FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Going zero waste in canteens: Exploring food demand with data analytics

Diogo Xavier Ribeiro Pereira

MASTER'S DISSERTATION

Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Assistant Professor Vera Miguéis

July 18, 2018

Going zero waste in canteens: Exploring food demand with data analytics

Diogo Xavier Ribeiro Pereira

Mestrado Integrado em Engenharia Informática e Computação

July 18, 2018

Abstract

Nowadays, almost all catering services' food demand management is based either on intuitive managers' guesses or on modeling customers' behavior only as a function of time, which in turn may arise problems such as food consumptions' underestimation or overestimation. The latter leads to food waste, which is a serious problem of today's society.

Therefore, in order to reduce such waste arising from mismanagement, this paper aims to describe a system capable of, under several circumstances, predicting daily food consumption, i.e number of meals to prepare for each dish type offered by a canteen - meat, fish or vegetarian. This system will be firstly designed taking into account the surrounding environment of the Faculty of Engineering of the University of Porto's (FEUP) canteen, from which specific factors, influencing food consumption, emerge. Aspects such as the menus available, weather conditions, proximity of holidays, students' timetable or even special weeks and events, are included in the system proposed. This data concerns a period of two years, 2016 and 2017. This study explores the use of advanced data mining techniques such as Random Forests (RFs), Support Vector Regression (SVRs) and Artificial Neural Networks (ANNs). Such techniques will be applied in order to find the three best performing models capable of predicting meat, fish and vegetarian daily consumptions.

The definition of the best performing models was conducted based on the mean absolute error (MAE). Thus, it was concluded that for predicting meat consumptions, RFs achieved the lowest MAE, whereas for predicting fish and vegetarian consumptions, SVRs outperformed the remaining. Such models predictions resulted in an average monthly waste five times lower than that of obtained by the canteen's management team. Thereby, it is possible to state that the predictive system developed contributes to a significant reduction of food waste in the canteen used as case study.

Resumo

Hoje em dia, na tentativa de encontrar um equilíbrio entre o aprovisionamento e a procura de comida, quase todos os serviços de *catering* se baseiam na intuição dos gestores desses mesmos serviços ou, por outro lado, tentam modelar o comportamento dos clientes unicamente em função do tempo. Tais práticas de gestão estão na origem de problemas como a subestimação ou sobrestimação da quantidade de refeições diárias a preparar. Por sua vez, a sobrestimação resulta em desperdício alimentar.

Portanto, por forma a reduzir o desperdício resultante das más práticas de gestão referidas anteriormente, este estudo propõe um sistema capaz de, sob várias circunstâncias, prever o consumo diário de cada tipo de prato oferecido, normalmente, por uma cantina - carne, peixe ou vegetariano. Este sistema foi desenvolvido tendo em conta o ambiente em que se insere a cantina da Faculdade de Engenharia da Universidade do Porto (FEUP). Como tal, fatores exclusivos desse mesmo ambiente e suscetíveis de influenciar o consumo diário, puderam ser identificados. Fatores como, condições climatéricas, proximidade a dias de feriado, horários dos alunos da FEUP ou mesmo semanas e eventos especiais, estão contemplados no sistema proposto. Estes dados dizem respeito a um período de dois anos, 2016 e 2017. Este estudo explora o uso das mais avançadas técnicas de *data mining* presentes na bibliografia. São elas as *Random Forests* (RFs), *Support Vector Regression* (SVRs) ou mesmo *Artificial Neural Networks* (ANNs). Estas técnicas foram aplicadas com o intuito de encontrar os três modelos mais precisos na previsão do consumo diário de pratos de carne, peixe e vegetariano.

Para que esses modelos fossem encontrados, foi calculado o erro médio absoluto para cada modelo em específico. Concluiu-se que, no que diz respeito à previsão do número de pratos de carne, foram as RFs que obtiveram o erro mais baixo, ao passo que, para a mesma previsão, mas desta vez no que concerne ao peixe e vegetariano, foram as SVRs que obtiveram o erro mais baixo. As previsões dos três modelos permitem contribuir para uma redução, em cinco vezes, do desperdício médio mensal, quando comparado ao desperdício resultante do método de gestão da procura atualmente aplicado pelos responsáveis da cantina. Assim, é possível concluir que o sistema desenvolvido contribui para uma redução basatante significativa do desperdício alimentar na cantina.

Acknowledgements

Despite the dissertation project, for its academic purpose, being an individual work, there are contributions, direct or indirect, that should be emphasized. For this reason, I would like to express my sincere gratitude:

To Professor Vera Miguéis, supervisor of this dissertation, for all the scientific, personal and interpersonal competence demonstrated throughout the development of this project. For all the constructive advices and criticism, for all the willingness in helping and clarifying doubts in a timely manner, for all the generosity and understanding. In conclusion, it would not be possible to achieve so successfully the main objective of this project without the masterful supervision of the Professor.

To my parents for all the unconditional support, understanding and motivation with which they presented me, not only during the development of this dissertation, but throughout my academic and personal life.

To my brother, who has gone through all this and who, as such, was able to advise me in the most effective and timely way possible.

To my remaining family for all the presence, both spiritual and in person, for all unconditional support, understanding and patience.

To the SASUP services team for having received me at their infrastructures and as such contributing for that all the data necessary to carry out this dissertation was duly collected. Especially to Ana for the continued availability and willingness to clarify any doubts about this same data.

To Luisa for supporting me in the good and bad moments. For having advised me in the most wise and responsible way possible.

To my friends for not misleading me during the development of this dissertation and for all their unconditional support, motivation and availability.

To all those mentioned above, in general, without their unconditional presence, it would not have been possible to conclude this dissertation in a successful way.

Diogo Pereira

In honor of my grandmother

"Antes de se ser bom, é ruim."
Carminda Portela

Contents

List of Figures

LIST OF FIGURES

List of Tables

List of Equations

Abbreviations

ANN	Artificial Neural Network
API	Application Programming Interface
ARs	Association rules
CART	Classification and Regression Trees
CICA	Centro de Informática Prof. Correia de Araújo
CPU	Central Processing Unit
DT	Decision Tree
FEUP	Faculty of Engineering of the University of Porto
FSC	Food Supply Chain
LR	Linear Regression
MAE	Mean Absolute Error
MLP	Multi Layer Perceptron
MLR	Multiple Linear Regression
MR	Multiple Regression
MSDT	Microsoft Decision Tree
NCEP	National Centers for Environmental Prediction
ReLu	Rectified Linear units
RF	Random Forest
SASUP	Serviços de Ação Social da Universidade do Porto
SES	Simple Exponential Smoothing
Sigarra FEUP	FEUP's platform and information System
SLR	Simple Linear Regression
SMA	Simple Moving Average
SVM	Support Vector Machines
SVR	Support Vector Regression
UP	University of Porto
U-statistic	Theil's U-statistic
WMO	World Meteorological Organization
WWO	World Weather Online

Chapter 1

Introduction

This work arises from the need of improving the way catering services - e.g canteens - are managed. Such management includes, in particular, the way canteens' managers plan their daily work. In order to perform efficient planning, canteen managers should be able to forecast daily food demand accurately [RS03].

Such forecasting, nowadays, is mostly made based on intuitive guesses by canteen managers, based on the analysis of historical demand or based on inference rules. These rules are based in the logical principle that if a given set of premises holds true then the conclusion is likely true. These simplistic ways of estimating demand can lead to problems such as underestimation or overestimation of the number of meals to be prepared, since daily food demand is affected by numerous factors.

On the one hand, the underestimation leads to the food running out before customer demand is satisfied. To overcome that, it is forced the improvisation of a menu as a whole or even of a specific item from a given menu, regardless of its quality or cost. That improvisation requires extra work by canteen's staff, which results in accumulated stress emerged from working under pressure or against time [RS03]. Likewise, customers do not see their primary menu choices being corresponded, which can decrease their satisfaction and therefore, in a long-term basis, the canteen's market share [RS03] due to declining demand by those same customers. On the other hand, overestimation leads to food leftovers, which in turn can be reused to integrate another day's menus, therefore reducing their quality. This integration process requires food to be stored for, eventually, several days, which increases its likeliness to contamination. Besides that, food resulting from leftovers might also be wasted. Hence, overestimation results in additional costs because of leftovers' handling - for instance their transport to landfills - and in the decreasing of customer satisfaction.

Thus, one can conclude that forecasting food consumption may affect canteen's financial and logistics management, customer satisfaction, employee morale and managers confidence [PM96].

Moreover, food waste resulting from canteens' leftovers contributes to the 1.3 billion tonnes of food annually lost or wasted [GCS+11] throughout the Food Supply Chain (FSC), that is, from initial agricultural production down to final household consumption. Such waste is equivalent to a third of the food produced globally for human consumption every year. It also entails the waste of resources used in its production such as land, water, energy and raw material.

Food waste has a great impact in the global warming, since when food leftovers are thrown to landfills, they start to rot due to anaerobic decomposition, producing harmful emissions to the environment, such as methane - a molecule twenty-three times more harmful to the environment than carbon dioxide.

