Int	Interval
tax	Int	Ratio
ptratio	Float	Interval
b	Float	Ratio
lstat	Float	Ratio
medv	Float	Ratio

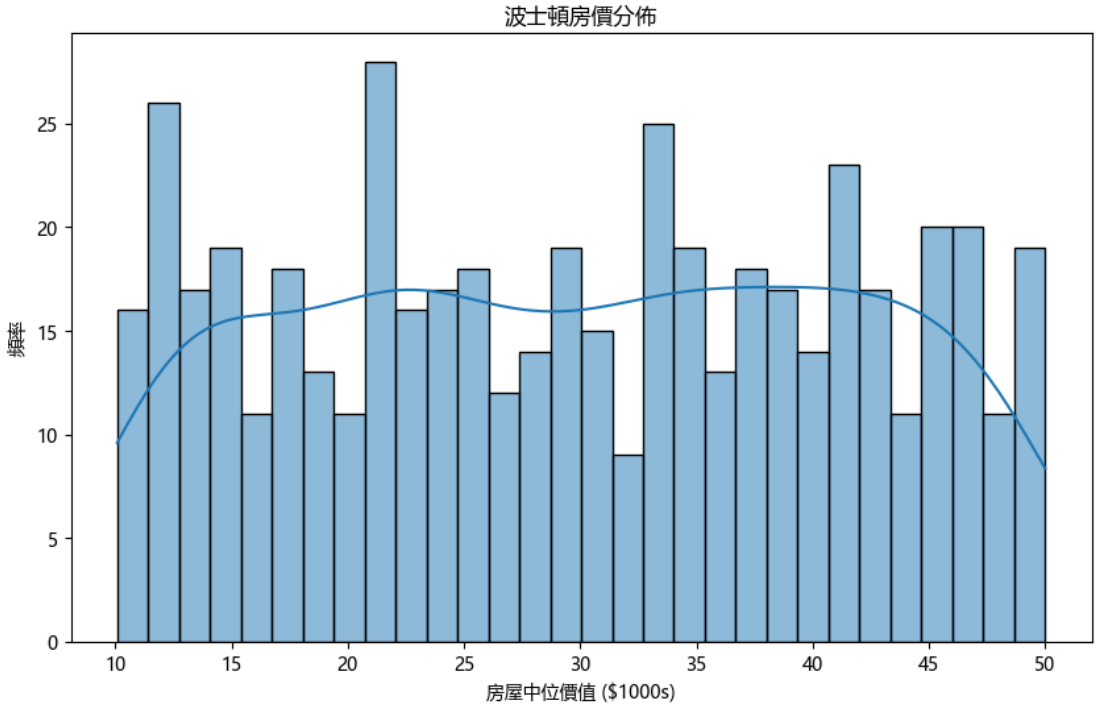

圖 5 波士頓房價分佈圖

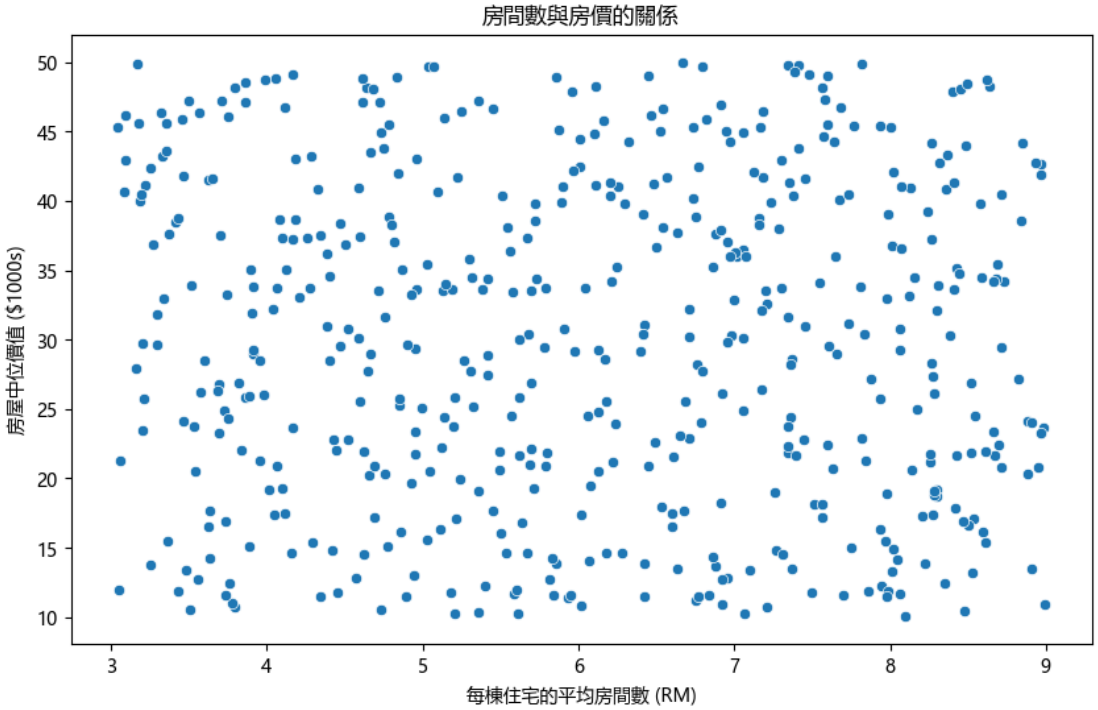

圖 7 房間數與房價之關係圖

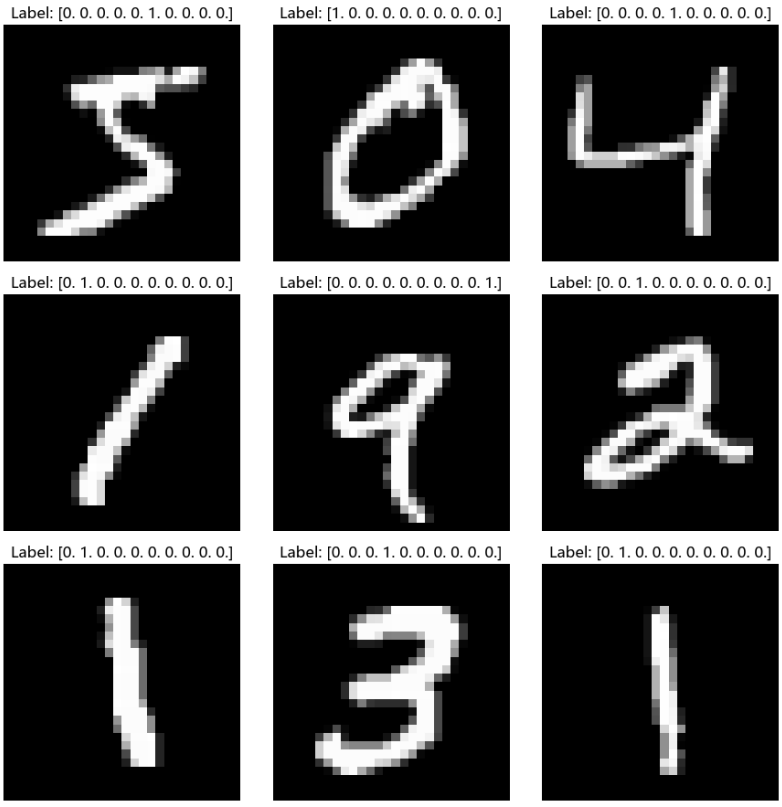

圖 6 MINIST 資料集

3.2 前置處理

在進行模型訓練之前，本研究對資料集進行處理，包括缺失值處理、移除和類別特徵編碼；使用 One Hot Encoding 對類別特徵進行編碼，轉換為模型可以處理的數值型數據。以及使用 Standard Scaler 對數值型數據進行標準化處理，使其均值為 0，標準差為 1。

3.3 實驗設計

本研究著重在實作練習神經網路的模型建立以及對超參數(Hyperparameter)的不同設定下，比較其績效表現。下面會針對類神經網路之超參數的設計做說明，包含輸入層(Input layer)、隱藏層(Hidden layer)及單元數、激活函數(Activation function)、優化器(Adam)、批次大小(Batch size)、訓練週期(Epoch)。

