Comp551 - Assignment 1

Members:
Charlotte Livingston (261043465),
Adam Dufour(261193949),
Gabriel Caballero (261108565)

September 30th 2025

Abstract

In this work, two different machine learning models have been studied: Linear Regression (LinReg) and Logistic Regression (LogReg), while also using mini-batch stochastic gradient descent (mb-SGD). The team was tasked with predicting key targets in two different datasets: one that investigated the diagnosis of Parkinson's disease (PD) and the other breast cancer (BC). The findings showed that the models predict these diseases with overall high accuracy. By comparing the coefficient weights, it was possible to determine for each instance which characteristics are deemed more important to the model. The team assessed the robustness of the models across different train/test splits. Implementing mini-batch stochastic gradient decreased computational burden of the models while maintain comparable performance. A comprehensive analysis of hyperparameter batch size and learning rate allowed for the training of two highly accurate and fast models. Ultimately, the team was able to identify the optimal model for diagnosis prediction in the PD and BC data set and obtain the following results. For PD, the linear regression model with polynomial features gave $R^2 = 0.94$ and RMSE = 1.91. The basic logistic regression model gave Accuracy of 0.95, Recall = 0.94 and F1 Score = 0.94.

Introduction

Two datasets were used to train and test the machine learning models. For the LinReg model, the Oxford Parkinson's Disease Telemonitoring Dataset was selected. This dataset connects speech features to the progression of Parkinson's Disease (PD) using analysis of voice recordings. It has been previously found that jittering and shimmering features are most correlated with PD progression[3]. The target of this dataset is the UPDRS, a score given by clinicians to assess the severity of the disease in patients. This score is continuous and there is a known linear correlation between a PD patient's speech and their UPDRS, making the dataset ideal for testing a linear model [4] [1]. The goal of this work is to predict the UPDRS based on the features of the patient's speech and to confirm the correlation of jittering and shimmering with the score. For the LogReg model, the Diagnostic Wisconsin Breast Cancer Dataset was used. This dataset connects the malignancy state of BC tumors and tumor features, with image metrics such as the fractal dimensions identified as being the most informative features[2] [5]. The target is a categorical value (malignant or benign), making it ideal for testing LogReg as a classifier. The goal of this work is to predict the diagnosis based on the features of the tumor and to confirm the correlation of the image metrics with diagnosis.

Datasets

In this section, the data sets used in the assignment will be described in order to understand the data and objectives of the developed models. Since these data sets contains biometrics from patients, anonymity, and ethical processes are crucial to collect and share the data.

Parkinson Disease Dataset [3]

The PD dataset used in this work compiled 5875 instances obtained from speech recording from 42 PD patients. In total, 19 features and 2 targets are contained in the data table. In this work, the data set was formatted and treated before being used to evaluate the performance of the linear regression models. The data set was stripped from any missing values or invalid data type and outliers (z-score$\geq$ 3) before being scaled between (mean = 0, std = 1) and one-hot encoded. The columns related to ID were dropped and the features and targets were divided. The data were then plotted for every feature across each target in order to obtain a sense for the potential early correlation between the features and targets. In addition, the features : age, sex and test time were plotted in order to have a better idea on the testing protocol (ie. sex distribution, age

tranches and times of recording speech). In order to have a numerical view on the features, the mean, standard deviation, minimum and maximum values were calculated. It was also decided to not drop the column total UPDRS as without it the performance was mediocre.

Breast Cancer Dataset [2]

The BC dataset studied in this document contained a total of 569 instances of BC tumors in patients. The set contains a total of 30 features and 1 target which is the ultimate diagnosis of the patient's tumor, malign or benign. The dataset was filtered and constructed in order to make it compatible with the logistic regression model. First of all, the set was stripped of any missing values, outliers or invalid data point, scaled and one-hot encoded as described in the PD data set section. As above, the data was then plotted and metrics to understand data distribution were calculated.