With the main focus on reducing food waste and therefore its environmental impact, this dissertation aims to describe a system capable of predicting daily food demand for a given dish type, that corresponds to the number of meals to be prepared in each day. Such forecasting varies from context to context, from where specific factors impacting daily consumption may arise.

Hence, this system will be built with respect to the FEUP's canteen. Demand in school canteens is affected by some exclusive variables, including semester-to-semester population changes, menus contained in the daily menu list, special student activities, and also by other common variables such as day of the week, weather conditions and others [SM95]. Thus, before the development of such system, all the relevant factors influencing customers' affluence to the FEUP's canteen, are identified, in order to increase system's prediction reliability. Since there are three main type of dishes available in everyday's canteen daily menu list - meat, fish and vegetarian - three predictive models are developed so that predictions on each dish daily demand can be made. These models were implemented using advanced data mining techniques, such as RFs, SVRs and ANNs [TSK05].

The models proposed, besides lowering the number of wasted resources associated with all the FSC stages, can contribute, in an indirect way, to food waste decrease - the main goal of this work -, which in turn reduces its impact in the global warming, and thereby, contributing to a more sustainable planet Earth. Furthermore, these models might contribute to an improvement in the canteen's management, finding the optimal balance between supply and demand. Herewith, resources, both financial and logistical, can be saved, which in turn are likely to be used in menus' confection innovation or even in new ones' introduction. This may enable, in a long-term basis, an increase in customers satisfaction as well as the canteen reputation, since customer affluence might increase and so is the canteen's market share.

Chapter 2 addresses the literature review related to this domain, whilst chapter 3 focuses in describing the methodology applied to solve the problem described and the data used in the study. Chapter 4 discusses the results achieved and its reliability when comparing them to the real and observed values. Discussions in chapter 5 are made to observe to which extent the predictive models proposed reduce food waste comparatively to the waste induced and verified by the canteen's management team. Finally, chapter 6 presents conclusions and also related potential future work.

Chapter 2

Literature review

Despite in the field, most catering services use very simple methods to estimate the demand, the literature reveals that, statistical models or even advanced data mining techniques are increasingly being used in order to improve production planning, that is, forecasting food consumption properly and efficiently [MMB91]. In such a way, food menus' underestimation or overestimation and its associated costs can be considerably reduced. Usually quantitative models outperform intuitive models or guesses [MMB91].

2.1 Existing methods for prediction

There are two types of quantitative models when it comes to food demand forecasting issues. Time series models, by one hand, focus on analyzing time series - a collection of data points distributed along time axis, for a specific variable, in this case, historical food demand data. This type of models may be useful to find out issues such as seasonality - periodic fluctuations in time series. They are the most appropriate when one can describe general patterns or tendencies regardless of the factors affecting the variable being predicted - daily food demand for a given dish type, in this specific case. On the other hand, there is a second type of methods which bear on a different approach, i.e. causal models, which are concerned on retrieving final predicted variable's value through identifying causal-effect relationships, between each predictor variable - factors affecting predicted variable's value - and the predicted variable. For instance, customer demand or food demand can be influenced by factors such as weather conditions, menu prices, seasonality, etc. An example of these models are those based in simple linear regression (SLR) or multiple linear regression (MLR). Such methods build a predictive causal model, mapping and capturing causal relationships between the variable being predicted and a set of predictors [SJH13].

A time series can be either stationary or non-stationary. If a time series is stationary then its properties, mean and variance, will not depend on the time. On the other hand, time series with trends are non-stationary, meaning that such trends will affect the value of the time series observed

in a certain instant of time. A time series can also include seasonality or not. [PA09] defines seasonality as "a repeating pattern within each year, although the term is applied more generally to repeating patterns within any fixed period". Since sales of products across food facilities are directly related to the choice or not - food consumption - of those products, seasonality can be caused by annual weather patterns, major holidays or festivals, regular occurrences of sport activities, etc [LBS+01].

Since most of the statistical methods, used for prediction, assume that time series is stationary, non-stationary time series are usually transformed - via mathematical transformations - into stationary time series. Such transformation then, makes prediction relatively easy to perform: one simply conclude that time series' statistical properties will be the same in the future as they have been in the past. Therefore, stationarizing a time series will allow one to obtain meaningful statistical properties - mean and variance -, which in turn could be useful in describing future behavior.

One may prefer to implement prediction systems through time series models or linear causal methods since they represent a low cost option and are easy to interpret. Although, given the randomness of today's real world problems, such models do not allow capturing relationships between predicted and predictors variables in such an accurate way. Therefore, nonlinear regression techniques, such as ANNs, can be seen as a suitable alternative in order to overcome such problems [GEM98]. These techniques, in order to perform the least erroneous models, usually have their hyper-parameters being tuned.

2.2 Related work

The literature already accommodates some studies that explore the potential of quantitative methods to support the estimated of food demand. For example, [MMB91] evaluated simple time series models, such as Naive-Method, Simple Moving Average (SMA) and Simple Exponential Smoothing (SES), to predict food demand. This study took place in an university dinning hall. Historical data, with three years long, regarding customer consumptions, were collected. Information concerning the choices of the customers, from a set of possible combinations of food items, was also considered. This data altogether enabled to determine the level of preference for a certain item, when available on the daily menu list. This preference measure was calculated by dividing the number of times that item was served by the total servings of the combination, over all replications, in which it was present. Then, giving that historical data as input to the methods referred in the beginning of this paragraph, it was possible to estimate the demand. Such estimation was then decomposed in the demand for each food item, through multiplying the demand by the its preference level.

[RS03] identified the most appropriate method to predict meal consumptions, during a whole semester, for an university food service facility. Most of the methods used were based in time series, such as naive model 1, 2, and 3, SMA and double moving average, SES, double exponential smoothing, Holt's and Winter's methods. SLR and MLR models were also applied. The

requirement to choose the most appropriate and accurate method was either its simplicity of utilization for canteen managers but also its accuracy level. Thereby, in order to estimate accuracy, metrics such as mean absolute deviation (MAD), mean squared error (MSE), mean percentage error (MPE), mean absolute percentage error (MAPE) and also root mean squared error (RMSE) are used. Theil's U-statistic (U-statistic) was also used, in order to compare naive models against the other methods (see [SCJ98] for further explanation in such forecasting and accuracy methods). Data with regard to 13 weeks of historical demand was collected. Such data was adjusted in order to take into account special circumstances such as weather, special events, holidays, etc. However, when applying time series models there were not considered variables other than the historical data of each dish demand. It was concluded that Naive-1 model had the worst accuracy since it predict next day's meal counts based on the previous day. LR had the second worst accuracy since this model considers only one variable as predictor and, consequently, it was not able to catch the change arising from seasonality. On the other hand, MLR had the best accuracy (MSE = 382 and U-statistic = 0.78) since it considers many predictors. In terms of accuracy, MLR outperformed other methods because it fitted the time series data with seasonality. In this study, the forecast method chosen as the most appropriate to use was based on its accuracy but also on simplicity, since a common canteen manager does not have much knowledge or time to explore the models. Thus, despite MLR have gotten the best accuracy, the one model chosen as the most appropriate was Naive-2 - uses the last week demand to forecast the next week demand -, obtaining the third highest accuracy and being the second easiest to use - easy to handle and implement.

[BS09] focus on three main goals. Firstly on identifying the top four factors affecting food consumption in refectories from an university. Then, to build a system capable of predicting daily consumption in the combination of the four types of dishes contained on the daily menu list. Finally this study aims to discover frequent associations in consumption. To accomplish so, supervised data mining techniques such as decision trees (DTs) for regression - Microsoft Decision Tree (MSDT) -, and association rules (ARs), were used. While ARs allow one to discover hidden relationships, correlations and descriptive attributes, DTs can help in predicting new cases based on old ones [TSK05]. Thereby, data for one year long, related to daily food consumptions, was collected. This data contained the four types of dishes, day of the week by name, the day and month numbers, and also a boolean variable indicating whether the day is an holiday or weekend, *1* if it is, *0* otherwise. In order to measure prediction performances, the R^2 metric was used (see section 3.3.3 for further details on this metric).

[BS11] once again in the same context, but this time, data collected corresponded to a 2-year time period. In addition to the variables used in the study described on the previous paragraph, data records kept track of each menu's calories. There were used also DTs for regression, such as Classification and Regression Tree (CART), Chi Squared Automatic Interaction Detection (CHAID) and MSDT. Since most of the variables were continuous or discrete, methods considering multiway splitting of attributes performed better than those considering binary-splitting, proving once again that the best method always depends on the problem's context and also that DTs for regression perform well when it comes to predict food consumption.

DTs are often compared to ANNs since both can catch nonlinear relationships between predictors and predicted variables, even if they are not predefined. Due to the lack of robustness to outliers, suboptimal performance and hard interpretation of DTs when lots of variables are given as input, RFs were proposed [Bre01] in order to try overcoming those issues ([AMR17]). Thereby, given the lack of studies applying ANNs and RFs in predicting food demand for a food service facility, such techniques were applied in this work. Allying that to the fact that there is no such thing as "the best method for all contexts", also SVRs were used so that accuracies' comparisons between predictive models can take place.

