Data Mining Techniques Assignment 1 - Basic

Nouri Mabrouk[2623401], Jayden Ly Vy[2683734], and Lucas Harlaar[2681328]

Group 56, Vrije Universiteit, 1081 HV Amsterdam, the Netherlands

In this report we analyse and apply different methods from the field of Data Mining. Three tasks are performed in order to get acquainted with the field.

Task 1: Explore a small dataset

Task 1A: Exploration

Exploratory data analysis (EDA) is performed to better understand the attributes of the ODI dataset. It consists of 313 DMT student entries who participated in an online questionnaire of 17 questions. Therefore the dimensions of the

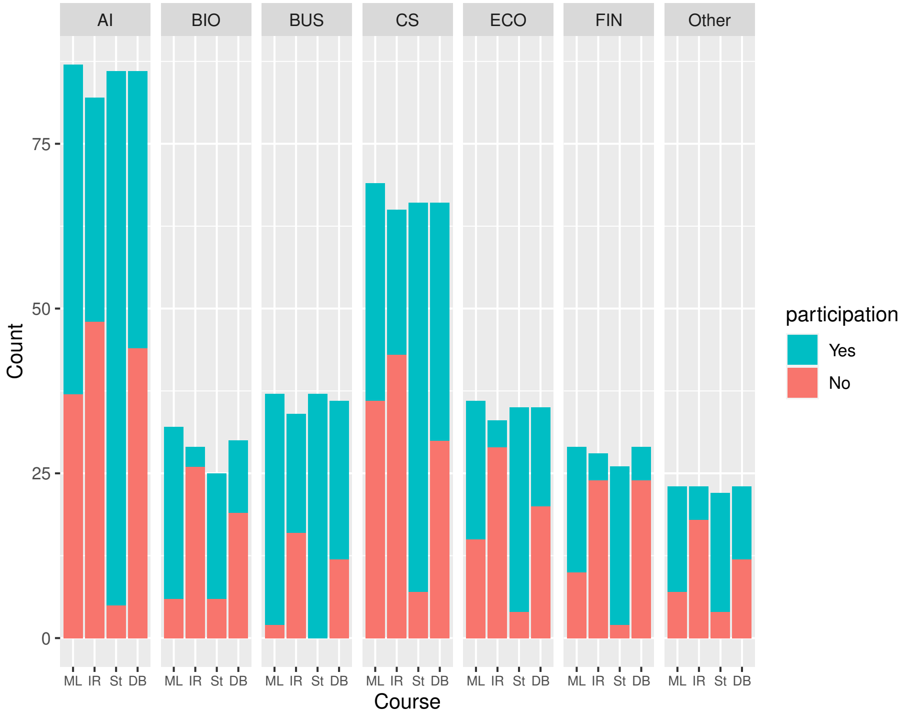

Fig. 1. ODI 2021 data: Course background information per program category. Courses are abbreviated as follows: ML = Machine learning, IR = Information retrieval, St = Statistics and DB = Databases.

dataset are (313 × 17). The 17 features consist of a mixture of multiple choice and open questions, resulting in both textual and numerical entries. To handle this diversity various plots of the different variables were made which gave a good indication on how to proceed. Then the data was cleaned by removing outliers in numerical entries, filling in missing entries with either the mode or median and categorizing text entries. All binary multiple choice entries of student's course backgrounds are presented in Figure 1. Most students come from an AI or CS program. It is noticeable that very few students in a Finance or Bio program have experience in Information Retrieval or Databases, compared to other program categories. Relatively, students in a Business program have the most background knowledge in general.

Figure 2 shows the correlation between each pair of numerical features. This detects similar features, which helps in the feature selection process. The correlations are not very high in general. Only $|\rho_{neighbours,reward}|$ and $|\rho_{stress,birthyear}|$ > 0.1. Interestingly the latter is negatively correlated indicating that younger students have a lower stress level in general.