Results

The machine learning models were implemented and their performance was tested on the cleaned datasets. All linear regression models were evaluated using a random 80/20 train/test split of the PD dataset and all logistic regression models were evaluated using the same split of the BC dataset, unless otherwise stated.

Linear Regression

As shown in Table 1, the analytical LinReg model testing showed a high R^2 score and a low root mean square error (RMSE), suggesting the overall success of the model's implementation.

Metric	Train Set	Test Set
RMSE (LinReg)	3.26	3.21
R^2 (LinReg)	0.91	0.91
F1 Score (LogReg)	0.988	0.964
Accuracy (LogReg)	0.991	0.973

Table 1: Performance of Linear and Logistic Regression models on Train and Test sets for an 80/20 split

Further analysis of LinReg weights post fitting revealed that the jitter features and NHR had the largest magnitude coefficient values (Figure 1). In contrast, demographic variables such as age, test time, and sex had lower weights. Analyzing Pearson correlation between features, Figure 2 shows that the jitter and shimmer features are strongly positively correlated with each other, while HNR is strongly negatively correlated.

LinReg performance on a sweep of train/test splits is summarized in Figure 3. The prediction accuracy is highest in the training subset at low splits and consistently decreases as the proportion of data used for training increases. The opposite is observed in the testing subset. The error bars increase with decreasing subset size, suggesting that the higher variation introduced by smaller subsets destabilizes the model.

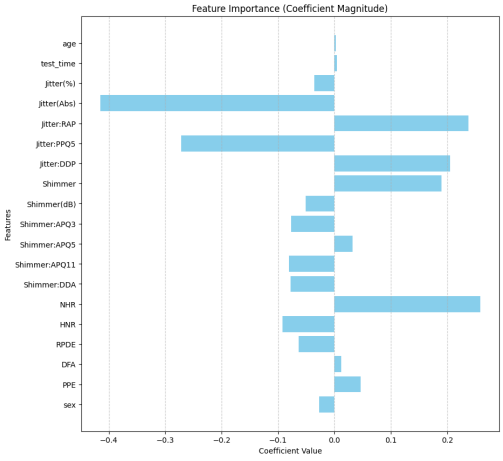

Figure 1: Weights obtained from the Linear Regression model on PD

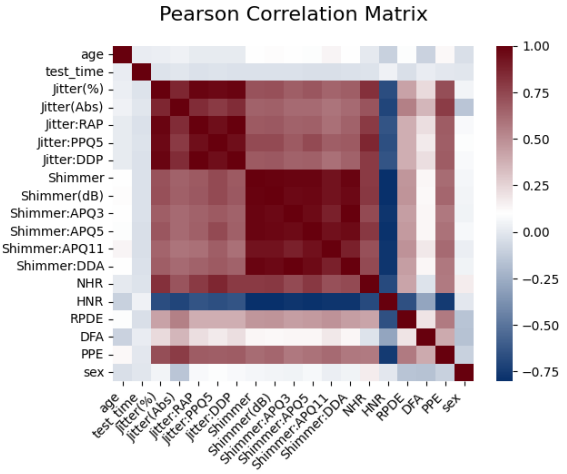

Figure 2: Weights correlated in a Pearson correlation matrix

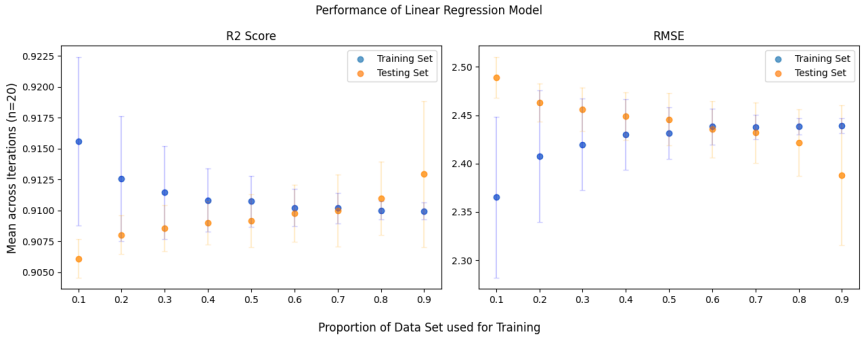

Figure 3: Performance obtained for the Linear Regression model during split sweep

