Data Mining Techniques
Assignment 1 – Basic

Group 121: Don Caleb Severin[2165596]

Vrije Universiteit Amsterdam
d.c.severin@student.vu.nl

1 Introduction

This document will traverse three very interesting and distinct tasks. For this, we will utilize various tools and techniques learned in this course. We will explore a small ODI (Own Data Initiative) Data set and develop our first primary classification, regression models. Following this, we would participate in the Titanic Survival competition in hopes of predicting the survivors and finally, discuss some more theoretical topics to showcase our understanding and further grow our understanding of the capabilities produced by Machine learning.

2 Own Data Initiative Dataset

The ODI Dataset, at first glance, seems to be a questionnaire posed to gather two different types of information; for one part, it collects information that points to the knowledge of the students, and the second gathers seemingly personal or random facts about the students.

2.1 ODI Exploration

Once the data set was downloaded and loaded into Jupyter Labs and a notebook was created, my first instinct was to view the first five lines of the data set and the last five lines. This way, we can see the various columns and some of the proceeding data. The shape of the data was (304, 17), which translates to 304 rows and 17 column headers or features. Seeing that this data set is smaller, I initially thought that model complexity and ensuring we manage complexity to mitigate over or under-fitting.
Looking through the data structure and quality, we can easily see much work needs to be done on this data to gain any real insight or meaningful visualization truly. We see that the "Dtypes" are all objects, meaning the columns hold multiple Dtypes, so pandas return the base object as a reference. Which will make plotting numerical values fail or unable to transform without using regular expressions. There are four or more missing values in the data, which is minimal. I observed inconsistency in date and time variables and also a language & character inconsistency throughout**,** as seen in

Table 1. This data set, though rich with interesting personal data metrics, would need a lot of attention, but let's focus on the things we could see.

Question 1	Question 2	Question 3	Question 1
Yes	0	Mu	Nee
Yes	0	sigma	Nee
Yes	1	Mu	Ja

Table 1: language and Character Input Errors

My interests are generally peaked in the different behaviors regarding the various gender types. Are there any differences, similarities or predictable instances that can be drawn from the data? We will be working in the column "What is your gender" to generate some visuals to better generate a working theory.

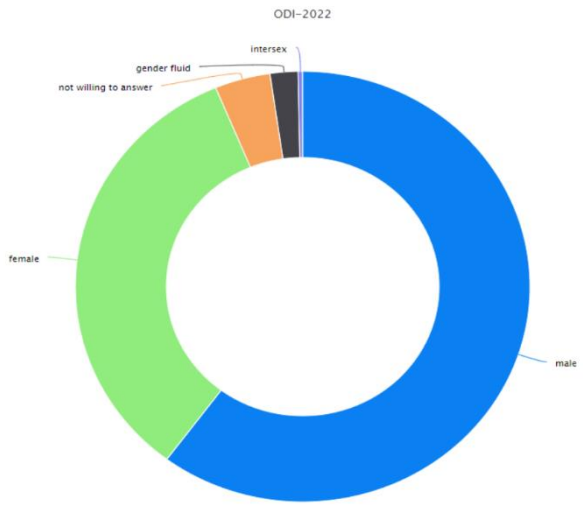

Figure 1: Gender distribution

In the above figure, we see a variety of genders; the majority is the male gender. However, we also see new genders in the minority, so it's still interesting to see. I understand that this column can be represented as a bias and potentially lead to a lower result in decision making, but the absolute simplicity and findings between them may be illuminating for future application of targeting and course suggestions. As we continued our exploration through the ODI dataset, we wanted to learn more, and, in fig 2, the "Tijdstempel" or "date and Time" columns were interesting to see how they corresponded with our gender column.