在本研究中，為探究不同價購及超參數對於預測績效之影響，本組將變化隱藏層數量，並比較 1、2 層時的模型性能；此外在固定隱藏層數量的情況下改變每層的單元數(例如：32, 64, 128)。以及比較不同批次大小(例如：16、32、64)。回合數部分則比較不同訓練週期數(例如：10、20、30)對模型性能的影響。此外，在隱藏層中使用不同的激活函數，隱藏層使用 ReLU；輸出層對分類使用 Sigmoid。

下一小節本組將談到本實驗所使用之績效評估指標，以及公式的呈現以及介紹。

3.3.1 績效評估指標

本研究在分類預測中以精確度(Precision)、召回率(Recall)、以及精確度與召回率的調和平均數(F1-Score)；以下為針對分類預測問題的績效指數公式。

$$Precision = \frac{TP}{(TP + FP)}$$

$$Recall = \frac{TP}{(TP + FN)}$$

$$F1\_Score = \frac{2}{(\frac{1}{Precision} + \frac{1}{Recall})}$$

而對於迴歸的預測則使用平均絕對誤差(Mean Absolute Error, MAE)、平均絕對百分比誤差(Mean Absolute Percentage Error, MAPE)、均方根誤差(Root mean square error, RMSE)。以下為本文使用之績效指標公式。

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \times 100\%$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}$$

3.4 實驗結果

首先本研究先以 MINIST 以及 Boston Housing Price 資料集之時訓練結果做討論，由下表 3, 4 可知，在 MINIST 分類預測上，當 epoch, batch sizes 參數改變時，對於分類績效有著微幅的影響。由本實驗可以發現到，無論是在回合數提高還是批次數量提高的情況下，對於測試資料都有著上升的趨勢，而在訓練資料上更是近乎完全分類正確的結果。

表 3 MINIST Hyper paraneter - Training Dataset

Training Data						
隱藏層數量	神經元數量	Epoch	Batch Size	Precision	Recall	F1-score
1	32	5	64	0.9663	0.9663	0.9665
			128	0.9601	0.9600	0.9601
		10	64	0.9761	0.9760	0.9761
			128	0.9708	0.9708	0.9708
2		5	64	0.9705	0.9705	0.9705
			128	0.9628	0.9627	0.9629
		10	64	0.9832	0.9832	0.9832
			128	0.9745	0.9744	0.9746
3		5	64	0.9718	0.9718	0.9720
			128	0.9653	0.9652	0.9656
		10	64	0.9791	0.9791	0.9793
			128	0.9784	0.9784	0.9785
1	64	5	64	0.9813	0.9813	0.9813
			128	0.9774	0.9774	0.9774
		10	64	0.9872	0.9872	0.9874
			128	0.9865	0.9865	0.9865
2		5	64	0.9810	0.9811	0.9813
			128	0.9768	0.9768	0.9771
		10	64	0.9904	0.9904	0.9905
			128	0.9909	0.9909	0.9909
3		5	64	0.9833	0.9833	0.9834
			128	0.9803	0.9803	0.9805
		10	64	0.9910	0.9910	0.9911
			128	0.9920	0.9920	0.9920

表 4 MINIST Hyper paraneter - Testing Dataset

Testing Data						
隱藏層數量	神經元數量	Epoch	Batch Size	Precision	Recall	F1-score
1	32	5	64	0.9627	0.9626	0.9626
			128	0.9543	0.9542	0.9541
		10	64	0.9679	0.9678	0.9678
			128	0.9640	0.9639	0.9639
2		5	64	0.9619	0.9619	0.9619
			128	0.9558	0.9553	0.9552
		10	64	0.9705	0.9704	0.9704
			128	0.9640	0.9638	0.9638
3		5	64	0.9664	0.9663	0.9663
			128	0.9574	0.9569	0.9568
		10	64	0.9668	0.9662	0.9662
			128	0.9646	0.9644	0.9644
1	64	5	64	0.9714	0.9713	0.9713
			128	0.9693	0.9693	0.9693
		10	64	0.9736	0.9734	0.9734
			128	0.9726	0.9725	0.9725
2		5	64	0.9725	0.9723	0.9723
			128	0.9687	0.9682	0.9681
		10	64	0.9749	0.9747	0.9747
			128	0.9761	0.9761	0.9761
3		5	64	0.9717	0.9715	0.9715
			128	0.9688	0.9685	0.9685
		10	64	0.9744	0.9742	0.9742
			128	0.9746	0.9746	0.9746