Logistic Regression

Looking now at the LogReg model, Table 1 shows the accuracy and success of the model. Figure 4 shows the confusion matrix for the train/test subsets. Figure 5 shows the data obtained from image analysis were highly weighted by the model. However, area1, concavity2, symmetry2, and smoothness3 are not valued as much as the other features. Figure 6 shows that LogReg's performance in the training subset remains high across all splits, while testing performance improves. Recall is stable, but precision increases, suggesting a higher likelihood of false positives with less representative training sets.

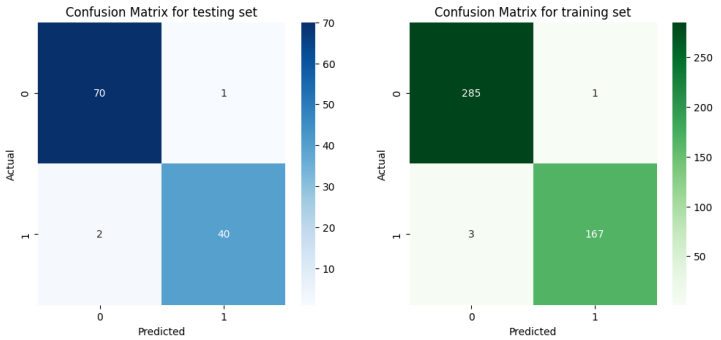

Figure 4: Confusion Matrix of Predictions made by LogReg

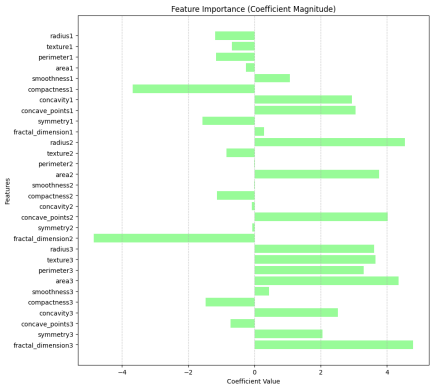

Figure 5: Weights correlated in a Pearson correlation matrix

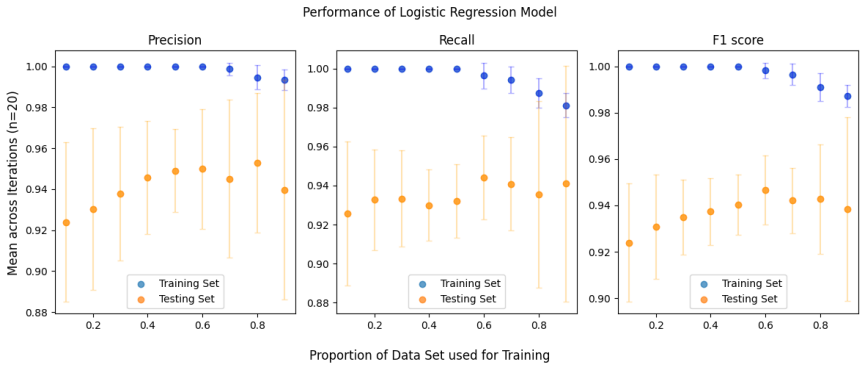

Figure 6: Performance obtained for the Logistic Regression model during split sweep

Linear Regression with Mini-batch Stochastic Gradient Descent

A variation of the LinReg was implemented with mini-batch stochastic gradient descent (LinReg SGD) and evaluated across a series of batch sizes and learning rates. As shown in Figure 7 R^2 and RMSE were consistent for batch sizes 8–32 but declined with larger batches, with the worst performance at full batch. Batch sizes 8 and 16 failed to converge within 2,000 iterations. Among converging trails, the mean convergence speed (calculated as max_iterations / iteration_of_convergence –

1) was slowest for batch size 64 and fastest for the full batch. Overall, larger batches converged faster but performed worse, with batch size 32 offering the best balance. Regarding learning rates, Figure 8 shows prediction accuracy metrics were consistently high for rates 0.001, 0.01, and 0.1 but declined for smaller rates. The largest rates 0.01 and 0.1 failed to converge before the maximum in all 10 trials, implying oscillations over the ideal fitting. The mean convergence speed was the fastest for 0.001. Therefore, of the rates tested, 0.001 was determined as the best learning rate for a batch size of 32.