[ANJ16] developed a literature survey about appropriate methods to predict customer demand. They also provided a list of variables that can be, eventually, used as predictor variables, such as weather (rainfall level, temperature, description), time (month, week, day of the week, hour), holidays (national, municipal and school holidays), special events and historical demand data. After that, such methods are briefly described and articles in which they were applied are also given. Therefore, Multiple Regression (MR) and ANNs were identified as the most appropriate methods to predict customer demand, since both are able to receive as input multiple predictor variables. Despite MR models might fail to clarify relationships between predictors and predicted variable due to multicollinearity - phenomenon in which one predictor can be linearly predicted by the other - among predictor variables, such models can translate also nonlinear patterns in data, whereas time series models, for instance, can not. The same holds for ANNs, even if it is hard to find a good enough neural network architecture, since it requires to find the appropriate number of layers and their amount of nodes, as well as selecting transfer functions of middle and output nodes [ZH].

In [ZH], the state-of-the-art regarding ANNs application reliability in forecasting is presented. Since today's real world systems are often nonlinear, methods capable of generalizing well on unseen data must be applied. That said, this study shows why ANNs may be preferable over traditional statistical methods when the underlying mechanism is nonlinear. There is given an overview about areas in which ANNs find applications. Finally, it includes insights about neural networks modeling issues and techniques, as a way of finding the optimal network architecture.

Similar to ANNs, SVRs are being increasingly used when it comes to forecasting. In [LP05] it is proposed a methodology to forecast customer demand, through SVR analysis. Such methodology is based on the statistical learning theory ([Vap95]) - inferring a predictive function through training data, as a mean of predicting unseen data - and its SVR applications. It is also given an overview of a method for tuning SVR hyper-parameters and finally some illustrative examples on SVRs applicability.

In [LS17], higher accuracies were obtained, when using Back-Propagation Neural Network (BPNN) to predict restaurant sales. The set of variables used was similar to the ones used in [BS09] and [BS11]. MAPE was used as the accuracy estimator.

2.3 Conclusions

To sum up, since it is expected that relationships between predictor variables and the predicted one - daily food demand for a given dish type - will be mostly nonlinear, advanced data mining techniques are the most appropriate when one wants to reach the highest accuracy possible, so that food waste is reduced as much as possible. The whole dynamics and randomness involved in this context implies the use of those techniques, despite some of them being hard to interpret, contrariwise to simple mathematical methods, for instance, time series models.

Once there is no clue about which data mining technique (ANNs, SVRs, RFs) would perform better in the context of the case study explored in this dissertation, all of them were applied in this study, being this an innovative aspect when it comes to predict food demand. The use of a considerable number of variables, most of them never explored before in the literature, also constitutes a contribution of this dissertation.

Chapter 3

Methodology

As mentioned before, this study proposes a system to predict daily food demand for a given dish type in FEUP's canteen. This model is intended to support canteen's managers in the definition of the quantities of the ingredients to buy as well as the quantities to prepare of each dish. This prediction system was designed under the characteristics of FEUP's canteen. Thereby, some details regarding their working service and timetable should be kept in mind. They are:

- It works as a lunch food service facility from 11:30 am to 2 pm, therefore, people come by only at the lunch time;
- It operates only on working days, thereby it is closed by the weekends;
- The customer is given a set of three type of dishes - meat, fish and vegetarian -, from which he can choose one or more, if bought separately.
- Usually, the weekly menu and its combinations offered by the canteen are repeated every eight weeks.

  Please note that in this study we do not distinguish the demand for a menu from the purchase of that menu. This is due to the fact that we do not have information on how the client behaves when the dish he/she wanted is not available. Thus, the quantity purchased is used as a proxy for the demand of the menu. Similarly, the term sales refers to the number of dishes sold/consumed.

This chapter describes the techniques used to predict daily food demand for a given dish type. Thereby, the way data used by such methods was collected and processed, is also described.

3.1 General description

This refers to a regression problem - predicting a continuous variable, that is, in the given context, the number of dishes of food to prepare for a specific dish type on a specific day. Three regression models are proposed, one for meat dishes, one for fish and another for vegetarian dishes.

Methods such as ANNs, SVRs and RFs are capable of building predictive models that identify nonlinear patterns in data - arising from causal-effect relationships between the predictor variables and the variable being predicted. Hence, such methods are very useful since they have the ability to learn that relationships from analyzing and processing data. It turns out that these methods clearly do not require the analyst to guess implicit laws and rules governing the systems from which data are obtained, which is not reasonable given today's real-world problems. Furthermore, being finished the learning process, such methods gain the capability of generalizing, that is, to infer, by itself, previously unseen data, hence predicting future behavior. That said, it was expected that their performance would have significant impact in balancing the demand for each dish type and the quantity produced, thus reducing food waste.

In order to achieve the goals proposed, there will be identified the three best regression models, one for each type of dish.

3.1.1 Data

Regressions methods distinguish two types of variables. They can be either, technically speaking, independent - predictor variables - or dependent - predicted variables. As previously mentioned in this document, predictor variables affect the predicted variable's value.

In the given context the variable being predicted is related to the daily food demand - intended as the amount of dishes served - for one of the three dish types. Therefore, in order to train those predictive models built by such methods, it was needed to collect historical demand data from each dish type. Such data was collected through the collaboration of the "Serviços de Ação Social da Universidade do Porto" (SASUP) institution, which have registered and stored daily food demand data from the last two years, 2016 and 2017, for each dish type. Such data records, for instance, how many times a given dish was served on that specific day - predicted variable (menu demand). Although, since the data collected relates to only two years, some independent variables, with much potential impact on food demand, were put aside. For instance, if there were more data available, the sales mean for each menu of a given dish type and the amount of sales verified for each dish type in the homologous day of the previous year, could have been selected as predictor variables.

Despite that limitations, relevant predictor variables that could impact daily food demand, were identified. Some of these variables are exclusive and strictly related to the whole environment surrounding the FEUP's canteen, while the remaining are inspired by the ones used in the literature, such as those referred in [ANJ16].

Data on meteorological conditions were collected through an Application Programming Interface (API), namely World Weather Online (WWO) [Wor]. This API provides hourly weather

historical data for worldwide locations as of July 2008. Such data is collected from a range of sources, such as World Meteorological Organization (WMO), Geostationary satellites, National Centers for Environmental Prediction (NCEP) and so on. Since some of the data, e.g. precipitation, is given on a hourly basis, it was decided to collect midday (12h) data, once it is the hour in between the canteen's lunch service period. Meteorological variables were collected since there were expected situations such as a sunny day having a negative impact on food demand for that specific day, given that, under such conditions, some people may prefer to have lunch on a terrace instead of indoors.

Also variables characterizing a given day were taken into account, such as its respective month, day of the week. The number of undergraduate, master, integrated masters and PhD students having classes during the different stages of the day (morning, afternoon or even both) were also considered. The latter variables were expected to impact food demand since students staying the whole day nearby the faculty are more likely to have lunch at the canteen, as opposing to the ones having classes only by the afternoon, expected to have lunch in a place other than the canteen. Variables related to holidays, special day's events or semester's special seasons were also recorded. Whenever a special day or season happens, the majority of people tend not to go to the faculty - in this case, to FEUP - and consequently do not attend the canteen, as stated in the chart 3.1. The data collected seems to show that there are preferences for some specific menus offered by the canteen. For example, although generally the meat dish type is the most selected, certain fish or vegetarian menus offered by the canteen have a huge preference from the canteen's customers. Thereby, menus descriptions from each dish type offered by the canteen across the two years were also registered as predictors.

Besides these variables mentioned above, the following were also considered:

- *Praxe*, this variable reflect if the day under consideration is a day in which "praxe" activities are running and thereby if students are going to have lunch at the canteen all together. Students that are involved in "praxe" activities are usually together after 1st year classes, one day per week, and, normally, if they have not have lunch yet, they have it all together in the canteen. It was then concluded that the days in which these groups gather could impact food demand in the canteen.

- *queimaWednesday*. This variable indicates if the day under consideration refers to the Wednesday of the "queima das fitas'" week. "Queima das Fitas" is a week, held every year, in which there are no classes and students celebrate their studies' final year. Tuesday of this week is the day with more people celebrating both afternoon and evening, hence the greater the likelihood that on Wednesday the students will wake up after the canteen closes. Thus, it is expected that on Wednesday there will be less affluence to the canteen.

- *If it was holiday one, two or three days ago*. These variables enable to capture the proximity of a given day to a past holiday. If for a certain day, for example a Friday, it is recorded that

there was a holiday the previous day, it is to be expected that there will be fewer people attending the canteen, since having no classes on Friday, the higher the possibility of students having gone to their hometown earlier, for instance, Wednesday.

- *If it is holiday in one day, two or three days*. These variables enable to capture the proximity of a certain day to a future holiday. Suppose we're on a Friday. If there is a holiday on Monday, that is, in three days, there may be people not having classes that same Friday, thus leaving before Friday for a mini vacation, for instance. Given these circumstances, one can expect a reduction in the amount of consumptions on that day.