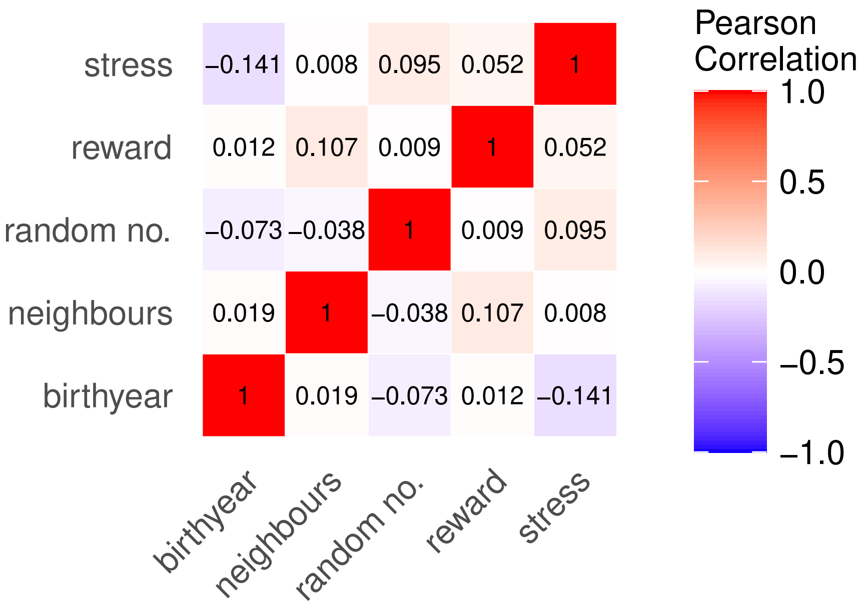

Fig. 2. ODI 2021 data: Correlation Matrix of the numerical variables in the dataset.

Figure 3 shows that most people think chocolate makes you fat and define a good day based on the weather conditions. Interestingly, the share of female participants answering slim on the chocolate question is the smallest. Furthermore, the share of male and female defining a goodday by the weather conditions is almost equally divided, a pattern which does not exist in any other categories. It indicates that these features could be useful in predicting gender, which is what we do in the next subsection.

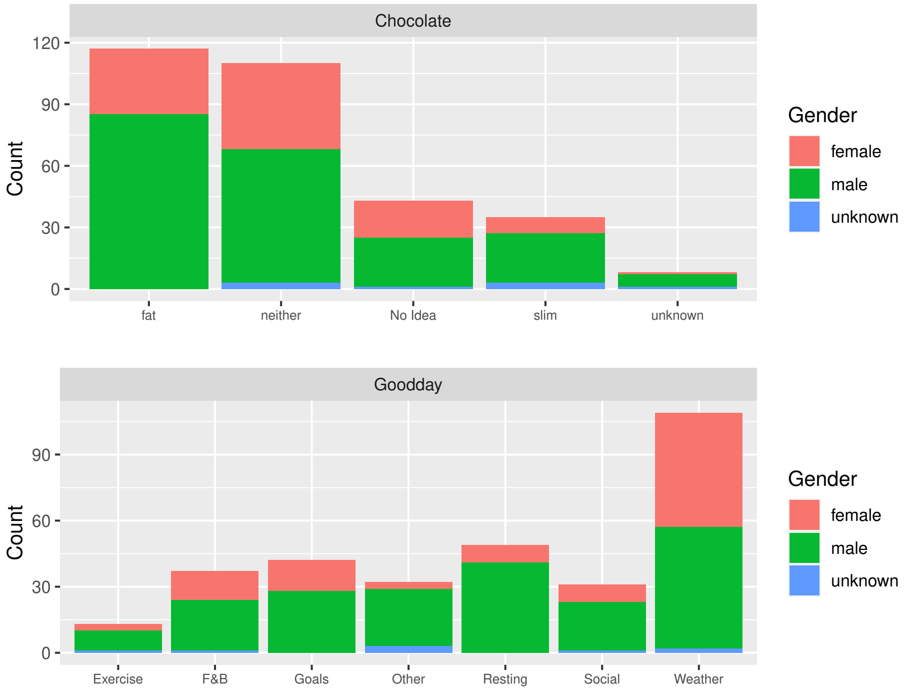

Fig. 3. ODI 2021 data: Student's opinion on the effect of chocolate on the human body and what defines a good day. Good day entries are categorized. F&B refers to the 'Food and Beverages' category.