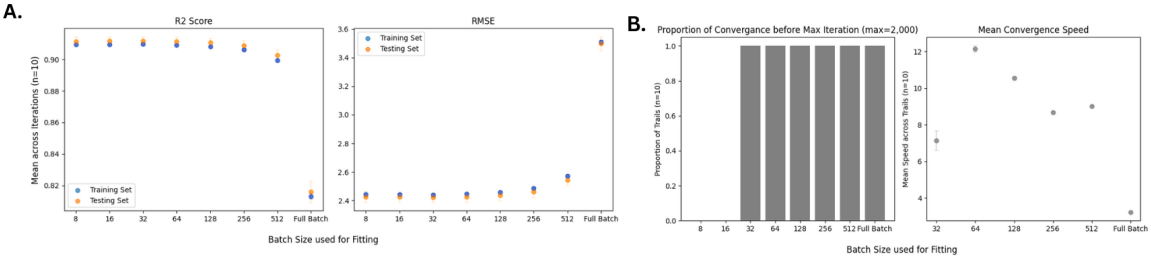

Figure 7: LinReg SGD Performance Across Batch Sizes

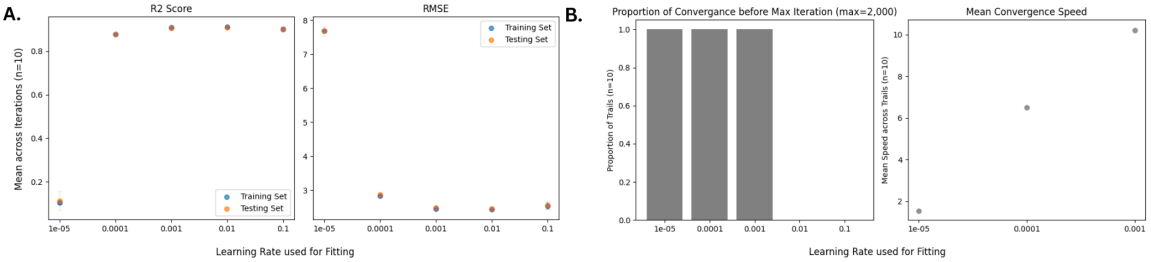

Figure 8: LinReg SGD Performance Across Learning Rates

Logistic Regression with Mini-batch Stochastic Gradient Descent

Similarly, a variation of the LogReg model was implemented using mini-batch stochastic gradient descent (LogReg SGD) and evaluated. Precision and F1 were consistent for batch sizes 8–32 but declined for larger batches, with the full batch having significantly worse performance (Figure 9). Recall remained uniformly high. Convergence was slowest for batch size 8, with 30% of trials not converged before the maximum iteration. Batch sizes 16 and 32 offered the best balance of performance and convergence speed. Now turning to learning rates. Figure 10 reveals precision and F1 were highest at learning rates 0.01 and 0.001. The fastest convergence occurred at 0.001, whereas 0.01 and 0.1 failed to converge. Therefore, 0.001 was selected as the best of the tested learning rates.