接著本實驗對 Boston Housing Price 房價預測之資料及進行神經網路訓練，激活函數分別使用 ReLU 以及 Sigmoid 測試其表現能力，其訓練、測試結果及績效如下表 5, 6 所示。由表格呈現結果可以得知，在任一神經元數量、隱藏層數量、激活函數之設定如何，大部分誤差最小的超參數設定為(Epoch, Batch_Size) = (150, 16)。

表 5 Boston Housing Price Hyper parameter - Training Dataset

Training Data							
隱藏層數量	神經元數量	Activation Function	Epoch	Batch Size	MAE	MAPE	RMSE
1	32	ReLU	50	16	18.9776	16.1307	4.3563
				32	33.6664	21.6405	5.8023
				64	108.8827	37.4655	10.4347
			100	16	10.3421	11.5466	3.2159
				32	20.5586	16.8121	4.5342
				64	29.2571	19.7311	5.4090
			150	16	8.6829	10.7819	2.9467
				32	15.0994	14.2506	3.8858
				64	21.2922	16.7615	4.6144
2	32		50	16	9.7544	11.1048	3.1232
				32	10.7898	11.4329	3.2848
				64	15.2220	14.6801	3.9015
			100	16	7.1662	9.7390	2.6770
				32	8.6489	10.6098	2.9409
				64	10.5579	11.6051	3.2493
			150	16	5.5546	8.4957	2.3568
				32	6.8736	9.4019	2.6218
				64	8.1677	10.4639	2.8579
1	64		50	16	13.8682	13.6498	3.7240
				32	24.5349	17.8818	4.9533
				64	47.6653	24.7208	6.9040
			100	16	8.7351	10.5770	2.9555
				32	13.4335	13.2142	3.6652
				64	23.5659	17.9021	4.8545
			150	16	7.0791	9.4406	2.6607
				32	9.8335	11.1914	3.1358
				64	15.1901	13.8859	3.8975

續下表

呈上表

2	64	Sigmoid	50	16	7.5022	9.8658	2.7390
				32	9.1506	11.0016	3.0250
				64	12.1886	11.9138	3.4912
			100	16	5.5184	8.4958	2.3491
				32	6.3458	9.0801	2.5191
				64	8.5804	10.4792	2.9292
			150	16	3.5315	7.0366	1.8792
				32	5.2522	8.2671	2.2918
				64	6.3937	8.8710	2.5286
1	32		50	16	41.1263	15.6681	6.4130
				32	174.7679	48.4941	13.2200
				64	382.5747	78.9179	19.5595
			100	16	26.8228	16.1575	5.1791
				32	37.8621	15.6611	6.1532
				64	139.6441	41.3743	11.8171
			150	16	21.2551	14.9941	4.6103
				32	31.4363	16.7106	5.6068
				64	62.9016	20.3083	7.9311
2	32		50	16	61.1069	20.2163	7.8171
				32	97.5443	25.2743	9.8765
				64	198.9509	43.3721	14.1050
			100	16	39.5331	17.4922	6.2875
				32	60.8825	20.4427	7.8027
				64	95.4311	24.4969	9.7689
			150	16	16.7578	12.5678	4.0936
				32	40.8567	18.7107	6.3919
				64	67.0762	20.2002	8.1900