- *ordinaryExamsSeason*, *supplementaryExamsSeason*. These variables specify whether a specific day is part of the ordinary exams season, supplementary exams season or not. Exams season is divided in two seasons, the ordinary season in which all the students are enrolled, and the supplementary season, for which only some students apply. In the first semester only students other than first year are enrolled, thus, there are fewer students standing by the faculty and, consequently, fewer students also going to the canteen. During these periods of the academic year, students do not have classes and as such, can study for exams in places other than the faculty, which can lead to not having lunch in the canteen. Supplementary season is a season for students who want to improve the exam grade they obtained at the ordinary season or who, on the other hand, were not successful in that season. With this, it is expected that the number of students with the semester to be completed is lower than in the ordinary season, and as such, fewer students are near the faculty, automatically reducing the daily consumption in the canteen during that season.

- *enrollmentWeek*, *1stYear1stWeek*, *1stYear2ndWeek*. These variables enable to identify the days that are part of the first weeks of the academic year. These are termed as the first three weeks of classes. In these weeks the students of the 1st year are invited to participate in extra-curricular activities, which means that these same students spend more time in the surroundings of FEUP. It is also in these weeks that there is a greater attendance in classes by the remaining students. Given all this, one could expect a greater affluence to the canteen at this time.

Data regarding "praxe" (see table 3.1) was collected just with respect to the five biggest groups, that is, the ones from Chemical, Civil, Informatics, Mechanical and Electrical engineering. These groups are the most active of the faculty and as such the ones that have more potential influence in the consumptions. Some dataset observations may have more than one value assigned to the "praxe" variable, since some of the groups gather the same day of the week. Such issue is bridged through the conversion from categorical to binary variables. Holidays' related variables refer to days where the canteen was not working.

For further details about the variables, the way and source from which they were collected, whether they are categorical (table 3.3), binary (table 3.4), continuous (table 3.2) or discrete (table 3.1), independent or dependent, please refer to the tables below.

Table 3.1: Summary of the discrete variables present in the dataset.

Variable	**Description**	**Mean (standard deviation)**	**Source**
day	Day number varying from 1 to 31.	15.72 (8.5)	Calendar
month	Month number varying from 1 to 12.	6.25 (3.53)	Calendar
maxTempC	Maximum temperature in celsius degrees.	20.12 (5,35)	WWO API
weekDay	Day of the week number, varying from 1 to 5 (from Monday to Friday).	2.95 (1.4)	Calendar
L_Manha, MI_Manha, M_Manha	Number of undergraduate students having classes in the morning, afternoon and all day long, respectively.	15.99 (19.08), 1012.93 (872), 5.97(9.87)	Sigarra FEUP
L_Tarde, MI_Tarde, M_Tarde	Number of integrated masters students having classes in the morning, afternoon and all day long, respectively.	19.65 (27.72), 598.27 (529.3), 17.48 (26.72)	Sigarra FEUP
L_Todo_dia, MI_Todo_dia, M_Todo_dia	Number of master's students having classes in the morning, afternoon and all day long, respectively.	9.92 (14.42), 773.76 (701.79), 4.96 (7.91)	Sigarra FEUP
D_Tarde	Number of PhD students having classes by the afternoon.	5.86 (7.98)	Sigarra FEUP
meat_sales, fish_sales, vegetarian_sales	Amount of meat, fish and vegetarian dishes sold, respectively.	395.96 (166.32), 90.74 (52.60), 43.99 (31.73)	SASUP

Table 3.2: Summary of the continuous variables present in the dataset.

Variable	**Description**	**Mean (standard deviation)**	**Source**
precipMM	Precipitation in milimeters.	0.44 (1.43)	WWO API

The dependent variables (marked as *bold*) were naturally discrete but were treated as continuous in order to treat the problem as a regression problem. This did not seem too out of place considering that customers require different amounts of food.

Table 3.3: Summary of the categorical variables present in the dataset.

Variable	Description	Most important categories' frequency	Source
praxe	Praxe refers to an organized group of students for extra-curricular activities, existing in almost all of the engineering fields from FEUP.	"civil": 8%; "química": 41%; "mecânica": 18%; "eletro": 26%; "informática": 29%	Each group's coordinator.
weekOfClasses	Week number along the semester, varying from 0 to 22.	"0thWeek": 15%; "1stWeek": 4,6%; "4thWeek": 4,3%; "3rdWeek": 4,1%; "11thWeek": 3,9%; etc.	Sigarra FEUP
weatherDescription	General weather description.	"Heavy_rain": 1,4%; "Light_rain": 17,2%; "Mild": 19,7%; "Moderate_rain": 4,8%; "Sunny": 56,9%	WWO API
daily_meat_menu	Menu's description for meat dish.	"strogonoff de porco com arroz branco": 2,8%; "hambúrguer de novilho c/ molho de tomate e massa": 2,8%; "perna de peru em vinho tinto c/ cogumelos e arroz": 2,6%; etc.	SASUP
daily_fish_menu	Menu's description for fish dish.	"sardinha assada com batata a murro e gaspacho": 3%; "solha grelhada com salada russa": 2,8%; "arroz de marisco": 2,6%; etc.	SASUP
daily_vegetarian_menu	Menu's description for vegetarian dish.	"pizza de legumes com ovo": 2,5%; "empadão de legumes e seitan": 2,5%; "jardineira de tofu": 2,5%; etc.	SASUP
daily_meat_category	Menu's category for meat dish.	"frango": 5,3%; "peru guisado": 5,3%; "hambúrguer": 3,9%; etc.	SASUP
daily_fish_category	Menu's category for fish dish.	"filetes pescada": 6,9%; "pescada assada": 6,5%; "pescada cozida": 5,5%; etc.	SASUP
daily_vegetarian_category	Menu's category for vegetarian dish.	"bolonhesa": 8,1%; "lasanha": 7,9%; "empadão": 6,7%; etc.	SASUP

Table 3.4: Summary of the binary variables present in the dataset.

Variables	**Description**	**Most important categories' frequency**	**Source**
First year initial weeks	Enrollment, first and second week of first year's students, respectively.	"enrollmentWeek": 2,3%; "1stYear1stWeek": 2,3%; "1stYear2ndWeek": 2,3%	Sigarra FEUP
Special days	Valentine's Day, super mega feup caffe and wednesday from "Queima das fitas"week, respectively.	"valentinesDay": 0,2%; "superMegaFeupCaffe": 0,5%; "queimaWednesday": 0,5%	Calendar/ Sigarra FEUP
Holidays weeks	Carnival, Easter and Christmas holidays, respectively.	"carnivalHolidays": 0,5%; "easterHolidaysWeek": 2,3%; "christmasHolidaysWeek": 2,6%	Calendar
Special weeks	Special weeks celebrating "Queima das fitas" and engineering, respectively.	"queimadasfitasWeek": 2,1%; "engineeringWeek": 1,9%	Sigarra FEUP
Exams seasons	Ordinary and supplementary exams' season, respectively.	"ordinaryExamsSeason": 14,6%; "supplementaryExamsSeason": 9,3%	Sigarra FEUP
First year exams seasons	Supplementary evaluation's period of preparation and season, respectively, for 1st year students.	"1stYearSupplementarySeasonPreparation": 2,1%; "1stYearSupplementarySeason": 4,4%	Sigarra FEUP
Holiday In	If it is holiday in a day, in two days or in three days, respectively.	"isHolidayIn_a_Day": 3,7%; "isHoliday_in_Two_Days": 3%; "isHoliday_in_Three_Days": 3,7%	Calendar
Holiday Ago	If it was holiday a day ago, two days ago or three days ago, respectively.	"wasHolidayOneDayAgo": 2,8%; "wasHolidayTwoDaysAgo": 2,1%; "wasHolidayThreeDaysAgo": 2,8%	Calendar

The variables mentioned above were considered as potentially impacting the food demand. However, in order to assess the true impact of each one on the final performance of each model, its importance was calculated as explained later in section 3.4 and illustrated in 4.2.

For each of the three models proposed three distinct datasets were used. Besides the dependent variable - number of meals consumed - having its specific values, also the number of records varies, since there was a period - July 31 to August 14, 2017 - where meat and fish dishes were served alternately - *one day yes, one day no*, whereas the vegetarian dish was served every single day along the two years. Also, and exceptionally, on May 25, 2018 there were served two menus of meat. Putting this all together, the meat, fish and vegetarian datasets comprise 428, 427 and 433 total observations, respectively. To have an insight about how a dataset observation looks like, see annex A.1.