Task 1B: Basic classification/regression

After the exploration of the ODI dataset, some classical classification models were trained to predict the gender of the students. First, a validation data set was created by splitting the dataset into two parts, 80% of which is used to train our models and 20% is reserved as the validation set. Eventually, a variety of statistical methods are employed to estimate the accuracy of the models when tested on the mentioned validation set. 10-fold cross-validation is used to estimate accuracy. This cross-validation method is an iterative process that splits our training set into 10 parts, train the model in 9 and test on 1. In the end all 10 parts function as testing part.

Since it is unknown which algorithms would be the best for the problem at hand, a mixture of simple linear (Linear Discriminant Analysis - LDA), nonlinear(k-Nearest Neighbor - kNN) and complex nonlinear methods (Support Vectors Machines - SVM and Random Forest - RF) are used. In total, 4 algorithms are employed and evaluated. Our classification setup is as follows: Gender is used as dependent variable, which is a factor variable consisting of 3 levels (Male, Female and Unknown). Independent variables are the opinion of the students on what is a good day, how they think about chocolate, which program of study they are following and their stress level. These features are selected because of the interesting patterns in Figure 3 where we observed that female and male students tend to have different opinions regarding chocolate and what defines a good day. Program of study and stress level were then added to provide more diverse information that could help in classifying gender correctly.

The different parameters being used in the models are as follows: For the Random Forest model, the tuning parameter used is mtry, which is the number of variables randomly sampled as candidates at each split. The final value used for the RF model was mtry = 2. For the kNN model, the parameter is k, which refers to the number of nearest neighbors to include in the majority of the voting process. The final value used for the kNN model was k = 9. For the SVM model, the final parameter values were sigma = 0.0415713 and C = 1. Lastly, there are no tuning parameters for the linear discriminant analysis model for classification.

Finally, a comparison between the model performance on the training set is made in order to select the most accurate one. The metric 'Accuracy' is used to evaluate model performance. This is a ratio of the number of correctly predicted instances divided by the total number of instances in the data set multiplied

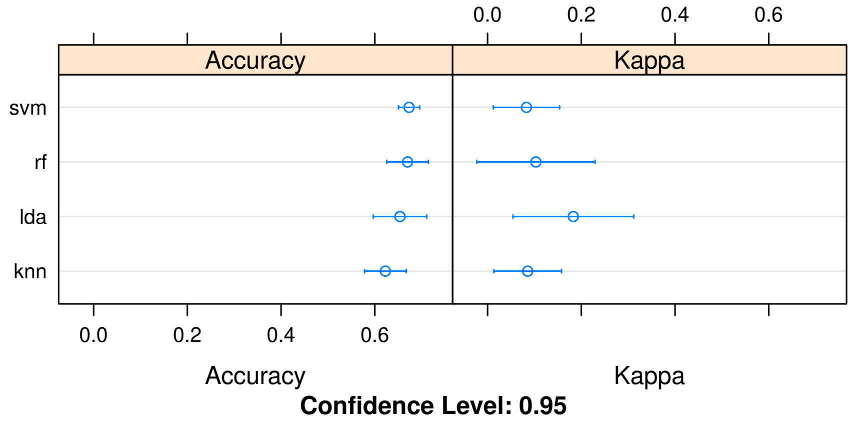

Fig. 4. ODI 2021 data: Comparison of Machine Learning Algorithms in R.

by 100 to create a percentage (i.e. 95% accurate). Figure 4 evaluates the four models by comparing the spread and the mean accuracy of each model. There exists a spread in the accuracy measures because each algorithm was evaluated 10 times (10 fold cross validation).

The SVM is the most accurate model based on the training set, since it has the highest accuracy level and lowest spread. The next step is to calculate the accuracy of the models on the validation set. This will serve as an independent final check on the accuracy of the best model. The models were run directly on the validation set and the results are summarised as follows:

Table 1. ODI 2021 data: Performance statistics of the classifier models on the validation set

Overall Statistic	LDA	RF	SVM	KNN
Accuracy	0.6271	0.6102	0.6102	0.5424
95% CI	(0.4915, 0.7496)	(0.4744, 0.7345)	(0.4744, 0.7345)	(0.4075, 0.6728)
No Information Rate	0.6102	0.6102	0.6102	0.6102
P-Value [Acc >NIR]	0.4508	0.5568	0.5424	0.8845

Note that the accuracy of the SVM model, while being tested on the validation test, is not the best model anymore. In this case LDA is the best performing model for unseen data. Interestingly, the accuracy of RF and SVM are identical, this could be because they are both complex nonlinear methods. The p-value reported here is described as: a one-sided test to see if the accuracy is better than the ”no information rate (NIR)” which is taken to be the largest class percentage in the data. This result is not good since the p-value of having the accuracy more than the NIR is not significant in all models. So instead of using these models, one can make a similar prediction with the majority class and still end up with similar or even better outcome.