表 6 Boston Housing Price Hyper parameter - Testing Dataset

Testing Data							
隱藏層數量	神經元數量	Activation Function	Epoch	Batch Size	MAE	MAPE	RMSE
1	32	ReLU	50	16	4.0237	21.7464	5.1358
				32	4.9639	27.0518	6.2734
				64	9.2561	42.9021	10.8668
			100	16	3.2204	16.7868	4.8836
				32	4.0690	22.2099	5.2578
				64	4.8037	26.0783	6.0910
			150	16	3.0313	15.3835	4.4642
				32	3.5837	19.0575	4.7269
				64	3.9844	21.3929	5.0061
2	32		50	16	3.0940	16.1160	4.5885
				32	3.1423	16.3869	4.7793
				64	3.7333	19.9323	5.0637
			100	16	2.8957	14.6782	4.4389
				32	3.0246	15.2589	4.6098
				64	3.2783	16.7625	4.7103
			150	16	2.6198	13.6320	4.0028
				32	2.9670	14.9856	4.8601
				64	2.9929	15.1487	4.6054
1	64		50	16	3.4304	18.1566	4.6712
				32	4.2386	23.6606	5.4151
				64	5.8604	30.3642	7.4462
			100	16	2.9055	14.7756	4.2994
				32	3.5946	19.1098	4.8521
				64	4.3929	24.2085	5.4648
			150	16	2.7512	14.5151	4.3724
				32	3.0172	15.5384	4.3594
				64	3.6431	19.5624	4.8104

續下表

呈上表

2	64	Sigmoid	50	16	2.9725	15.0040	4.8060
				32	3.0863	15.7367	4.6318
				64	3.3189	16.6926	4.6580
			100	16	2.6538	13.0456	4.2195
				32	2.7227	13.6889	4.3352
				64	3.0367	15.3980	4.6500
			150	16	2.4434	12.4005	4.0111
				32	2.6235	13.5619	4.1762
				64	2.7577	13.6930	4.4576
1	32		50	16	3.9652	16.5208	6.2081
				32	11.6217	48.5237	13.4913
				64	18.3654	78.8408	20.0401
			100	16	3.5868	16.8253	5.0624
				32	3.8792	16.5465	5.9330
				64	10.0179	41.1425	11.9909
			150	16	3.4693	16.4653	4.8569
				32	3.7256	17.3422	5.3893
				64	5.4198	21.2453	7.8786
2	32		50	16	5.0646	22.5228	7.7519
				32	7.3231	30.8755	10.1131
				64	12.0454	47.5425	14.5728
			100	16	4.0855	18.7052	6.2060
				32	5.0626	22.6369	7.7346
				64	7.1489	29.7853	9.9893
			150	16	3.2541	16.2117	4.7921
				32	4.2443	20.1027	6.2693
				64	5.3304	22.9989	8.1955

而本實驗最後一階段則是針對人口普查資料集”Adult dataset”做神經網路訓練，做分類預測、迴歸預測及其績效評估。本組將激活函數設為 ReLU，並比較在不同超參數下之績效表現，如下表所示可知，在激活函數為 ReLU 之下，無論神經元、隱藏層數、回合數、批次大小設定為何，其績效表現皆相當，並無太大之差異，導致此現象的原因本組認為可能在資料特性上，數據分布較為均勻少有極端離群值出現，使得不同超參數設定下計算出來的績效相近。同樣地在下表分類預測上的表現也並不是特別優異。

表 7 Adult Hyper parameter Regression Predict - Training Dataset

Training Data							
隱藏層數量	**神經元數量**	**Activation Function**	**Epoch**	**Batch Size**	**MAE**	**MAPE**	**RMSE**
1	32	ReLU	10	32	40.2217	99.4543	42.0519
				64	40.2134	99.4217	42.0447
			20	32	40.2001	99.3941	42.0311
				64	40.2047	99.4092	42.0344
	64		10	32	40.1502	99.2211	41.9846
				64	40.1836	99.3231	42.0175
			20	32	40.2123	99.4246	42.0434
				64	40.2113	99.4226	42.0423
2	32		10	32	40.1521	99.2466	41.9841
				64	40.2066	99.4098	42.0378
			20	32	40.1965	99.3844	42.0266
				64	40.2397	99.5003	42.0719
	64		10	32	40.2119	99.4289	42.0420
				64	40.2368	99.4875	42.0688
			20	32	40.1973	99.3854	42.0277
				64	40.1811	99.3300	42.0132
3	32		10	32	40.2068	99.4102	42.0375
				64	40.1545	99.2519	41.9872
			20	32	40.2150	99.4322	42.0464
				64	40.1920	99.3657	42.0231
	64		10	32	40.2084	99.4074	42.0397
				64	40.2252	99.4636	42.0559
			20	32	40.2035	99.4055	42.0333
				64	40.1756	99.3313	42.0056