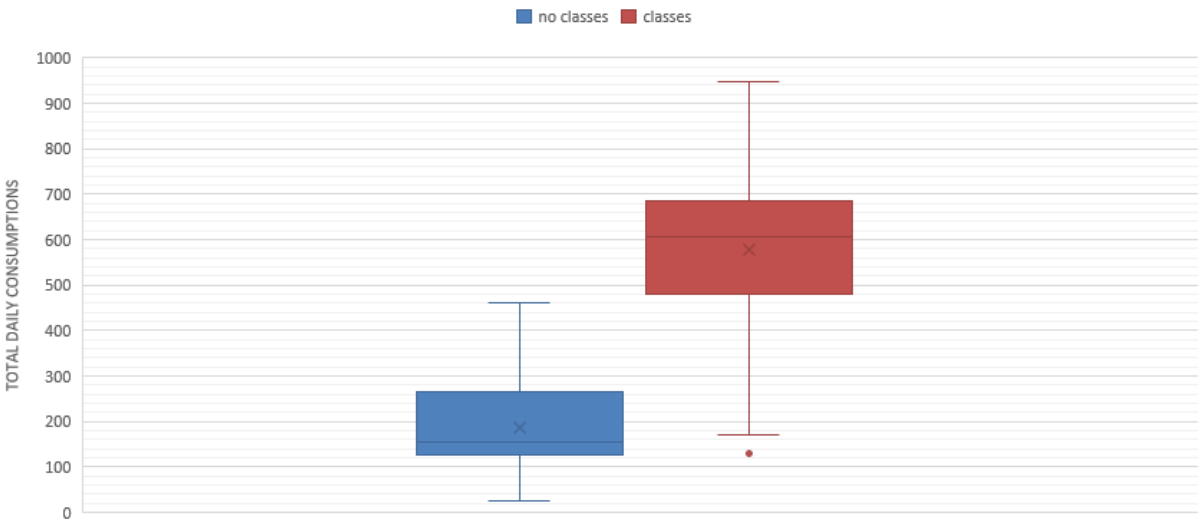

Figure 3.1: Average consumptions on days with and without classes.

The original dataset, as stated previously in the present section, includes all the menus descriptions, offered by the canteen over the years 2016 and 2017, as predictor variables. Although, in hopes of achieving better results, a new dataset was proposed. Instead of being composed by the menus in full, those were grouped into main food categories , both for the meat menus, as well as for the fish and vegetarian ones, so that the amount of predictor variables and datasets' sparsity were reduced.

Categorization of menus

Regarding the categorization of the menus (see list of vegetarian categories in table D.2), this was done in order to group the menus and as such reduce the sparsity of the dataset. These menus were grouped according to not only the similarities between menus descriptions (check the vegetarian menus and their related statistics in table D.1) and ingredients but also according to the menus

with a similar average number of sales. For instance, when it came to categorize pizza menus, always offered as a vegetarian dish, the corresponding menus - *pizza com ovo*, *pizza calzone com ovo* and *pizza de legumes com ovo* - were grouped together in the same category since they have "pizza" in its description and their sales mean - *88,2*, *79,7* and *74,4* respectively - are similar.

3.1.2 Preprocessing

Many of the previously referred variables are categorical variables. Thus, such variables, upon the preprocessing of data, were subjected to a conversion of its several categories into dummy variables - or binary variables. Furthermore, since the data mining techniques used in this study are quite sensitive to outliers [TSK05], these were removed, more specifically the observations related to the period from April 13 to April 15, 2016, during which a special external event, guaranteeing lunch at FEUP's canteen, took place. During this period, pikes in each dish consumptions, upon comparison to the remaining data, were verified.

In order to reduce as much as possible datasets' dimensionality, all the binary constants, for which their values were all equal to 0, have been removed. As such, the shape of each matrix representing a given dataset may vary.

Scaling of numerical variables was also performed. It is a common preprocessing method, but not a functional requirement. Given the wide range of values coming from continuous and discrete variables collected for this study, scaling was applied to such variables when building models with ANNs and SVRs. For instance, ANNs are steepest descent algorithms, therefore, if the input values are too large, the function updating weights, may impair convergence to a local minimum - the one minimizing error -, being also prone to overflow values - numeric value that is outside of the range that can be represented with a given number of bits by the Central Processing Unit (CPU). Therefore, data given as input to ANNs was scaled to a range in accordance to the activation function used in the hidden layer - Rectified Linear Units (ReLu) [ZH] - (for further details on why this function was chosen over the remaining, see next section 3.2). Similar to ANNs, for SVRs it is also appropriate to have its data scaled, since the cost function will tend to behave well and therefore finding the optimal solution more easily. On the other hand, RFs were not subjected to scaling since they are tree partitioning algorithms and as such do not include a coefficient impacting the values of the independent variables. That said, data was scaled when it came to apply those algorithms, namely SVRs and ANNs. Since most of the numerical variables' values were not distributed normally, standard scaling was used. Scaling, when applied to these algorithms, was proven to improve running time when training models, as well as their performance and generalization error reduction.

3.2 Techniques

As previously stated, three main data mining techniques were used in this study. Since the use of these techniques is scarce when it comes to predict food consumption, in this study they are

applied and compared against each other in terms of performance, as a way of contributing to the existing forecasting literature.

3.2.1 Random Forest

Random Forest is a class of ensemble methods - intend to improve final prediction by aggregating predictions of multiple classifiers [TSK05] -, designed particularly for decision tree classifiers, but also used with decision tree regressors, as described in [Bre01]. Each tree has its own random vector associated, independent from all the vectors related to each of the remaining trees, but with the same distribution. This random vector injects randomness into the model-building process through many approaches. The one used in this study is by selecting, at random, F predictor variables to split at each node of the decision tree, which may help reducing its bias [TSK05]. Each vector is composed by its randomly chosen variables. The tree is grown whenever a node is splitted. This split is performed through the variable giving the highest quality split, that is, the variable which reduces the most the value of the metric chosen, for instance, Mean Squared Error (MSE) or MAE (see section 3.3.3). Once all the trees are grown, the final prediction is obtained by averaging the values calculated at the final node of each of them (see Figure 3.2). Please note that the number of trees used to constitute the forest is a hyper-parameter predefined by the user (see section 3.3.2).

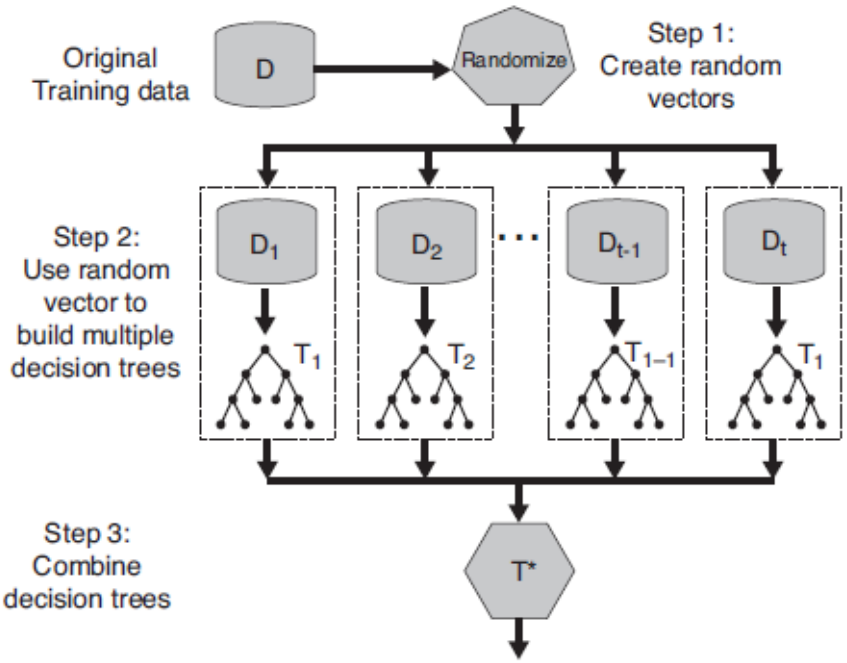

Figure 3.2: Example of a Random Forest [TSK05].

3.2.2 Support Vector Regression

SVR is an analogous technique for regression, as SVM is for classification. They are based in statistical learning theory, since they bear on finding a predictive function based on data. In SVR it is also stated the maximal margin hyperplane concept, which defines the extent to which a training example falls out without being *misclassified*. As implicit in Figure 3.3 this margin is defined by a subset of training examples - the so-called *support vectors* [TSK05]. Thus, the wider the margin the lower the willingness to generalization error. Therefore, SVR formulation, similarly to SVM, aims to minimize an upper bound for the generalization error, instead of minimizing the prediction error on the training set [CW07]. This generalization error is increased if the training errors are bigger than an hyper-parameter - predefined by the user - known as the margin of tolerance. When the data is nonlinear, SVR performs a nonlinear transformation - *kernel trick* -, in order to map the data from its original feature space into a new space where the decision boundary becomes linear [TSK05].

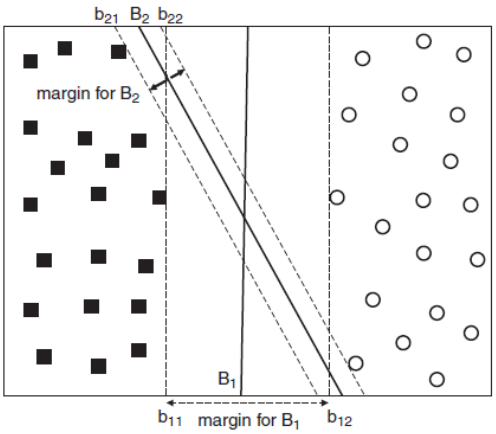

Figure 3.3: Margin of a decision boundary [TSK05].