Task 2: Compete in a Kaggle competition to predict Titanic survival

Task 2A: Preparation

The test and training set were obtained from Kaggle and combined to explore the full dataset as a whole. When training and testing, they are split up again and the training set itself is also split into a train and test part. The data consist of 1309 observations in total of which 891 observations belong to the training set and 418 observations belong to the test set. The features are a mix of text and numerical variables. Some equivalence and obvious correlations can be noticed which serve as justification for the following transformations of the data: adding the number of sibling/spouses and parents/children together can be considered as family size. The deck level is extracted from the cabin numbers and dropping the room number.

Furthermore, the name variable was dropped because most information contained in the name correlates highly with other features such as age and sex. In the same line of reasoning ticket fare was dropped, because class, age and deck level are correlated with that feature. Lastly, it was decided upon to drop the ticket number and embarked features because they are unnecessary for the learning algorithm and a more parsimonious model is preferred. The descriptive statistics of the selected features in the training set after transformations are shown in Table 2. The age and deck variables have many missing values. Imputation of these missing values is done by means of Predictive Mean Matching. Now the data is fully prepared to be used in the classifier algorithms which is described in the following subsection.

Table 2. Titanic dataset: descriptive statistics of selected features in training set.

Survived	Pclass	Sex	Age	FamSize	Deck
0:549	Min. :1.00	male :577	Min. : 0.42	Min. : 0.00	C : 59
1:342	1st Qu.:2.00	female:314	1st Qu.:20.12	1st Qu.: 0.00	B : 47
	Median :3.00		Median :28.00	Median : 0.00	D : 33
	Mean :2.309		Mean :29.70	Mean : 0.9046	E : 32
	3rd Qu.:3.00		3rd Qu.:38.00	3rd Qu.: 1.00	A : 15
	Max. :3.00		Max. :80.00	Max. :10.00	(Other): 18
			NA's :177		NA's :687

Task 2B: Classification and Evaluation

In order to evaluate the performance of two classifiers on the data, the original and cleaned training set itself is divided in a 80/20 train and test set. These sets contain 713 and 178 observations respectively. It allows training and testing of the algorithms on data of which we know the actual outcome and hence measure their performance directly. We select the random forest and support vector machine algorithms to do so. Table 3 shows the confusion matrices and performance measures after training and testing both classifiers. We notice that the random forest classifier has an accuracy level of 85% and slightly outperforms the SVM. Therefore, the former model is selected and applied on the original and cleaned test set obtained from Kaggle.

Uploading the solution on the platform resulted in a score of 0.75, indicating that a correct outcome was predicted for 3/4 passengers. This is 10% lower than expected based on the results from our training phase and 10-fold cross validation. This can be explained by the different makeup of the test set on Kaggle. The ratio of survived vs died is different between the training set and the test set on which the solutions are based. This results in the model being more accurate at predicting the holdout set using cross validation, than at predicting the test set. It is expected that an optimized (overfitted) solution for a subset of data performs less on extrapolation.

Table 3. Titanic data: classifier train and test evaluation

Confusion Matrix				
	Random Forest		SVM	
	Predict. deaths	Predict. survivals	Predict. deaths	Predict. survivals
Actual deaths	104	11	105	10
Actual survivals	15	48	17	46
Statistics				
Accuracy:		0.8539		0.8483
95% CI:		(0.7933, 0.9023)		(0.787, 0.8976)
No Information Rate:		0.6461		0.6461
P-Value [Acc >NIR]:		4.605e-10		1.449e-09

There are numerous strategies which might have improved the score. In our research, the name and fare (after imputation of age and deck) variables were dropped. A more rigorous analysis would have extracted more value from these variables and thus result in a higher accuracy score. Moreover, the imputation of missing values was done using predictive mean matching, which may not be optimal for the situation at hand. Furthermore, many more modeling strategies and specifications may have been compared in order to select a better performing model. Since the fitted model is not the most sophisticated one, it was unexpected beforehand to see an accuracy of 0.9 or higher. Lastly, because the models are well performing and use informative variables to predict survival odds, they are expected to outperform the No Information Rate by far. Hence, 85% accuracy on the holdout set and 75% on the test set aligns with our expectations.