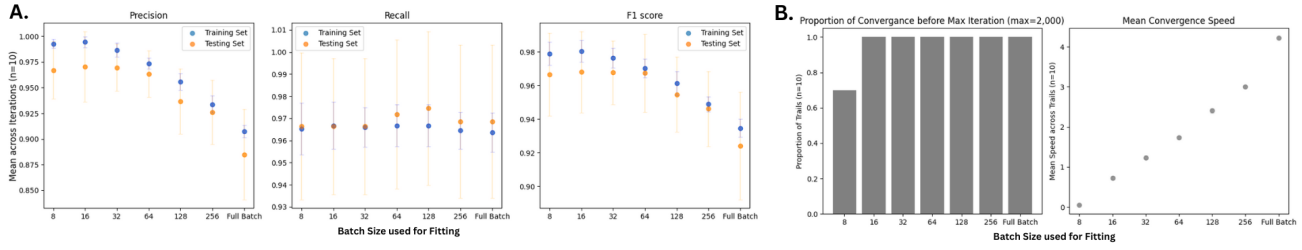

Figure 9: LogReg SGD Performance Across Learning Rates

Additional Investigation

Figure 11 shows the test accuracy when the learning rate and lambda changes. The behavior of lambda shows that at a high value, the model does worse than at an in-between value such as 0.01, same for learning rate, where the middle value gives the best results. These result are in accordance with the theory since extreme values can constrain the model too highly or make the model step too far per iterations.

We additionally investigated adding polynomial features, discovering that they significantly improved the linear regression model by capturing non-linear patterns in the Parkinson's data (Figure 12). This implementation resulted in our best preforming linear regression model with $R^2 = 0.9449$ and RMSE = 1.9091. However it added no accuracy to our logistical regression models working with the Breast Cancer dataset.