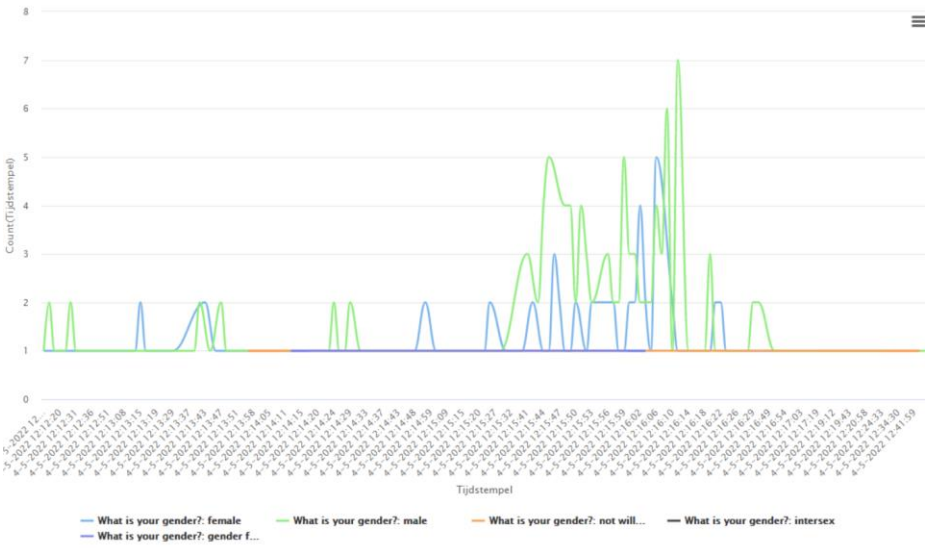

Figure 2: Time frequency in relations to gender

From this plot, we can see most of the questionnaire interactions happen between 12:15:21 - 12:16:29. The highest interactions are from the male gender, but we see the most consistent and early forms of interactions happening by the female gender. Does this indicate a more preparedness for the woman? Or does it show a tremendous interest in the questionnaire from the men? Does this preparedness or interest in scientific courses relate to the high-stress levels? Let's keep exploring these assumptions further.

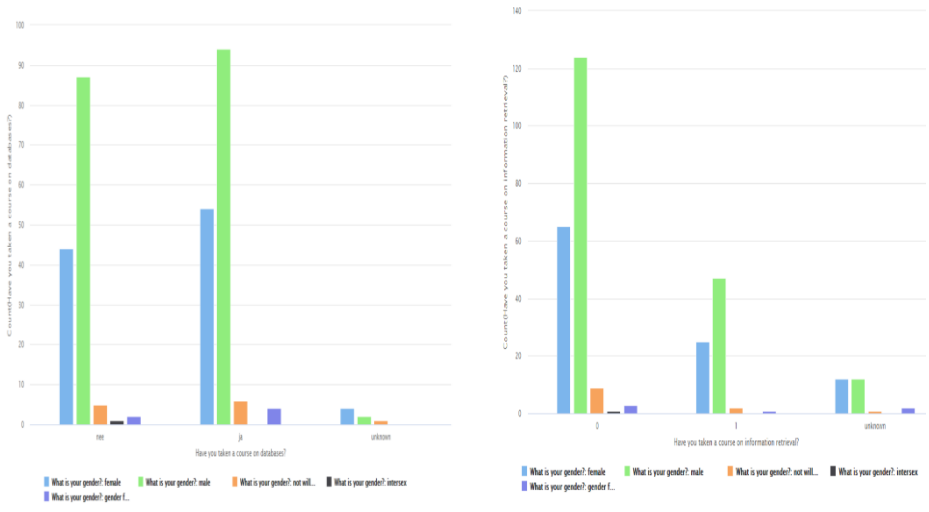

Figure 3: DB & IR Knowledge base on gender

We wonder what courses are being taken by the respective genders in the above fig 3 shows the number of men and women that have completed a course in Databases and Information Retrieval. They look very similar in Yes and No distributions.

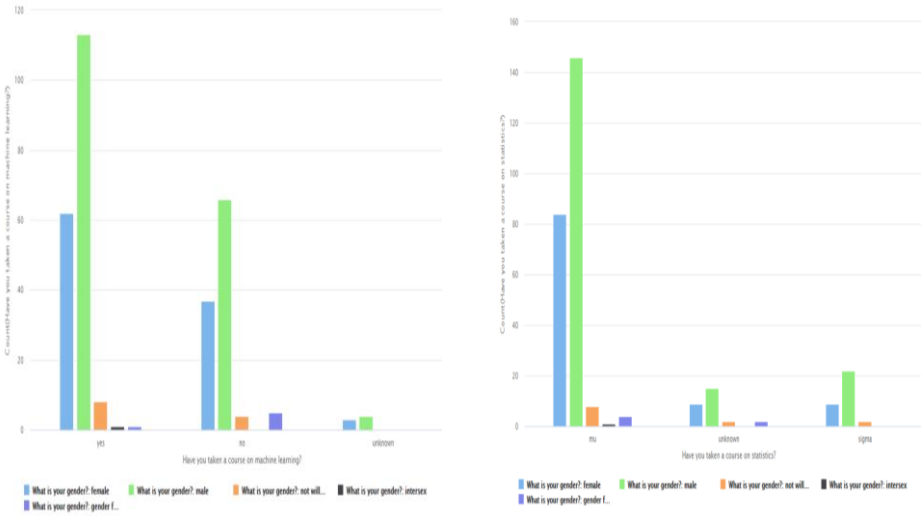

Figure 4: ML & Stat Knowledge Base

We see in fig 4 the metrics for men and women completing a course in Machine learning and Statistics. We see that most men and women take statistics, which is valid for machine learning. The data shows that men and women are taking the same classes and courses.
With this understanding of various similarities between the two significant genders, let's try to make predictions.