Task 3: Research and theory

Task 3A: Research - State of the art solutions

For this task a data mining competition that finished 8 years ago was analysed: Amazon.com - Employee Access Challenge on Kaggle. For this competition participants were asked to predict employee access privileges for various company specific applications, based on historical data regarding employee roles and the corresponding resources to which they have access. The competition used the area under the ROC curve as measure to evaluate the submissions and determine the winner, which is a team consisting of Paul Duan and Benjamin Solecki.

The winning approach of Paul and Benjamin focused on stability of the models, because the dataset contained only categorical variables with a large number of categories. Hence, it is important that the models are relatively robust to changes in the composition of the dataset, which is what distinguishes them from the other top competitors. Furthermore, they spent more time on feature extraction rather than feature selection, because features would be highly dependent on the composition of the dataset. Their winning approach consists of a weighted average of their individual methods. Ben used a weighted average himself of a logistic model and a mixture of tree-based models i.e. Random

Forests, GBMs and Extremely Randomized trees. Paul on the other hand used a classifier consisting of a large combination of models, which were combined by computing their predictions using 10-fold cross-validation and putting them in a second model, which was then trained to determine dynamically which model should be trusted the most [1]. Hence, the fragmented structure of their approach made it an flexible approach that could handle composition changes in the data really well.

Task 3B: Theory - MSE versus MAE

The mean squared error (MSE) and mean absolute error (MAE) are two performance measures for numeric prediction,

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2, \qquad MAE = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \hat{Y}_i\right|, \tag{1}$$

where n is the sample size, Y_i is the observed value and $\hat{Y}_i$ is the predicted value. The MSE is the most commonly used measure, since it is a well behaved function and easy to manipulate mathematically. Both measures do not take into account the sign of the individual errors. However, the MSE is more sensitive to outliers compared to the MAE, since it squares the difference between the observed and predicted values. Hence, any estimator that tries to minimize the MSE can be heavily influenced by the occurrence of such outliers. Therefore, when suspecting outliers occurring in the data one should opt for the MAE as metric for average model prediction error, because it is more robust [2].

There exist unlikely conditions in practice under which MSE and MAE are equal.

Proof. Let $r_i = \left|Y_i - \hat{Y}_i\right|$ and let $\boldsymbol{r} = [r_1, \ldots, r_n]'$ be a vector with the absolute residuals for all n data points. Furthermore let $\mathbf{1}$ denote a $n \times 1$ vector of ones. Setting MSE equal to MAE gives:

$$(\boldsymbol{r} - \mathbf{1})'\boldsymbol{r} = 0, \tag{2}$$

after rearranging. This has two obvious solutions, namely $\boldsymbol{r} = \mathbf{0}$ and $\boldsymbol{r} = \mathbf{1}$. However there exist infinitely many solutions. By completing the square, equation (2) can be rewritten as follows:

$$(\boldsymbol{r} - \frac{1}{2}\mathbf{1})'(\boldsymbol{r} - \frac{1}{2}\mathbf{1}) = \frac{n}{4}, \tag{3}$$

which describes an n-dimensional sphere centered at $[\frac{1}{2}, \ldots, \frac{1}{2}]'$ with radius $\frac{1}{2}\sqrt{n}$.

Hence in practice, when the absolute residuals of any predictive method lie on the surface of this hypersphere, MSE and MAE will give identical results.

