Assignment 1(Basic) - Data Mining Techniques

L. Prast 2628427, H. Choudhary 2754741 and N. Levi 2696717

Group 109: Vrije Universiteit Amsterdam

1 Introduction

In this report we will first explore a data set that was made during the lecture about us and give some insights about the data of ourselves. Afterwards we will discuss our findings of the famous Titanic data set and participate in a Kaggle competition of this data set. Lastly, we will give some insights in the more theoretical parts of data mining.

2 TASK 1: Explore a Small Dataset

In this section we explore the ODE dataset and use Machine Learning models to train and deduct insights from the data.

2.1 TASK 1A: Exploration

The given dataset has 304 instances with 17 feature columns. The features related to Machine Learning, Information Retrieval, Statistics, Databases, Gender, Chocolate, and position of the survey participant are categorical and rest of the attributes are numerical. We clean the data-set as follows -

- Remove instances where the feature column related to stress is empty or not a number. Mask the values so that numbers less than 0 are replaced with 0 stress-level and numbers greater than 100 are update with 100 stress-level.
- Remove the columns that don't contribute in analyzing our data-set. So, we remove features related to number of neighbors, whether people were standing or sitting while filling in the survey and timestamp.
- We map different words representing the same meaning for column related to course and activities/things that make a good day.

Stress Level Analysis

We analyze the distribution of Machine Learning(ML) knowledge and stress levels. Around 60% of the people in the class have taken a course on machine learning before. We evaluate mean stress levels for people who did a course versus people who had no prior ML course. The aim is to understand if there is relation between the two. Mean stress level for people with ML course is 44.5 whereas for students who didn't take ML is 53.8 We perform a z-test to check

if the difference is significant and with 95% confidence level we reject the null hypothesis that mean of both type of students(ML vs no ML) is same. Therefore, people who had some experience with machine learning models were less stressed. From 1 we see that there are more number of people in low stress level range who took a course in ML.
A similar experiment is conducted based on people who had a Database course versus who didn't. Around 52% people had a prior experience in Databases and mean stress level for these students vs with no prior course is around 44 and 53 respectively. A z-test is implemented and with a confidence level of 95% we can say that database feature has an impact on the stress level of the students.
A partial explanation for the above results could be that people who have worked with databases or some machine learning algorithms before could be more confident to play around with the data and train it to find some insightful results.

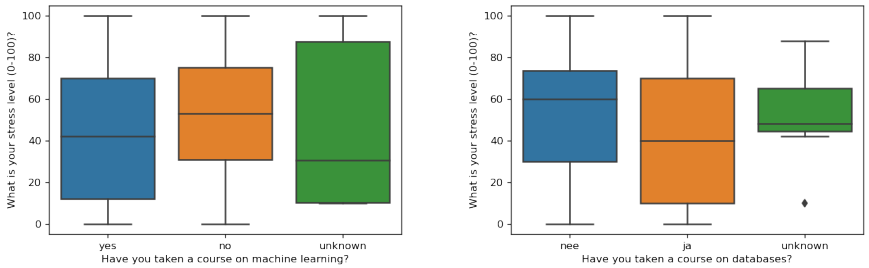

Fig. 1: Left Figure - Distribution of stress levels with respect to Machine Learning feature and Right Figure - Distribution of stress levels with respect to Database feature

Then, we explored whether there is a connection between the program the student is taking and the reported stress level. The results are displayed in 2

Top Mood Boosters

After cleaning the feature column "What makes a good day?" we plot the frequency plot of this attribute 2. Here, we can see that weather, food and sleep constitute as the most frequent mood boosters for the students.

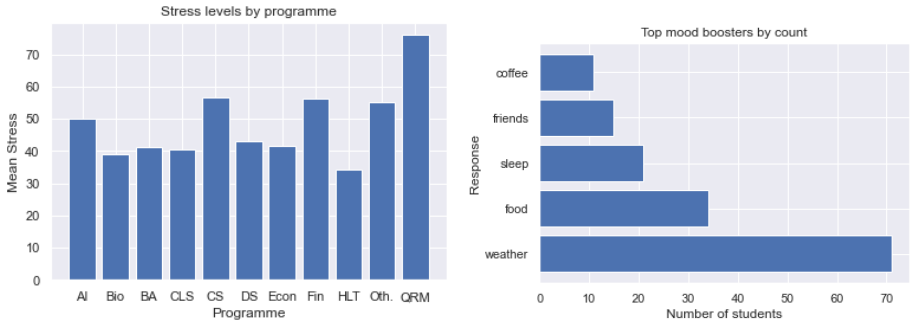

Fig. 2: Left Figure - Distribution of stress levels with respect to program and Right Figure - Frequency of mood boosters

