COMP 551: Mini-Project 1: Machine Learning

Luis Medrano, Manoj Krishna Venkatesan and Israel Hernandez

Abstract— **For this Mini-Project, the research team investigated the performance of two linear classification models: Logistic Regression and Naive-Bayes. In addition to the basic construct of the algorithms that implement both of these models, two benchmark datasets were selected and pre-processed with strategies for improving the overall performance. The first dataset consisted of numerical features and the second one of a mix of numerical and categorical features. Both of them were adjusted in order to be easily processed by the implemented models. Two additional datasets were treated and tested with the above models. After applying our tests, it was found that Naive-Bayes is more computationally efficient than Logistic Regression and that logistic regression needs intensive computation time.**

I. KEYWORDS

Logistic Regression, Naive-Bayes, Data Processing, Normalization, One-Hot Encoding, Data Distribution, Data Training, Predictions, Accuracy Optimization.

II. INTRODUCTION

The following are the tasks that were carried on for this Mini-project: a) Investigating the performance of the Logistic Regression and the Naive-Bayes classification models; b) Exploring techniques to optimize the performance of such models through the pre-processing of the datasets.

Two main datasets, obtained from the UCI Machine Learning Repository (1), were specified by the requirements of the project: Ionosphere and Adult datasets. On the Ionosphere dataset Sigillito (2) presents the radar data collected by a measuring system in Goose Bay, Labrador, where each of 17 pulse numbers are represented by 2 attributes, corresponding to complex values that describe the studied electromagnetic signal. Zhou et al. (3) found this dataset essential for testing a novel decision tree algorithm titled as Neural Ensemble Based C4.5. The natural distributions of the Ionosphere dataset allowed them to demonstrate the good comprehensibility of their C4.5 neural network. For the adult dataset, Becker et al. (4) extracted more than 40,000 instances from a Census database of 1994. Most of these instances are clean records that provide categorical information of several people. The main goal of this extraction was to carry on the exercise of determining if a person earned more than 50,000 units of currency per year. Kohavi (5) proved the, as he describes, surprising accuracy of Naive-Bayes induction algorithms by working with the adult dataset. This last reference serves as an indicator to corroborate that this mini-project is following the right direction. Two additional categorical datasets were used to add test scenarios: The Breast Cancer Original dataset and the Electrical Grid Stability Simulated Data. On the Breast Cancer dataset, Wolberg (6) reported clinical cases from January 1989 to November 1991. Zhang (7) found this dataset useful to implement instance-based learning and test classification models. On the Electrical Grid dataset, Arzamasov (8) implemented the Decentral Smart Grid Control concept. In his journal, Arzamasov (9) explains how the studied system is able to implement demand response with minimal changes of the infrastructure.

The logistic regression model was built with base on the procedure described on the chapter 8 of the Machine Learning book written by Murphy (10) and the Naive-Bayes model was composed based on the steps mentioned by Parsian (11) in his data algorithms book. Since the early development stages of these models, there was also development of algorithms pre-process the Ionesphere and the Adult datasets, which would be used to test Logistic Regression and Naive-Bayes respectively. These pre-processing algorithms are further discussed in the DATA ANALYSIS section of this document.

Although in the final stages, both models were implemented for all the chosen datasets, it was found out that Logistic Regression is the most accurate for continuous data, as in the Ionosphere dataset, while Naive-Bayes is better when handling discrete data, as in the rest of the chosen datasets. Another important characteristic that was found out, is that logistic regression is more

[2]The authors are with the Department of Electrical and Computer Engineering, McGill University, Montreal, QC H3A 2K6, CANADA Project Repository at $github.com/vmk94/Team29_01$

time consuming, as it can take several iterations to compute, while Naive-Bayes is a direct method that delivers result in a shorter time.

III. DATA ANALYSIS

A. Dataset Pre-Processing & Discarded features

In the case of the Ionosphere dataset, the pre-processing consisted in extracting the binary labels (good or bad) column; erasing the second column, which was a feature with only zeros and wouldn't really add up to the analysis; checking that all the elements of the features matched the expected data type (float); and finally, normalizing the dataset so its values ranged only from 0 to 1.

For the adult dataset the nationality and race features were removed, since the values "United States" and "White" were repeated throughout more than 80 % of the instances. Also, the feature "Study" was eliminated since the dataset was encoded through "number labeling". Meanwhile, the missing categorical values were replaced with the mode of its respective feature. Another point to notice is that these categorical features were One-Hot Encoded. With this, the dataset was ready to be processed by the classification models.

For the Breast Cancer dataset the first column was eliminated as it was an ID instance number and the last feature (*target*) was binomially encoded. Finally, for Electrical Grid dataset, the headline row was removed and then, the features *tau1, p1, g1* were eliminated because these were non predicted values that were dependent of the rest of the data. Additionally, the feature *stab* was eliminated since this was the real number characteristic equation that identified the classification of the data.