To experiment with these performance measures, different regression methods were applied to an economic panel dataset[1] because it lies in our field of interest

[1] https://data.world/data-society/asia-pacific-economic-outlook

as econometrics students. It consists of 40 Asian and Pacific countries and 10 corresponding yearly economic variables between 2013 and 2020. $\%\Delta GDP_{2017}$ and $\%\Delta Export_{2017}$ were used as independent variables and $\%\Delta CPI_{2017}$ as dependent variable. Table 4 shows the resulting MSE and MAE of the regression methods. We notice that the Polynomial regression performed the best, because both performance measures are the lowest in comparison. This method uses a curved line to determine the best fit through the data points, instead of a straight line like Linear regression or Bayesian glm. Intuitively, this makes sense. GDP and Export are good indicators to predict CPI, however their relation is more likely nonlinear. For example, if economic growth is really high, it indicates that country is in a more volatile economic situation. Hence inflation could also be more volatile than in a natural economic growth equilibrium of around 2%. A polynomial regression would describe the variation in CPI better.

Table 4. Performance measures of different regression methods

	Linear regression	Lasso regression	Polynomial regression	Bayesian glm
MSE	4.608351	4.006503	2.908942	4.608269
MAE	1.658279	1.674198	1.38248	1.658291

Task 3C: Theory - Analyze a less obvious dataset

When the data only contains regular texts, the modelling technique that would be suitable to use is Natural Language Processing (NLP). NLP is a machine learning technique that allows computers to break down and understand text as a human being would. Within the NLP field, there is a specific technique that can be applied to our problem at hand, called text classification. Text classification is the process of assigning predefined tags or categories to unstructured text. It's considered one of the most useful NLP techniques because it's so versatile and can organize, structure, and categorize any form of text to deliver meaningful data and solve problems [2].

In order to analyse our text with machine learning the data needs to be preprocessed. By using a number of NLP techniques, the data is transformed as follows: First, the text data column was tokenized, the strings of characters were broken up into semantically meaningful parts that can be analyzed (i.e words), transforming these tokens into lower case and discarding meaningless part such as white space. Second, stopwords were removed to provide a more accurate automated analysis of out textual data, meaning we left out the words that provide very little semantic information or no meaning at all (i.e. ”a”, ”and”, ”or”, ”the”, ”etc.”). Third, punctuation and numbers are left out to retain only the most meaningful tokens for our purpose of classifying between ”ham” and ”spam” messages. Fourth, a stemming operation was carried out on the textual data, which is the process of chopping off the ends of words in order to reduce inflectional forms and derivationally related forms of a word to a common base form. Finally, derivational affixes were removed.

After pre-processing our data for the text classification, the Regularized Regression technique is applied. Regularized regression is a classification technique where the category of interest is regressed on text features using a penalized form of regression. Here a specific type of regularized regression, the Least Absolute Shrinkage and Selection Operator (or LASSO), is applied. In the LASSO estimator, the degree of penalization is determined by the regularization parameter λ [2]. Cross-validation was used to find the optimal value for λ. The labels variable indicates whether an sms is classified as a "spam" or a "ham". Again, 80% of the data was used as the training set and a regularized regression classifier was employed on this subset. In the second step, the class for the remaining test set is predicted. Table 5 shows the confusion matrix and performance statistics of the LASSO regression method.

Table 5. SMS Data: Classifier train and test evaluation

Confusion Matrix		
	Predicted label:	
Actual label:	spam	ham
spam	106	23
ham	1	977
Statistics		
Accuracy:		0.9783
95% CI:		(0.9679,09861)
No Information Rate:		0.9033
P-Value [Acc >NIR]:		<2.2e-16

The confusion matrix shows that the model produces slightly more false ham classifications than false spam classifications, but the majority of labels are correctly predicted. From the statistics, we obtain that the classifier predicted correctly in 97.83% of the cases. It indicates a strong performing classifier with a very significant p-value.

Hypothetically, the quality of the model is good enough for implementation. However the model can be further improved upon by drawing random sample of the sms text and its labels to create better train and test set. Additionally, one can use different algorithm like SVM, since the results are usually better than those obtained with regression model, but note that computational resources are needed for SVM. The use of deep learning is another option, which is a set of algorithms and techniques that use artificial neural networks to process data like the human brain does. Even though this method can make more accurate predictions than traditional machine learning models, it requires huge amount of training data to be effective and give better results than the methods used here.

References

1. Duan, P.: Winning solution code and methodology (2013), https://www.kaggle.com/c/amazon-employee-access-challenge/discussion/5283
2. Witten, I.H., Frank, E., Hall, M.A., Pal, C.J.: Data mining: practical machine learning tools and techniques. Morgan Kaufmann, 4th edn. (2017)