Table 1: Responses to the "What makes a good day?" question and associated mean stress and requested prize levels

Good day response	Mean Stress level	Mean Prize request(€)
Other	53.196581	20921.802083
alcohol	43.857143	46.142857
chocolate	54.375000	1.500000
coffee	48.100000	133.444444
exercise	**60.090909**	16029.750000
food	44.288235	96.642857
friends	45.800000	39.750000
sex	**28.750000**	30.000000
sleep	54.857143	60.368421
weather	39.661972	52.622951

From the table, it is evident that respondents who reported 'sex' to the question of what makes a good day were least stressed. Furthermore, a response of 'exercise' was associated with the highest stress levels.

Association

A set of experiments are conducted to see if there any association among the features of the dataset. We find that 84% of the people who took Database course also took Statistics with support = 44%. Therefore, is a strong association between students took a Database course and Statistics course.
Database Course $\implies$ Statistics Course
Similarly, Database Course $\implies$ Machine Learning Course with support = 37% and confidence = 70%.

2.2 TASK 1B: Classification

After exploring the data-set, we were interested in seeing whether we could use some of the variables to classify another. The variables of interest were *Age, Stress , ML exp., IR exp., DB exp. and Stats exp.*. Before we could proceed, we made a small transformation to the data, on top of what was already described in 2.1. Namely, we transformed the date of birth variable into an age variable, calculating the difference in years between the current date and the reported date.

We then proceeded with a simple correlation test, which indicated that experience with Machine learning showed the highest correlation with the remaining variables, and so we proceeded with a classification where Machine Learning Experience was the response variable.

To begin with, we used a *Logistic Regression* algorithm. We split the data into test and training sets based on a 30/70 test and used a 10-fold cross-validation. We then continued with a Naive Bayes classifier and compared the results:

Table 2: Classification on the ODE data

Model	Accuracy(train)	Standard deviation(train)	Test accuracy
Logistic Regression	0.65	0.089	0.6744
Naive Bayes	0.63	0.13	0.6

As can be seen from the table, the Logistic Regression model was able to classify the ML experience variable better than the Naive Bayes. It should be noted that the result is still quite low, which we believe to be due to the low number of instances in the data-set, which were further reduced while cleaning the data from non-sensical answers.

3 TASK 2: Compete in a Kaggle competition to predict titanic survival

In this part of our report we are going to try to predict if passengers of the titanic have survived the titanic on the basis of their data. We will first explore the data then we will prepare the data and afterwards we will train multiple classification models to make our prediction of the passengers. To objectively test the model the total data set of the passengers of the Titanic is divided in a training part and testing part. We will train our model with the models, see which ones work best on the training data and afterwards use the best of those models on the testing data.

3.1 Exploration of the data

We started with checking if there was data missing in the data set. We decided to remove the cabin column entirely, since their is so much missing data. In the

age column a quarter of the data was missing, which we will fill later when we prepare the data. Afterwards we explored the data.

In fig. 3 we saw that the passengers on the Titanic were relatively young and that people in the the higher classes were older than people in the younger classes. We will use this to fill in the missing data of the age of the passengers when we prepare the data.

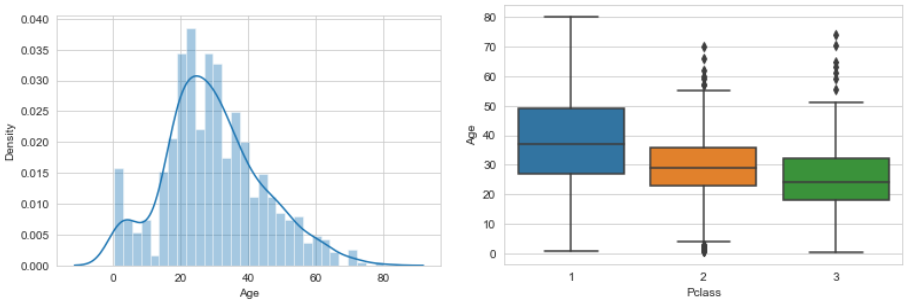

Fig. 3: In the left figure the distribution the age of the distribution can be seen and in the right figure the distribution of the age of the passengers in the different classes can be seen

In fig. 4 we can see in the left figure that women were more likely to survive the Titanic disaster and in the right figure we can clearly see that persons in the more expensive first and second classes had a larger survival chance.