2.2 ODI Classification\Regression

With this understanding of various similarities between the two significant genders, let's attempt to make a prediction on our data using the rapid miners Auto Model Feature to streamline the process and gain takeaway information. I decided the stress levels of students were more important even though there were concerns about the data ranges and the system's ability to predict based on data fed to the model.

We began by choosing our target feature and the "What is your stress level (0-100)" column. Some modifications were made to this field by re-representing some of the data to complete numbers or replacing scientific notation with standard numbers.

In fig5 we see the prediction statistics, and although the numbers are disappointing where the Naïve Based Model scored 7,7% in the fastest time. However, the score could be expected due to the quality of the data not being as robust.

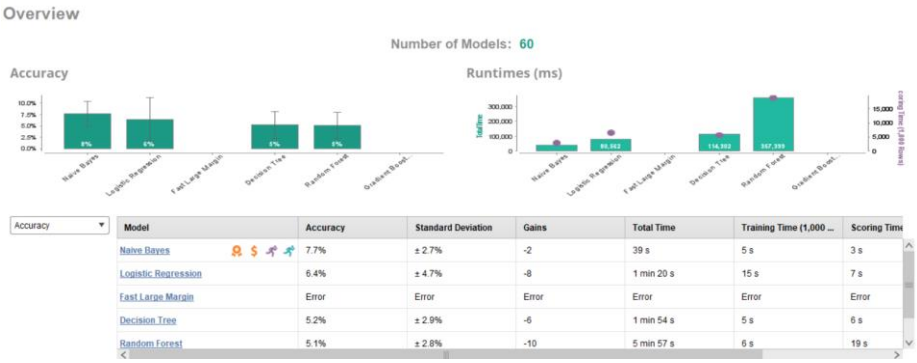

Figure 5: Model Accuracy and Measurements

Finer tuning and transformations to the data set could have raised the score along with outside metrics added to the dataset, but for this report, we have demonstrated exploration and classification modeling using rapid miner.

3 Kaggle Competition: Survival Prediction

For our Kaggle competition, we are tasked with investigating the provided train data of survivors on the Titanic and trying to make an accurate prediction using machine learning techniques.

3.1 Data Preparation

The Titanic Train Dataset, at first glance, looks clean and organized, but let's take a closer look at the 17 attributes and 819 rows to see the quality. This way, we can then understand the transformations needed to be made. Categorical features in the data set are Survived, Sex Embarked, and Pclass is Ordinal. The numerical(continuous) data is Age, Fare Numerical(Discrete) SibSp. There are alphanumeric data types in Ticket and Cabin. Immediately we see Age, Cabin, and Embarked holding 866 missing values collectively. I used pandas to look at the distribution of the data to get some insights; see table 1, we can see the distributions of figures showing some relevant information.

We see a 38% survival rate; we have a mean age of 29 which can suggest a relatively low percentage of older passengers. If that is true, why is the survival rate so low? Do we look at the "Pclass" = Ticket Class, and more than 50% of passengers were in the lower third, does this account for the survival rate? Is this the part of the ship that sank first, and people in these classes had a more challenging time escaping?

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

Figure 6: Descriptive Analysis Titanic Data

We find more compelling evidence to support our hypothesis that the “Pclass” &” Gender features can help us determine the survival of our passengers. For example, we see in the below fig7 that “0 = Not Survived” and “1 = Survived” that males and females in the 3rd class had more deaths than those in the 1st class.