表 8 Adult Hyper parameter Regression Predict - Testing Dataset

Testing Data							
隱藏層數量	神經元數量	Activation Function	Epoch	Batch Size	MAE	MAPE	RMSE
1	32	ReLU	10	32	40.1799	99.4571	42.0511
				64	40.1720	99.4298	42.0439
			20	32	40.1594	99.4049	42.0310
				64	40.1638	99.4180	42.0339
	64		10	32	40.1091	99.2317	41.9842
				64	40.1419	99.3252	42.0169
			20	32	40.1724	99.4304	42.0444
				64	40.1706	99.4276	42.0428
2	32		10	32	40.1114	99.2539	41.9841
				64	40.1671	99.4199	42.0390
			20	32	40.1565	99.3923	42.0277
				64	40.2006	99.5094	42.0740
	64		10	32	40.1721	99.4352	42.0432
				64	40.1974	99.4980	42.0703
			20	32	40.1572	99.3854	42.0292
				64	40.1415	99.3351	42.0151
3	32		10	32	40.1668	99.4198	42.0384
				64	40.1142	99.2623	41.9877
			20	32	40.1742	99.4372	42.0469
				64	40.1530	99.3703	42.0254
	64		10	32	40.1683	99.4145	42.0405
				64	40.1849	99.4724	42.0565
			20	32	40.1642	99.4106	42.0354
				64	40.1359	99.3392	42.0073

表 9 Adult Hyper parameter Classification Predict

(Epoch, Batch_size)	Precision	Recall	F1-score
(5, 32)	0.7190	0.6261	0.6694
(5, 64)	0.7248	0.6152	0.6655
(10, 32)	0.7111	0.6279	0.6669
(10, 64)	0.7445	0.5962	0.6621

四、結論

在這份實驗作業中，我們通過構建類神經網路模型，使用 Python 語言和 Keras & Tensorflow 框架，對各個資料集各別進行迴歸預測、分類預測，並使用分類指標：Precision、Recall、F1-score；以及迴歸預測指標 MAE、MAPE 和 RMSE 等對模型的性能來進行比較。在實驗過程中，我們學到了如何應用神經網路之超參數設定，來讓不同組合搭配下可以使得模型能夠產生出更好的結果。並學習了如何對模型的性能進行評估和分析。這將有助於學生在未來的機器學習領域中應用和實戰所學的知識以及技能。

五、參考文獻

HO-HSUN (2017)前饋式神經網路 https://ithelp.ithome.com.tw/articles/10194255

Johnny. Deep Learning 基本功：認識 MNIST 資料集與損失函數。 https://datasciocean.tech/deep-learning-core-concept/mnist-dataset-and-cost-function/#google_vignette

LUFOR129 (2020). Tensorflow、Keras 傻瓜式安裝教學。 https://lufor129.medium.com/%E5%82%BB%E7%93%9C%E5%BC%8Ftensorflow-keras%E5%AE%89%E8%A3%9D%E6%95%99%E5%AD%B8-730b235275d

Ann. (2021). Activation Functions — Sigmoid & ReLu & tahn & LeakyReLu & ELU. https://medium.com/@adea820616/activation-functions-sigmoid-relu-tahn-20e3ae726ae

nancysunnn (2021). 回歸模型的衡量標準：MSE. RMSE. MAE. MPE。 https://ithelp.ithome.com.tw/articles/10274551

1240117300(2020). Pytorch+MINIST 實現手寫數字識別。 https://github.com/1240117300/MINIST/blob/master/test2.py

Hunter-P (2019). tensorflow-minist. https://github.com/Hunter-P/tensorflow-minist