B. Exploratory analysis - feature distribution and correlation

As part of an exploratory analysis, the team decided to study the distribution and correlations of the features of the datasets; however, because of the great amount of features that are included in these datasets, and after some literature research (12), we decided to employ the weights obtained in the logistic regression model to decide what features would be the most representative of the dataset and thus focus on exploring their distribution.

In the Ionosphere dataset, a feature with the greatest weight was that of the first column, with a weight of:

$$\beta = -4.750686$$

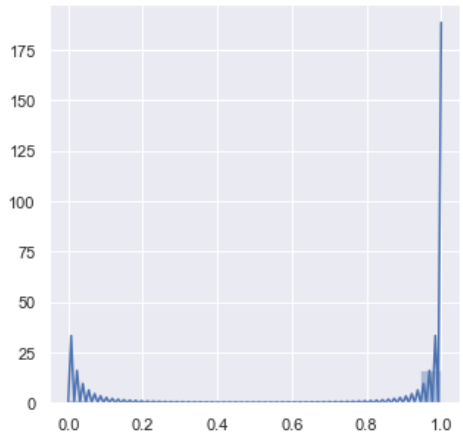

Fig. 1. Distribution for most representative feature of the Ionosphere dataset

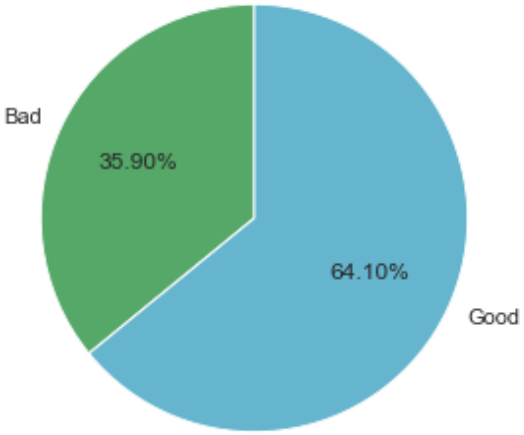

Fig. 2. Binary distribution (Good vs Bad) of Ionosphere dataset.

As seen in Figure 1, the distribution of this feature is highly charged towards the value of 1, which is also the mode of this set. The binary distribution of the whole data set can be seen in Figure 2. For the case of the adult dataset, we can see how the dominant feature of the set is the Capital Gain, with a weight of:

$$\beta = -25.461216$$

This feature is the one that mostly influences the dataset to have the binary behaviour (earning more than 50k per year) that is shown in Figure 3.

In order to employ the useful study of correlation, the correlation matrices were built for all of the datasets. Figures 4 and 5 show the correlation matrices for the datasets of Breast Cancer and Electrical Grid, respectively.

So as it can be appreciated in the correlation matrices

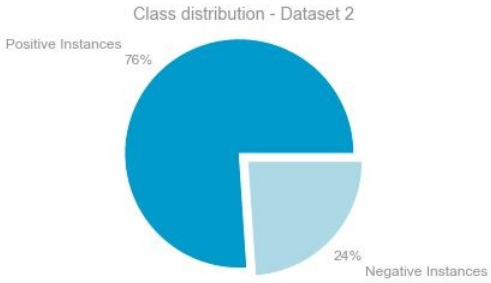

Fig. 3. Binary distribution for whether the individual makes more than 50k of income per year, for the adult dataset.

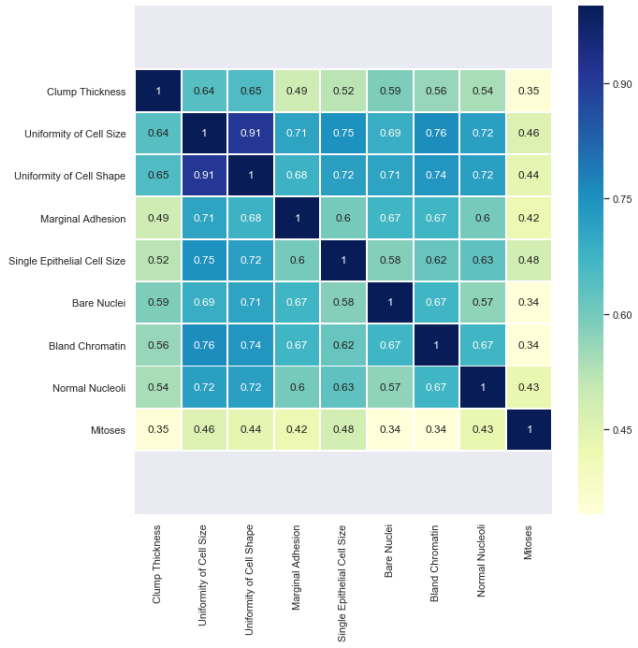

Fig. 4. Correlation matrix for the features of the Breast Cancer dataset

of these last two datasets, that there is no extreme correlation between the features and hence there are no redundant features in these datasets.