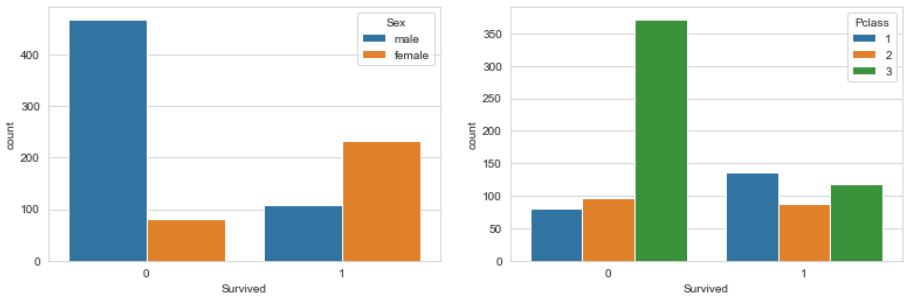

Fig. 4: In the left figure the distribution of the passengers that survived the sinking of the Titanic is seen separated by sex. In the right figure the distribution of the passengers that survived the sinking of the Titanic is seen separated by the class the passengers were in. In both graphs 0 indicates that a passenger has not survived and 1 that a passenger has survived.

3.2 Preparation of the data

We started by filling missing ages of the passengers with the mean age of the class that the passenger was in, which was 38 for class 1, 30 for class 2 and 25 for class 3.

In the dataset we have numerical variables like Pclass, Age, SibPp, Parch and Fare, but we also have categorical variables like Name, Sex, Ticket, Cabin and Embarked. We need to transform the categorical variables to some numerical feature in order to perform the classification task or we need to remove them. From the names of the persons we got the title and removed the names since the names are different for all persons except for maybe a couple. However, the titles are the same for a lot of passengers. In fig. 5 we can clearly see that the title of a person did matter for their survival chances.

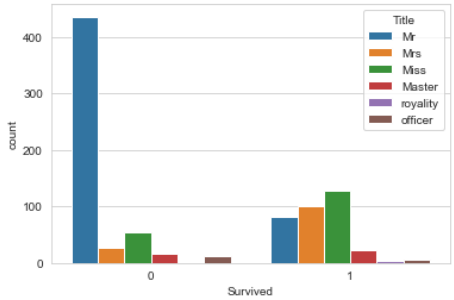

Fig. 5: The distribution of the passengers that survived the sinking of the Titanic is seen separated by the title of the passengers. 0 indicates that a passenger has not survived and 1 that a passenger has survived

For every title we made dummy variables and added them to the training data and the test data. We also made dummy variables for the categorical features sex and embarked. As said earlier we removed the cabin data of the passenger since too much data was missing and we also removed the ticket information of the passengers. There was no missing information about the tickets of the passengers, but when we made it in a numerical feature it did not increase our accuracy.

3.3 The models

Before we did submit our models to Kaggle to check their accuracy on the test data, we did a K-cross validation with $K = 10$ on the training data to see which models could predict the best if a passenger of the training data did survive the Titanic disaster. The results of this cross validation can be seen in table 3.

Table 3: The accuracy of the different models using 10-fold cross validation on the training data provided by Kaggle

	Accuracy	Standard deviation	95 % Confidence interval
Naive Bayes	0.804	0.043	+/- 0.032
Decision Tree	0.781	0.061	+/- 0.046
Logistic Regression	0.827	0.033	+/- 0.025
Gradient Boosting	0.831	0.048	+/- 0.036
Random Forrest	0.803	0.046	+/- 0.034

As seen can be seen in table 3 logistic regression and gradient boosting did predict the most accurate which passengers did survive. Therefore we trained those models to use them on the test data, which resulted in an accuracy of 0.768 for the logistic regression model and an accuracy of 0.780 for the gradient boosting model. The accuracy of our best model the gradient boosting model on the training data was 0.831 with cross validation, while the accuracy on the test data was 0.780. The explanation for this can be that in our training data provided by Kaggle there is a different percentage of people that survived compared to the test data, so our model could not predict as well for the training data as for the test data.