3.2.3 Artificial Neural Network

ANNs are inspired by the human brain. That is, neurologists have found that the human brain learns, by updating the connection strength between neurons, when stimulated repeatedly by the same impulse [TSK05]. Generally, an ANN is composed by a set of nodes - neurons - and layers, being connected by direct and weighted links - axons. Therefore, ANNs intend to simulate the brain functioning, updating links' weights accordingly to the input given from historical data, iteration after iteration, until they have learn the relationships between predictor and predictive variables [ANJ16]. ANNs only find a local minimum in the function of minimizing sum of squared errors, since the final link weights chosen will be the ones for which the function value is minimum. To find the minimum, the function's gradient is calculated. This gradient determines weights' values - between input and hidden and between hidden and output layers -, which are

being constantly updated, as well as the errors in the output layer are being back-propagated to the hidden layer - *back-propagation*. Such process is repeated until the final values are obtained, those for which the function minimizes the error the most.

ANNs are data-driven self-adaptive methods, that is, they learn from training examples - data samples - and capture subtle functional relationships among those, even if the underlying relationships are unknown or hard to describe [GEM98]. There are many possible network structures in the family of ANNs, from the simplest one called *perceptron* to the most popular and widely-used when dealing with forecasting problems, the *multilayer feed-forward ANN* (see Figure 3.4), also known as *Multi layer Perceptron* (MLP). The latter is the one used in the present study.

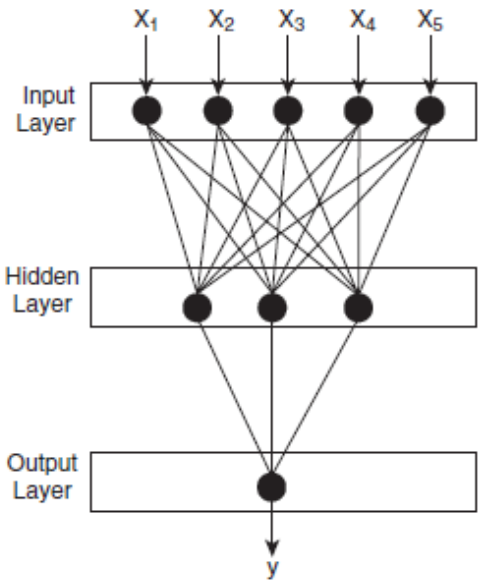

Figure 3.4: Example of a multilayer feed-forward ANN [TSK05]. Each node in the input layer corresponds to a predictor variable, $X_1 ... X_5$, whereas the one node in the output layer corresponds to the predicted variable.

3.3 Performance evaluation metrics and methods

Upon training a model with data, one must test its final performance in another partition of data completely exclusive and apart from the one used for training [TSK05]. Given the context of the present study, such data were randomly splitted - a *seed* was set for reproducibility purposes across all the experiments - in two equal training/validation and training/test sets. Furthermore, a model with the ideal complexity is one that produces the lowest generalization error - expected error of the model when predicting previously unseen records ([TSK05]). Thereby, the next sections describe techniques used to find out such complexity.

3.3.1 Cross-validation

Cross-validation is a commonly used technique in nowadays' machine learning models. It is used to validate the stability of such models, that is, it provides a robust way of testing whether a model generalizes well or not on previously unseen data. There are many variants on this technique. The one used in the present study is *K-fold cross-validation*, with k set to 10. This method divides the dataset into k-subsets, where $k-1$ are used to form a training set and the remaining one is used as test set. Afterwards, such process is repeated k times, such that each time one of the k subsets are used as a test set. Therefore, the final error estimation is made by averaging all the errors coming from each of the k rounds [TSK05]. Cross-validation is a technique that neutralizes the risk of over-fitting or under-fitting, since most of the data is used for fitting the model, which in turn leads to low-biased predictions. In this study, it was used on the training/validation and training/test sets, separately. The final model's performance was obtained through the results coming from running 10-fold cross-validation on the training/test set.

3.3.2 Hyper-parameters tuning

Hyper-parameters are parameters optionally set by the user before training the model with a given algorithm. However, the appropriate choice of such parameters - also known as *grid search* -, regardless of the data, leads to model's final performance improvement and as such to its vulnerability to over-fitting and under-fitting reducing [CW07]. Therefore, such tuning was taken into account upon training models through the data mining techniques used. Tuning was made while the model was being trained on the training set (see section 3.3). In order to make models more robust and "over-and-under-fitting-proof", for each set of hyper-parameters values given as input (using a grid search approach), it was applied a 10-fold cross-validation (see subsection 3.3.1 for further details on this technique) to infer which set of parameters gives the lowest model's accuracy on the training/validation set. Afterwards, and this time using the best hyper-parameters, cross-validation was used once again to obtain model's performance on the training/test set. The closer the both performances - using training/validation and training/test set - the greater the certainty that the model is not prone to either under-fitting or over-fitting.

Since there were used three data mining techniques, different hyper-parameters for each technique were considered. Thereby, below is given a summary about which parameters - and its meaning - were chosen to be optimized for each technique:

- **Random Forests**: According to [Bre01], in order to grow each of the trees - number of estimators - a random selection of a subset from the existing variables - or features - should be made. This subset is the one considered when splitting decision trees' nodes. Thereby, the size of such subset was tuned. It was also tried an approach on tuning the number of decision trees composing the forest. The larger the amount the better will perform the algorithm, but also the longer it will take to compute. Thus, it was concluded, upon trading off runtime and performance, that its default value - 1000 - was enough.

- **Support vector regression**: Many SVR hyper-parameters should be set properly so that reliable performances are achieved [CW07]. The ones tuned in this study were:

  - *Kernel function*: This function is used to construct the non-linear decision boundary on the SVR input space. The one chosen in this study was the Gaussian function, which according to [SS04] is the one performing better.

  - *Regularization parameter, C*: This parameter defines the penalty value upon "misclassifying" a training sample [TSK05].

  - *Gamma*, γ: This is a free-parameter in the Gaussian function, meaning the influence of a given support vector in another's classification, regardless of their distance. Thereby, large values of gamma imply low variance but an highly-biased model, and vice-versa.

- **Artificial Neural Network**: ANNs are known for being very difficult to train, since its structure is very problem-dependent [ZH]. MLP is the network structure used in this study, which is basically composed by many nodes distributed over an input layer, one or more hidden layers and an output layer. There are several methods in the literature when it comes to tune such hyper-parameters. Although, none of them can guarantee an optimal network configuration. The most reliable and easy to implement technique is indeed grid search. Therefore, the following parameters were tuned and analyzed:

  - *Number of nodes in the hidden layer*: This is a critical parameter in the final model performance because it is the hidden nodes that captures pattern in data and performs non-linear mapping between independent and dependent variables. This layer requires an *activation function* in order to introduce nonlinearity between variables [ZH]. Thereby, before tuning the hyper-parameters, such function was defined as being the standard one. Hence, *ReLu* was chosen given its ability converge and train faster than the *sigmoid* or *tanh* function, and to rectify and avoid vanishing gradient problem. This is a recurrent problem when training ANNs, which occurs mainly when values calculated in previous - hidden - layers have a really tiny or even none effect on the output.

  - *Learning rate*: A crucial hyper-parameter for back-propagation algorithms, such as MLP, since it determines to which extent weights' updates impact values calculated in the hidden layer. For highly complex data, as it is in this case, low values should be chosen, despite its restricted and allowable range being [0,1] [ZH].

The next table gives an insight on the hyper-parameters' range of values tested upon grid search.

Table 3.5: Hyper-parameters - range of values tested with grid search.

Data mining technique	**Hyperparameters**
Random Forests	*max_features*: [1, number_of_features_in_dataset, **step** = 1]
Support Vector Regression	C: [2^0, 2^{20}, **step** = 2] γ: [2^{-10}, 2^8, **step** = 2]
Artificial Neural Networks	*number_of_nodes_hidden_layer*: [20, (number_of_features_in_dataset)*2 + 1, **step** = 4] *learning_rate*: [10^{-3}, 10^0, **partitions** = 10]

3.3.3 Metrics

When it comes to forecast a continuous variable - regression problem - there are two main metrics, MAE and MSE, used in order to measure model's forecasting error, and, consequently, its performance [RS03]. MAE measures the mean absolute difference between observed and predicted values (see equation 3.1). Such metric was the one used upon applying the techniques described in previous sections, such as cross-validation (see section 3.3.1) or feature selection (see section 3.4). R^2 was also used to the same extent as MAE did. Such metric measures how correlated are both observed and predicted values (see equation 3.2), that is, how much variance arising from the observed values can be "captured" by the predicted values. This metric was used to evaluate how well models behave, comparing the regressed and observed lines.

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j| \tag{3.1}$$

where:

n = number of samples

y_j = observed value

$\hat{y}_j$ = predicted value

$$R^2 = 1 - \frac{\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}{\sum_{j=1}^{n}(y_j - \bar{y})^2} \tag{3.2}$$

where:

n = number of samples

y_j = observed value

$\hat{y}_j$ = predicted value

$\bar{y}$ = mean of the observed data, given by:

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_i \tag{3.3}$$

3.4 Feature selection

Since there were collected many variables to handle this problem, the dataset may be seen as sparse matrix and may contain a lot of irrelevant information since most of its entries are equal to 0. Consequently, feature selection was run to handle this potential problem. This technique has three main advantages. It makes it easier to interpret the model, variance and proneness to overfitting are reduced and last but not least, time required to train the model decreases significantly. Basically, this technique is used to select the most important features and as such removing the least important, those impacting in a negative way the model's performance - in this case, those which increase its MAE. The technique applied in this study, is also known as *permutation accuracy importance* [SBZH07]. It was used with each of the data mining techniques used in this study, and it is implemented as follows:

1. After performing hyper-parameters' tuning, as described in section 3.3.2, an object of the data mining technique being used should be instantiated with its best set of hyper-parameters.