IV. RESULTS

As part of the task of investigating the performance of our classification models, we invested time in implementing the 5-fold method to measure the accuracy. As soon as we had this measurement prepared and our models running, we executed experiments to evaluate how the accuracy is impacted by two essential parameters: the number of training set instances and the convergence speed (number of iterations). Additionally, we compared the accuracy of the Logistic Regression model against the accuracy of Naive-Bayes. Lastly,

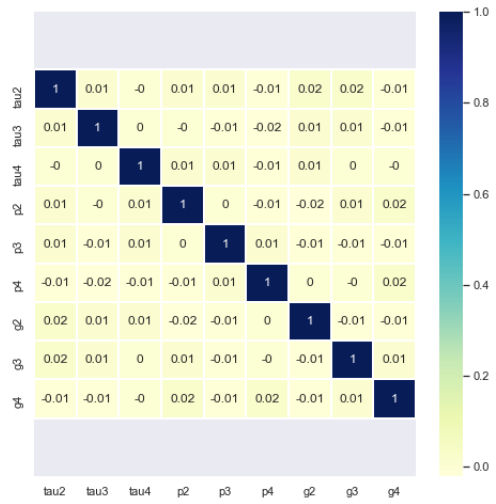

Fig. 5. Correlation matrix for the features of the Electrical Grid dataset

we looked into how our pre-processing of the datasets and additional modifications to the model algorithms, improved the overall performance.

A. Convergence Speed of Logistic Regression

When it comes to the the convergence performance of Logistic Regression, the learning rate of the gradient descent method plays a central role. By testing different learning rates, while implementing Logistic Regression on the Ionosphere dataset, we noticed that the convergence speed can increase when the learning rate goes into any of its extremes. A learning rate that is too low will naturally make the model to take more iterations to reach convergence; in the other hand, a learning rate that is too high, maybe cause overshooting and this will increase the converge at a slower pace, if it ever converges. Thus, an ideal learning rate will be somewhere between these polar opposites and will allow the algorithm to be fast enough without causing overshoots.

B. Impact of Train set size on Accuracy

In this exercise, the accuracy of the Logistic Regression model was tested on the Breast Cancer and the Electrical Grid datasets by varying the size of the training data sets from 0 to 1 proportion of training instances with respect to the total amount of instances, so 1 represents a 100 % of training instances. Figures 6 and 7 illustrate the relationships of these proportions

with the accuracy of the model for the Breast Cancer and the Electrical Grid datasets, respectively. These figures clearly suggest two key aspects: a) Increasing the number of training instances will improve the accuracy and b) there exists a range in which the relationship between these two variables can be modeled with a linear function of a positive slope.

Fig. 6. Proportion of training sets (horizontal-axis) against accuracy (vertical-axis) for Breast Cancer dataset.

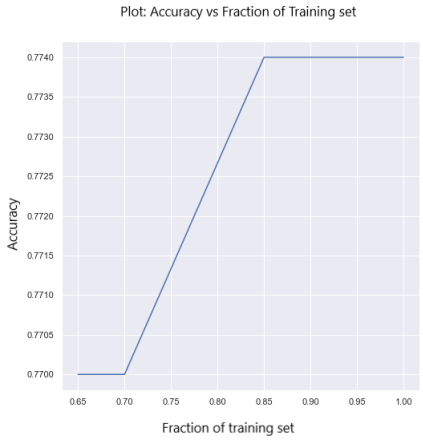

Fig. 7. Proportion of training sets (horizontal-axis) against accuracy (vertical-axis) for Electrical Grid dataset.

C. Accuracy Comparison: Logistic Regression vs Naive-Bayes

Table 1 allows us to compare the optimal accuracies that were obtained for each of the four datasets with Logistic Regression (LR) and Naive-Bayes (NB). As it can be seen in Table 1, there is not a clear winner. Although Logistic Regression has a greater accuracy on the majority of the datasets, the difference is only considerable for the Ionosphere dataset.

	Logistic Regression	Naive-Bayes
Ionosphere	0.9056	0.7788
Adult	0.8401	0.8180
Breast Cancer	0.9714	0.9599
Electrical Grid	0.7673	0.7812

TABLE I
OPTIMAL ACCURACIES OF LR AGAINST NB

D. Impact of Dataset pre-processing

As it was stated in the first sections of this document, one of the main tasks of this project was to work on techniques to improve the performance of our two classification models. Different results were obtained for each classification model after going through the tests. At least for the four chosen datasets, the improvements in the convergence speed and accuracy were negligible when using Logistic Regression. Removing the features or applying normalization did not have a considerable effect on the performance of the model. An equivalent result was obtained with the Naive-Bayes model. As observed in Figure 8, the effect in the accuracy is negligible.