Our best score placed us a bit above the 75th percentiel of all competitors of the competition. So, there is still room to improve the data preparation and the model itself. Improvements could be for example using the information of the tickets in other way than we did or use different models and perform hyper parameterization on them.

4 TASK 3: Research and Theory

4.1 TASK 3A: Competition

In this section we will review a competition relating to data mining, and go over the approach of the winner.

The competition of interest is the Inria Aerial Image Labelling Dataset [1], originally proposed by Maggiori et al in [3]. We chose to look into this competition because of our interest in applying machine learning on remote sensing problems, which this competition addresses - Namely, detection of buildings from sattalite imagery.

The data-set consists of 810 km^2 of satellite imagery ranging from densely populated urban areas to small towns with a low population density. The data-set includes ground-truth labels for two semantic classes - building and not building.

Furthermore, performance in the challenge is measured as follows:

- Intersection Over Union of the building class
- Accuracy, measured by the percentage of correctly classified pixels.

The original paper employed a U-Net Convolutional Neural Network, and resulted in an accuracy of 93.54% and 55.82% for the IOU.

The submission that achieves the highest rating was given by Chatterjee in [2]. In his approach, Chatterjee proposed a novel network with the underlying architecture of a fully convolutional network, infused with feature re-calibrated Dense blocks at each layer [2], .The novel network that was used to expand on the original U-Net model used by the creators of the challenge in [3]. It does so by utilizing two aspects of the U-net architecture, namely; The *localization accuracy* of the U-Net architecture - A property that helps detect the "where" in the image, as opposed to the "what", and the *reduced feature redundancy* of the U-net architecture, which helps simplify the image properties and extract the most relevant feature. The novel ICT network uses these two aspects of the U-net model, and expands on that by using dynamic channel-wise feature re-weighting of the Squeeze-and-Excitation(SE) blocks. The ICT-Net architecture resulted in the highest score in the challenge, boasting an accuracy of 97.14% and and IOU of 80.32%.

We took the liberty to try working with the data provided in the challenge to train a U-net model. Our approach included a post-prediction processing of the output, where we used a cutoff value for each pixel to decide whether it belongs to the building class or not. The resulting model was then used on an unseen aerial image of a business district in Amsterdam centred around the Rembrant toren. The result is shown below:

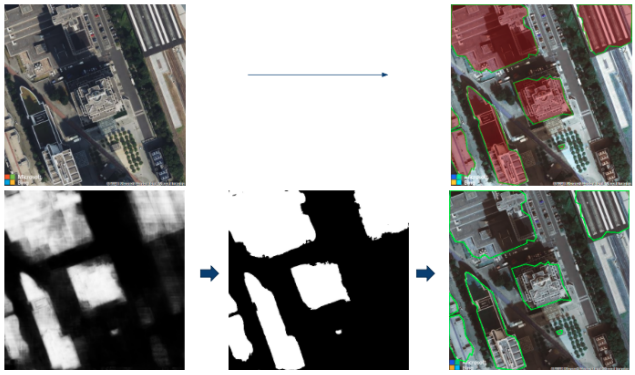

Fig. 6: Building detection by a U-net network

4.2 TASK 3B: Theory - MSE versus MAE

Mean Squared Error(MSE) and Mean Absolute Error(MAE) There are two types of machine learning problems - Classification and regression. They both have different loss functions/evaluation metrics. For classification a fit of the model can be evaluated by finding accuracy of the model and is quantified as fraction of mismatches in the target variable predictions. For regression models,

we can use Mean Squared Error, Mean Absolute Error, Root Mean Square Error etc. In this section we analyze two of the most widely loss functions MAE and MSE -

$$MAE = (\frac{1}{N})\Sigma_{i=1}^{N}|y_i - \hat{y}_i|, \qquad MSE = (\frac{1}{N})\Sigma_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{1}$$

Here, y_i is the true value and $\hat{y}_i$ is the predicted value of the target variable.