2. Split, randomly - this time without a *seed* -, the dataset into two equal parts, one for training, the other for testing.

3. Get predictions on training set through 10-fold cross-validation, as per described in 3.3.1.

4. Calculate difference between y values from training set and predicted ones - termed *original_mae* from here on.

5. Iterate over the dataset variables and shuffle the values from the variable corresponding to the i^{th} iteration - where i is an index on the variable's list according to its disposal in the dataset. This shuffling process is also known as the *permutation of values*.

6. Once the permutation is completed, the MAE on the modified dataset is calculated, analogously to what is done in steps 3 and 4. Such MAE is termed *shuffled_mae* from now on.

7. Permutation importance is now calculated through $\frac{shuffled\_mae - original\_mae}{shuffled\_mae}$ and saved to a dictionary - data structure mapping a key to set of values -, storing for each variable a list of its permutation importances.

8. Repeat this whole process, from steps 2 to step 7, a number of times proportional to half the size of the dataset. Each iteration has a different random split of the data, as stated in 2.

9. Get the permutation's importances mean for each variable. If the mean is lower than 0, the corresponding variable must be removed from the dataset, otherwise it is stated as an important and impactive variable and therefore should not be removed.

This whole process was run against the original datasets - the one with menus descriptions as variables and the one with menus categories as variables (see section 3.1.1). After running it, hyper-parameters tuning, as described in 3.3.2, was run once again, this time against the dataset without the least important variables.

RFs already include feature selection. The variable used to split a node is the one giving the highest decrease in MAE or MSE, for regression problems. So, by pruning the tree below a particular node one can have an insight about which features are the most important, since those are the first to be used when growing the tree. Therefore, it was expected that this technique would not have much impact in improving model's performance once it was built with RFs. Although, it was verified a great improvement when it came to build models both with ANNs and SVRs, as stated in section 4.2.

3.5 Software

Datasets were kept in excel files, since they store each row as a data record. Then, through using Python, those datasets were easily retrieved to software code with the help of many existing libraries, especially designed with data analytics and machine learning kept in mind. The main packages used were *xlrd*, *pandas*, *scikit-learn* and *numpy*.

In addition, Python is a language inter-operating way better than other much used options - R programming language -, since, in case of one wondering to integrate the predicting system into others institutions' systems - mostly designed in languages such as Python, C#, C++, etc -, latency or maintenance issues will be minimal. Moreover, the author of this dissertation is used to this language - one from which syntax is intuitive and easily readable -, which is yet another strong reason to use it.

Chapter 4

Results

This chapter focuses on illustrating the datasets and its main variables as well as the results achieved upon building models with the different data mining techniques described in section 3.2. Experiments were done with two different datasets, one with the menus descriptions as variables and the other with their main categories as variables, both with and without feature selection. The following sections give a detailed comparison between such scenarios.

4.1 Descriptive analysis

In order to detect outliers, seasonality and other patterns, charts regarding the daily demand for each dish type - meat, fish or vegetarian - over 2016 and 2017, were generated. By looking at these charts (see charts 4.1 and 4.2) one can observe seasonality. For instance, the levels of consumption per month seem to be quite similar. More specifically, for example in October, the level of demand is always very high. This is due to the fact that it is the beginning of the academic year, where there is a greater affluence to the faculty by the students, and as such to the canteen. In order to try capturing such variance, it was mapped a variable indicating to which week of classes a given observation concerns (please refer to the table 3.1).

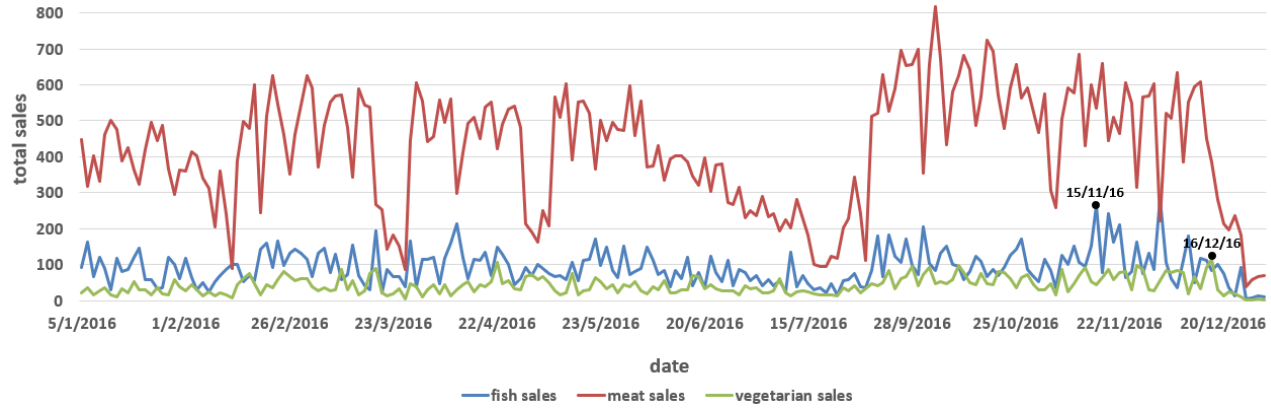

Figure 4.1: Meat, fish and vegetarian sales observed over 2016.

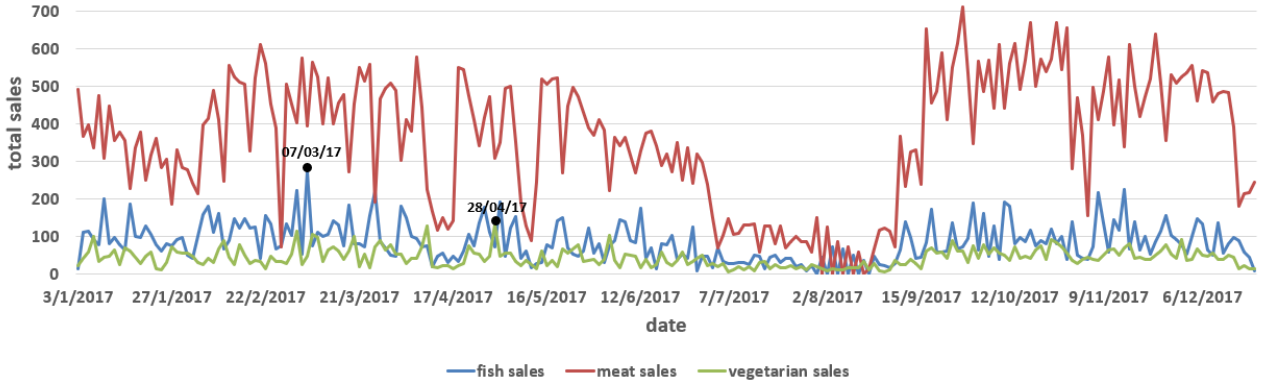

Figure 4.2: Meat, fish and vegetarian sales observed over 2017.

As the charts above demonstrate, data for sales in 2016 and 2017 have a similar pattern and behavior. Also noteworthy is the stark difference between meat sales and both fish and vegetarian sales (see tables 4.1 and 4.2 for their mean, standard deviation and coefficient of variation values). This is due to the fact that most customers, given the three menus in the daily menu list, usually have preference for the meat dish, as referred in section 3.1.1.

As can be seen in the sales charts for 2016 and 2017, 4.1 and 4.2 respectively, the peak sales of fish dishes take place on November 15, 2016 and March 7, 2017 , which correspond to the menu category "pataniscas bacalhau", the fish dish that most influences the total daily consumption (see Figure 4.11). On the other hand, with regard to the peak sales of vegetarian dishes, this occurs on December 16, 2016 and April 28, 2017, both of which correspond to the sale of the "pizza" menu category, which is the vegetarian dish with the greatest influence on total daily consumption (see Figure 4.10).

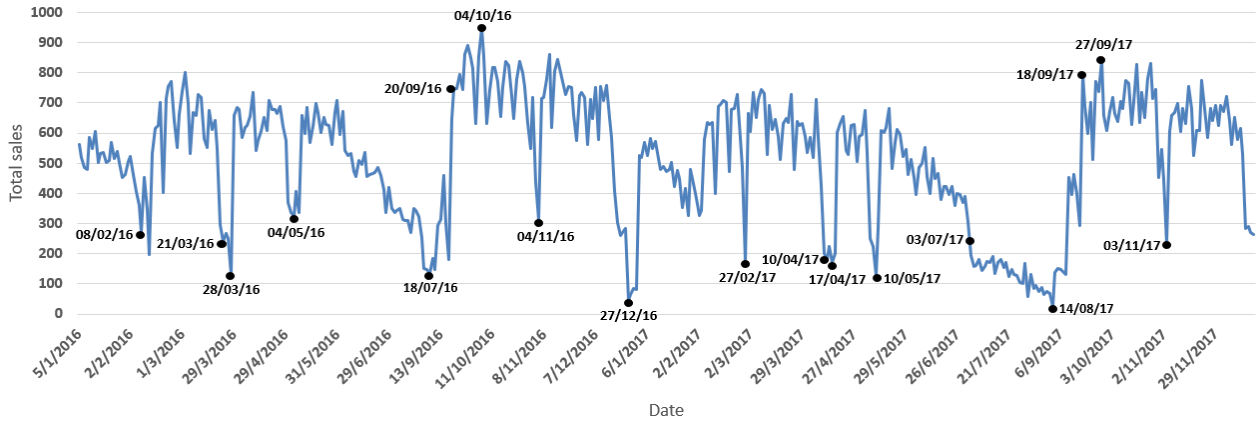

Figure 4.3: Total sales verified over 2016 and 2017 years.