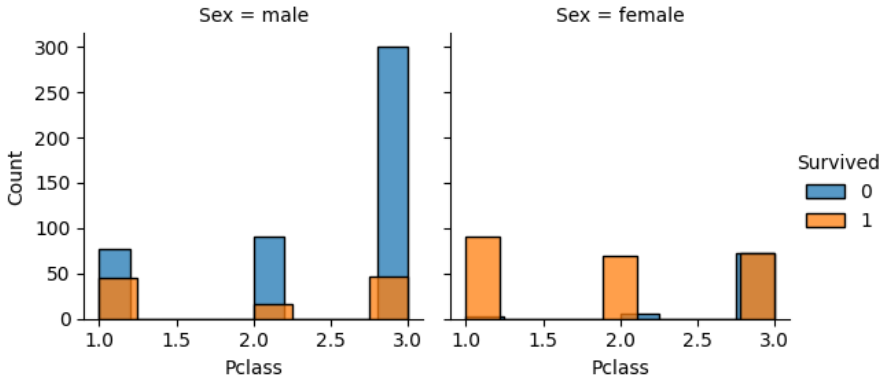

Figure 7: Survival Metrics of Pclass and Gender

The following figures take a ball parch approach to make some conclusions that I still find relevant for the “SibSp” and “Parch” features; in fig 8, we uncover that between the aged of 20-40, the more Parent-Child relationships, the less likely survival becomes for some gender group (I am gearing to the male demise in this analysis.

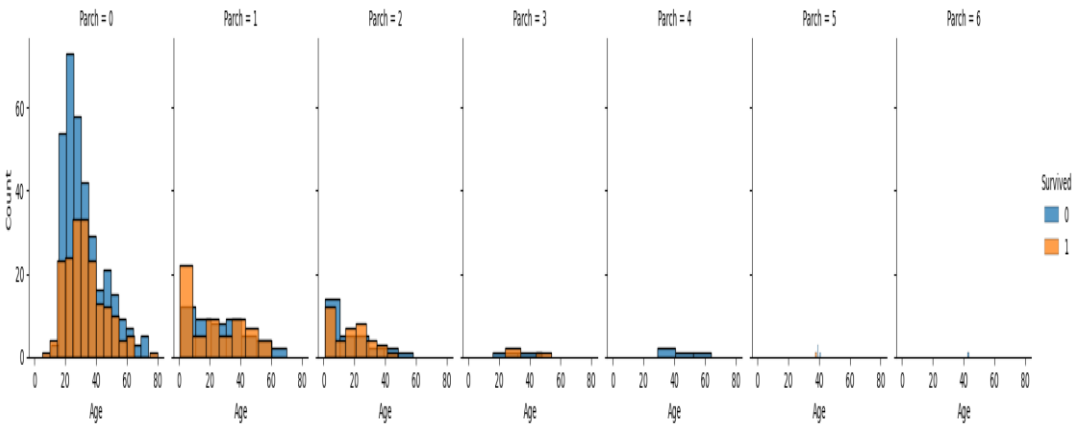

Figure 8: Survival Metrics Parch and Gender

This analysis also bears weight when we look at the same relationship, but with Sibling Spouse additions, being alone on the titanic may have given you a higher chance of survival.

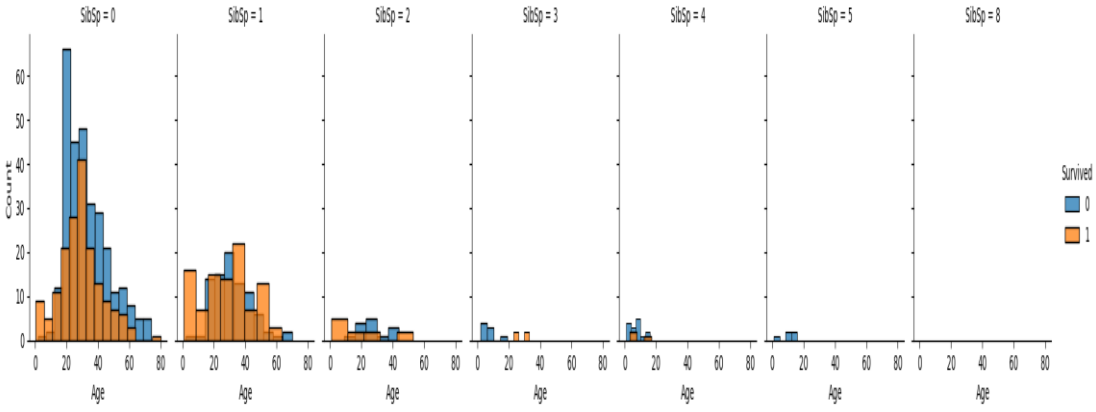

Figure 9: Survival Metrics SibSp and Gender

For this data set, we want to perform some transformations and preprocessing before moving into building the classification models. If we look at our findings, we have to manage the missing Values. However, we have firstly deemed some features unnecessary to our bottom line. Hence we will drop the following features:

- Embarked
- Cabin

- Name
- Fare
- Ticket

These features, though, may allow us to engineer new features, for now, little empty weight in the survival of our passengers. Removing these features has already eliminated most of our missing data, and we can interpolate for the age data since we are missing less than 50% of the age data. We will also be modifying the "Sex" feature from a non-numerical value as it is observable but not measurable, which I think is essential for our task since it bears weight in the survival rating.