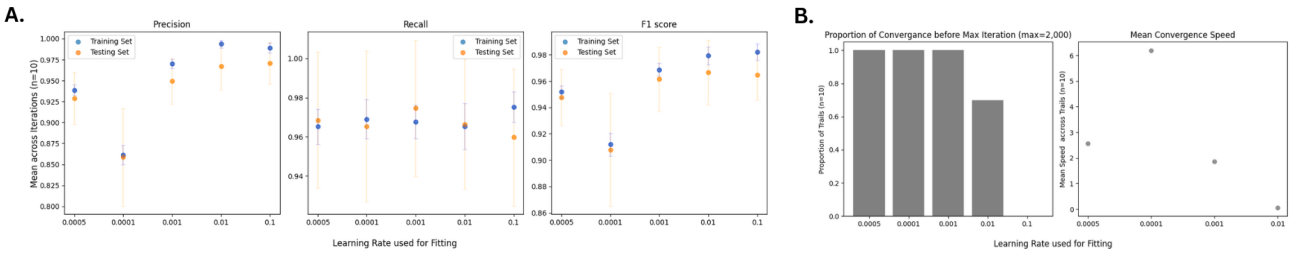

Figure 10: LogReg SGD Performance Across Learning Rates

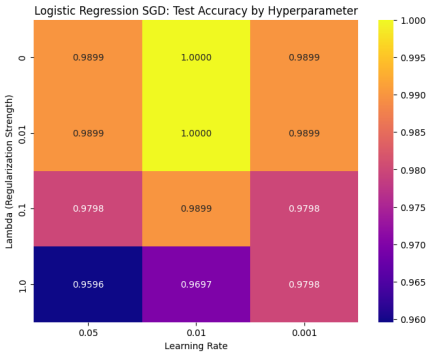

Figure 11: Hyperparameter modulation impact in Logistic Regression SGD

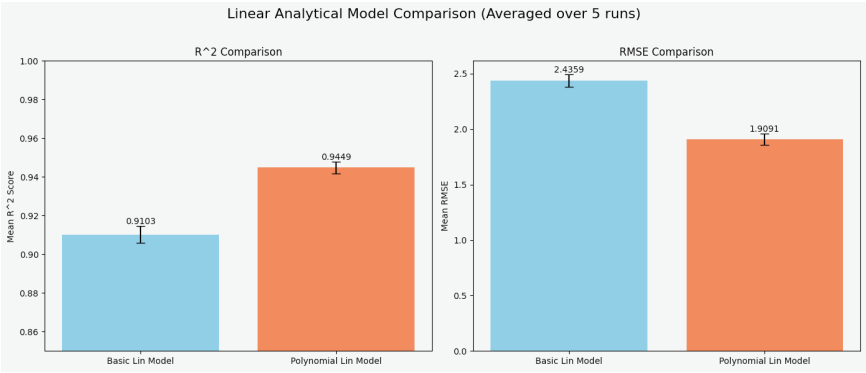

Figure 12: Comparison of Performance of Polynomial Linear and Basic Linear Model

Discussions and Conclusions

Looking first at the LinReg model. The key features identified by the LinReg model (jitter, NHR, and shimmering) match previous analysis of the same PD dataset and allow the model to accurately predict the patients' state. Analysis of various test/test splits revealed LinReg model is not robust across all splits and makes the most consistently accurate predictions with 70/30 or 80/20 split. This is likely because the model requires a minimum proportion of the dataset for training in order to be representative, and a test set large enough to avoid inflated performance results.

The LogReg model identified and highly weighted specific metrics of the image analysis matching previous analysis of the BC dataset. It is observed that while overall the accuracy of the model is high, there is a tendency to false positives. A testing and training set size study showed that LogReg performance is robust across a range of splits.

Including mini-batch SGD for LinReg resulted in reduced performance compared to the analytical solution, since the analytical solution provides the theoretically optimal result for linear regression. However, the SGD model converges to a solution with nearly identical performance. Critically, the accuracy of the SGD solution and convergence speed is dependent on both learning rate and batch size. This demonstrates the necessity of testing sweeps before assigning these hyperparameters. So while the analytical solution is faster and more direct for smaller datasets, the SGD approach is an effective method for larger datasets. Regarding mb-SGD for LogReg, the results obtained in this work are a testament to the impact and use of models such as LinReg and LogReg. Simpler models like these are quite powerful and it would be insightful to expand their use to other diseases or other modes of image analysis to aid in the diagnosis efforts.

Statement of Contribution

Charlotte Livingston: Results, Discussion, Testing of LinReg + LogReg v. train/test split, Testing of LinRegSGD + LogRegSGD v. batch size + learning rate

Adam Dufour: Abstract, Introduction, Dataset analysis, Linear Regression model, model weights, Discussion

Gabriel Caballero: Implementation of LogReg + LinRegSGD + LogRegSGD, Discussion, Implementation and testing of polynomial feature models

References

[1] Hamideh Ramezani, Hossein Khaki, Engin Erzin, and Özgür Baris Akan. Speech features for telemonitoring of parkinson's disease symptoms. *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3801–3805, 2017.

[2] W. Nick Street, William H. Wolberg, and Olvi L. Mangasarian. Nuclear feature extraction for breast tumor diagnosis. In *SPIE Proceedings*, volume 1905, pages 861–870. SPIE, July 1993. https://proceedings.spiedigitallibrary.org/proceeding.aspx?articleid=1008972 ; http://ui.adsabs.harvard.edu/abs/1993SPIE.1905..861S/abstract ; https://www.spiedigitallibrary.org/conference-proceedings-of-spie/1905/1/Nuclear-feature-extraction-for-breast-tumor-diagnosis/10.1117/12.148698.short?SSO=1.

[3] Athanasios Tsanas, Max A. Little, Patrick E. McSharry, and Lorraine O. Ramig. Accurate telemonitoring of parkinson's disease progression by noninvasive speech tests. *IEEE Transactions on Biomedical Engineering*, 57:884–893, 2009.

[4] Athanasios Tsanas, Max A. Little, Patrick E. McSharry, and Lorraine O. Ramig. Robust parsimonious selection of dysphonia measures for telemonitoring of parkinson's disease symptom severity. In *International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, 2011.

[5] William H. Wolberg, Street Wn, and Olvi L. Mangasarian. Image analysis and machine learning applied to breast cancer diagnosis and prognosis. *Analytical and quantitative cytology and histology*, 17 2:77–87, 1995.