MSE Vs MAE Squaring part of MSE magnifies the large errors. Therefore, MSE is not robut to outliers and when there are outliers in the data or distribution is skewed, MAE should be preferred. Therefore, if the aim of the model is to have a good performance on majority of the data then MAE error is used as all the errors are weighted on the same linear scale. However, MSE can be preferred for ensuring that the trained model has no outlier predictions.

$$MAE = MSE \implies |y_i - \hat{y}_i| = (y_i - \hat{y}_i)^2 \tag{2}$$

Hence,

$$(y_i = \hat{y}_i \ \vee \ y_i = \hat{y}_i + 1 \ \vee \ y_i = \hat{y}_i - 1) \, \forall i \in N \tag{3}$$

Therefore in the above three situations MAE and MSE will be the same.

Regression model analysis with MSE and MAE Now we analyze MAE and MSE with a real dataset. We import the dataset "winequality-red.csv" from UCI Machine Learning Repository[4]. The data has 11 features that can affect the quality of the wine. This dataset has 1599 instances of real data. All the features have real numbers. We train the model to predict the 'Quality' of wine and analyze models based on the evaluations metrics MSE and MAE. Attribute Information - Input variables (based on physicochemical tests): fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol and Output variable (based on sensory data) : quality (score between 0 and 10).
We train the data using Decision Tree(Model-1) and Random Forest(Model-2) Regression models. We split the training data with test split = 0.3 and evaluate the MAE and MSE using k fold cross validation(k=10). In Model-1, we get MSE $\approx$ 0.57 and MAE $\approx$ 0.45. Then we train the data with a Random Forest regressor and the MSE and MAE are approximately 0.32 and 0.41 respectively. Therefore, we get low values of MSE and MAE as compared to model-1. This means Random Forest model fits the training data better than Decision Tree. However, this can also be an overfit so we need to evaluate our model on the test dataset as well. Models-1 gives MAE and MSE of around 0.53 and 0.74 on test data. respectively. Whereas as Model-2, the forest regressor has MSE $\approx$ 0.46 and MAE $\approx$ 0.40. Hence, Random Forest Regressor is a better model to predict the wine quality than Decision Tree.

4.3 Task 3C: Theory - Analyze a less obvious dataset

The data consists of text message from the UK which are already labeled as spam or ham. In this case ham means that is a bonafide message and spam means that is an unwanted message. To analyze text data with a classification algorithm we will need to transform the texts to numerical feature vectors. There are multiple modeling techniques to approach this type of data as Piek Vossen told us in our Text Mining lecture like decision trees with a vectorization called 'closest to rules', naive Bayes with a 'bag of word', support vector machines with 'data delineation' and decision trees with 'distance in feature space' [5]. In our study we used the naive Bayes modeling technique with the 'bag of word' vectorization.

Before we used this method we needed to do some transformations to prepare the data set. We removed all the punctuation in the texts and the most common stop words in English using the NLTK library, since both are similar for ham and spam. The goal of this transformations is to keep the words that are unique for each message to and then perform the classification algorithm on this stripped data.

There are a lot more data transformation that can be made like stemming, but we choose not to do this, because stemming is not guaranteed to work properly on text data, because in this type of data there are a lot of abbreviations or shorthand.

The stripped messages were vectorized with the bag of words method from the SKLearn library. Afterwards we also used the term frequency-inverse document frequency (TF-IDF) from SKLearn to give each word a weight, which is a statistical measure used to evaluate how important a word is to a document.

Now the data is transformed enough to use. We used the Naive Bayes classification algorithm to first train the model with the train data and afterwards use it on the test data. The confusion matrix can be seen in table 4 when we compared the prediction for the test data to the real information about if the text message were spam or ham.

Table 4: The confusion matrix of our Naive Bayes model for the prediction if the sms messages are spam or ham

	Predict ham	Predict spam
It is Ham	1207	0
It is spam	44	142

We have a good accuracy of 0.97 for our model, but what is also very important in the case of predicting ham or spam is that we do not have false negatives. In this case that would mean that we would have predicted that a text was spam and acted accordingly, while it actually was ham. The accuracy could be improved by using the earlier mentioned stemming or other text normalization methods.

References

1. Inria Aerial Image Labeling Dataset, `https://project.inria.fr/aerialimagelabeling`
2. Semantic segmentation from remote sensor data and the exploitation of latent learning for classification of auxiliary tasks, Chatterjee, Bodhiswatta and Poullis, Charalambos, Computer Vision and Image Understanding, Elsevier
3. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark, , 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 3226–3229, IEEE
4. Wine-Data, `https://archive.ics.uci.edu/ml/datasets/wine+quality`.
5. Lecture Notes, Text Mining Guest Lecture