3.2 Titanic Data Set Classification

We will be moving into the machine learning aspect of this competition and applying predictive modeling to our dataset. In our prediction, we have chosen four very different kinds of classification due to the success rate in other projects, and overall popularity in the Skit learn community.

- Decision Trees
- Generalized Linear models
- Random Forest
- K-Means

We will discuss our highest scoring predictors in this paper, and in Table2, you can see the overall scores of the four classifiers.

Classifier	Accuracy	Kaggle Accuracy
Decision Tree	83.73%	N/A
Random Forest	84.62%	77.51%
Gradient boosted Trees	83.95%	72.96%
Generalized Linear models	79.01%	N/A

Table 2: Model Accuracy

Random Forest and Gradient Boosted Trees both scored the highest in our simulation. However, once loaded into Kaggle, the score dropped in accuracy. It is unknown why, but one can speculate it may be the way Kaggle scored in these competitions. However, Random Forest is a robust classification algorithm capable of working with large volumes of data. The random forest also employs bagging techniques over the Gradient Boosted Trees Boosting. Bagging translates it to low model complexity and limited overfitting in this experiment leading to a high score over Gradient Boosted Trees. This is due to slight differences in how things are accomplished in each algorithm.

4 Research and Theory

In this final part of our paper, we will go into the theoretical understanding of how these algorithms work and how they are used to make societal differences

4.1 State of The Art Solution

We look at a Kaggle competition for personalized H&M fashion recommendations. This competition started on February 7th, 2022. H&M Invites developers to develop a recommendation system based on data from previous transactions, as well as metadata such as garment type and customer age, text data of product description, and product images. So we assume there will be some complex algorithms for natural language processing and deep learning applications. The competition Winner goes by Sinkin from Tokyo, Japan, shared his solution. Our candidate approach in his documentation mentions the importance placed on retrieval strategies, feature engineering, and choosing a Gradient Boosted Decision Tree combination. In reading the overview of his document, we see the candidate having a proper understanding of the domain problem and identifying seasonal trend options, and making decisions based on this. Our candidate also utilized optimization techniques like TreeLite to accelerate interface speeds. When we compared this to the second-place candidate in this contest, we didn't see adequate care in feature generation or domain analysis which I think were the most important aspects of our first candidate's solution.

4.2 MSE Versus MAE

Mean Squared Error indicates how far a set point is from the regression line. This constitutes the amount or distance of errors and then squaring them, which is needed to remove negative signs. This error calculation lets you know how close you are to finding the best possible fit.

Formula:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Mean Squared Error indicates how far a set point is from the regression line. This constitutes the amount or distance of errors and then squaring them, which is needed to remove negative signs. This error calculation lets you know how close you are to finding the best possible fit.
Formula:

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

These two calculations may seem similar, but their motivations are different. For example, one may guide you in selecting the best fit and provide accuracy in depicting the "True" difference between two indexes of different or varying levels. Nevertheless, we see them producing similar results when exploring the performance of predictive models and allowing you to make tweaks and better predictions by developing feature-rich models.

4.3 Analyze Less Obvious Dataset

For our SmsCollection data set, we are asked to figure out what techniques would best suit this type of Text data, and we have found three valuable variations. The first is Naïve Bayes, a popular choice for classifying text-based data. You can use the Multinomial Bayes algorithm for smaller data sets like this and get good results. This algorithm works well because it calculates conditional occurrences of two events. So simply, it would calculate the occurrence of each tag in any text and output the highest probably tag.

For this to be accurate, the underlying text or vector representing this text must have substantial metadata on the probability of this text being used in a category. Finally, when all else fails, we turn to our following technique, Deep learning architectures like Convolutional Neural Networks and Recurrent Neural Networks. Though this method is much more suitable for Much larger data sets, it is still noteworthy at the end of the spectrum.

In terms of data transformation, if we look at the data set, we see a lot going on—misspellings within the data asset, alphanumeric properties and currency symbols, encodings, and capitalization inconsistencies. So we would have to account for missing data and the lack of features there off to engineer better ones for predictions.

References

1. Author, A. M., & Author, S. G. (). Introduction to Machin Learning with Python. Location: Boston, Beijing, Farnham, Sebastpool, Tokyo
2. H&M personalized fashion Recommendations. Kaggle. Retrieved June 6th, 2022, from https://www.kaggle.com/competitions/h-and-m-personalized-fashion-recommendations/discussion/324070