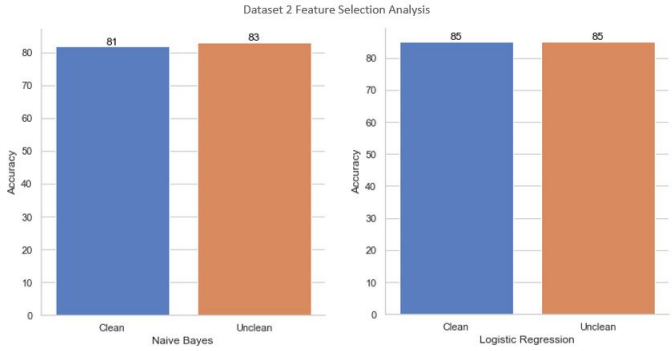

Fig. 8. Comparison of Accuracy changes after pre-processing the datasets and then implementing Naive-Bayes.

V. CONCLUSIONS

After running through the entire exercise we were able to complete a detailed investigation of the performance of the Logistic Regression and Naive-Bayes Classifications models. From our observations, we can tell that there is no way to generalize and state that any of these models is always better than the other and that it really depends on the characteristics of the dataset. For example, the Logistic Regression seems to work better for datasets that are mainly formed

by continuous data than Naive-Bayes; however, we would need to dedicate a more extensive test to confirm this hypothesis. In terms of convergence speed, it is clear that Naive-Bayes takes the lead because of its direct nature (no need of iterations). All in all, choosing between these two models implies a trade-off between accuracy and speed. For datasets in which the difference of accuracy is not great (which was the case for all datasets except for the Ionosphere dataset), Naive-Bayes might be more convenient if the amount of instances requires a considerable increase.

Because our data pre-processing seemed to be ineffective, the first improvement that we could do in the future is in this aspect. Data regularization and SMOTE for dataset imbalance are procedures that are popular and that we could first learn more about in order to decide if they could have a greater impact to the techniques we already applied. In terms of improving the models, future work should focus on implementing variations of the gradient descent, such as SGD and ADAM.

STATEMENT OF CONTRIBUTIONS

Although the work was split between the team, everyone was involved in the three main tasks in which we split the project: Data Pre-Processing, Model implementation and Documentation. Luis was mainly in charge of working the Adult Data Set and the Naive-Bayes model. Manoj was in charge of working with the Breast Cancer and Electrical Grid datasets. Manoj was also responsible of developing the algorithms for implementing the Logistic Regression Model. Israel worked on the Ionosphere Dataset and on documentation. As said before, every team mate would help the others in the tasks that were mainly assigned to each.

REFERENCES

[1] Dua, D. and Graff, C. UCI Machine Learning Repository: Adult Data set. University of California, Irvine, School of Information and Computer Sciences, 2019.

[2] Zhi-Hua Zhou and Yuan Jiang, "NeC4.5: neural ensemble based C4.5," IEEE Trans. Knowl. Data Eng., vol. 16, no. 6, pp. 770–773, Jun. 2004.

[3] Zhi-Hua Zhou and Yuan Jiang, "NeC4.5: neural ensemble based C4.5," IEEE Trans. Knowl. Data Eng., vol. 16, no. 6, pp. 770–773, Jun. 2004.

[4] R. Kohavi and B. Becker. UCI Machine Learning Repository: Adult Data Set. University of California, Irvine, School of Information and Computer Sciences, 2019.

[5] O. L. Mangasarian and W. H. Wolberg. UCI Machine Learning Repository: Breast Cancer Wisconsin (Original) Data Set. University of California, Irvine, School of Information and Computer Sciences, 1990.

[6] O. L. Mangasarian and W. H. Wolberg. UCI Machine Learning Repository: Breast Cancer Wisconsin (Original) Data Set. University of California, Irvine, School of Information and Computer Sciences, 1990.

[7] J. Zhang, "Selecting Typical Instances in Instance-Based Learning," Machine Learning Proceedings 1992, pp. 470–479, 1992.

[8] V. Arzamasov, UCI Machine Learning Repository: Electrical Grid Stability Simulated Data Data Set. University of California, Irvine, School of Information and Computer Sciences, 2018.

[9] V. Arzamasov, K. Böhm, and P. Jochem, "Towards Concise Models of Grid Stability," in 2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm), 2018, pp. 1–6.

[10] K. P. Murphy, Machine learning: a probabilistic perspective. MIT press, 2012.

[11] M. Parsian, Data Algorithms: Recipes for Scaling Up with Hadoop and Spark. "O'Reilly Media, Inc.," 2015.

[12] M. Parsian, Data Algorithms: Recipes for Scaling Up with Hadoop and Spark. "O'Reilly Media, Inc.," 2015.