In order to make a more detailed analysis of consumption trends in the canteen, a chart (4.3) was plotted to give an insight about the total amount of consumptions/sales verified in both 2016 and 2017.

Periods from 21 to 28 March 2016 and from 10 to 17 April 2017 correspond to the Easter holidays, that is, days in which there are no classes, hence the reduced number of consumptions observed.

Both October 4, 2016 and September 27, 2017 had the highest sales amount in its respective year. Both correspond to a day of the second week of the first semester classes. Thus, since in the first weeks of classes it is usually observed the highest attendance levels in classes by the students and also great attendance in *praxe* groups, it is expected a great affluence to the canteen in that same period.

Both November 4, 2016 and November 3, 2017 correspond to the last day of the engineering week, a week where there are no classes and in which the engineering area is celebrated. Both correspond to a Friday, preceded by the academic Thursday, where many students go out at night, it is expected that in that days the same students wake up after the canteen closing, hence the abrupt drop in sales compared to the previous days.

Since both 18 July, 2016 and 3 July, 2017 correspond to the Monday after the last day of the second semester, it is expected that few people go to faculty from those days until the beginning of the next school year, hence the significant drop in consumptions thereafter. In the beginning of the next school year, both the 2016/2017 - 20 September 2016 - and the 2017/2018 - 18 September 2017 -, there is a sudden increase in consumptions as can be noticed in chart 4.3.

The 27th of December 2016 and the 14th of August 2017 were the days when the lowest consumptions were recorded. These days corresponded to days of Christmas holidays and summer holidays, respectively. Therefore, daily consumptions are expected to be lower, since the majority of the faculty staff do not work at that time, as well as students do not have classes.

The 4th of May 2016 and the 10th of May 2017 correspond to the Wednesday of the week of *Queima das Fitas*, week in which there are no classes. As mentioned before, the day before the Wednesday, there is a kind of a parade of all the courses from the University of Porto (UP), which

occurs from early afternoon until dawn. As such, it is to be expected that the next day there will be a significant sales decline, since there is a greater number of students waking up after the canteen closes, in addition to having preference for other places to have lunch.

On 27 February 2017 there was a large decrease in the total sales value - about 175 - since it was a day related to the carnival holidays, so as there are no classes on that day, it was expected that fewer people would have lunch in the canteen. On the same day corresponding to the carnival holidays of 2016, 8 February 2016, there was a less abrupt descent - total sales of 270 -, because at that time supplementary exams season was taking place, thus it was expected more people nearby the faculty, and consequently having lunch in the canteen.

Table 4.1: Mean and standard deviation values for 2016 sales.

	meat	**fish**	**vegetarian**
mean	422,34	92,57	42,21
standard deviation	136,25	37,72	19,51
coefficient of variation	32,26%	40,74%	46,22%

Table 4.2: Mean and standard deviation values for 2017 sales.

	meat	**fish**	**vegetarian**
mean	359,62	83,44	42,27
standard deviation	143,22	40,33	18,95
coefficient of variation	39,83%	48,33%	44,83%

4.2 Predictive models

As stated in the beginning of this chapter, there were used two datasets in this study - one with the menus descriptions and the other with their categories - for each type of dish. Such datasets were subjected to a technique that removes irrelevant variables and as such selects the most important ones (see section 3.4 for details on how such technique works). That being said, for each type of dish there are four different datasets. This gives us a total of 12 datasets, which were given as input to the three data mining techniques - RFs, SVRs and ANNs -, in order to construct the predictive models.

A total of 36 models were trained and built, from which three were chosen as the most accurate on predicting meat, fish and vegetarian's dishes demand. Since the main goal of this study is to reduce food waste, we aimed at finding a model which could minimize the differences between the observed and predicted daily food demand. Therefore, the model performing better is the one obtaining the lowest MAE.

After training the models, it was concluded that, whatever the dish type demand to be predicted - meat, fish or vegetarian sales -, the dataset for which a given model obtained the best performance was the one including menus' categories and feature selection. Thereby, in the following tables,

there are summed up the R^2, MAE, positive and negative deviation, and hyper-parameters' values obtained for the three best models predicting meat, fish and vegetarian sales. Such values concern to those obtained upon running such techniques against the test set, with 10-fold cross-validation, as stated in 3.3.1.

Meat

Table 4.3: Results on meat dataset with menus categories, with feature selection.

	Random Forest	**Support Vector Regression**	**Artificial Neural Network**
R^2	0,82	0,79	0,64
MAE	<u>53,87</u>	59,00	82,73
Positive deviation	26,60	30,50	45,21
Negative deviation	-27,27	-28,50	-37,52
Hyper-parameters	*max_features* = 11	C = 8192 γ = 0,00195	*nodes* = 140 *learning_rate* = 0.464

Note: *nodes* refers to the number of nodes in the hidden layer. Underlined values concern to the data mining technique that best predicts meat sales.

Fish

Table 4.4: Results on fish dataset with menus categories, with feature selection.

	Random Forest	Support Vector Regression	Artificial Neural Network
R^2	0,37	0,47	0,44
MAE	30,76	27,83	29,43
Positive deviation	15,84	14,73	20,87
Negative deviation	-14,92	-13,10	-8,56
Hyper-parameters	*max_features* = 75	C = 2048 γ = 0,00195	*nodes* = 120 *learning_rate* = 0.215

Notes: *nodes* refers to the number of nodes in the hidden layer. Underlined values concern to the data mining technique that best predicts fish sales.

Vegetarian

Table 4.5: Results on vegetarian dataset menus categories, with feature selection.

	Random Forest	Support Vector Regression	Artificial Neural Network
R^2	0,55	0,69	0,64
MAE	13,02	10,51	11,06
Positive deviation	6,62	5,46	7,57
Negative deviation	-6,40	-5,05	-3,49
Hyper-parameters	*max_features* = 4	C = 8192 γ = 0,00049	*nodes* = 200 *learning_rate* = 0.100

Note: *nodes* refers to the number of nodes in the hidden layer. Underlined values concern to the data mining technique that best predicts vegetarian sales.

MAE, as mentioned before, refers to the mean of the absolute differences between the observed and predicted values. Such differences' values can be either positive or negative. Thereby, given that differences, **positive** and **negative deviation** points to how much, averagely, the predictive model has predicted a value below or above the observed value, respectively. Assuming that the number of meat, fish and vegetarian dishes to be prepared for a given day will be governed by the model, if it has predicted above, the difference would reflect leftovers, the so-called **waste**. Otherwise, the model has underestimated the demand, meaning that more *backup* food would be needed to produce. For an explanation on the R^2 meaning, please refer to the section 3.3.3.

Tables 4.3, 4.4 and 4.5 show that RFs were the technique performing better when predicting daily meat consumptions, whereas SVRs outperformed all the other techniques upon predicting daily fish and vegetarian consumptions. When comparing the results obtained by each technique in the datasets entailed by these tables and the remaining datasets (see annexes B), it is possible to conclude that the technique of feature selection, when applied in datasets subsequently given as input to RFs, provided few improvements in MAE values. This is due to the fact that the RF

algorithm itself already makes feature selection, that is, given the F variables randomly selected (see section 3.2.1), the algorithm chooses the variable that will give the largest decrease in the MAE when splitting the node. On the other hand, when datasets subjected to feature selection are given as input to both SVRs and ANNs, improvements in MAE values are notorious.

Since the main goals of this dissertation is the reduction of food waste, special attention should be paid to the values obtained for the negative deviation. As noted earlier, this deviation is what food waste translates to. As shown in Table 4.3, the best predictive model for daily meat consumption achieved an average daily waste of 27 meat dishes. Regarding the prediction of daily fish consumption, the best model got an average daily waste of approximately 15 fish dishes (see Table 4.4). Finally, with regard to the daily vegetarian consumption of vegetarian, the best model obtained an average daily waste of only 6 vegetarian dishes (see Table 4.5). To have an insight about the monthly waste over 2017, either induced by the models proposed or by the method of food demand management currently applied by the canteen, check Figure 5.2.

In the further charts, it is possible to compare the observed and predicted values obtained for the above and so-called *best models*. The ten most important variables for building the model are also illustrated. Such importance values were obtained through running the algorithm described in 3.4.

Meat

- **Observed** *versus* **Predicted**

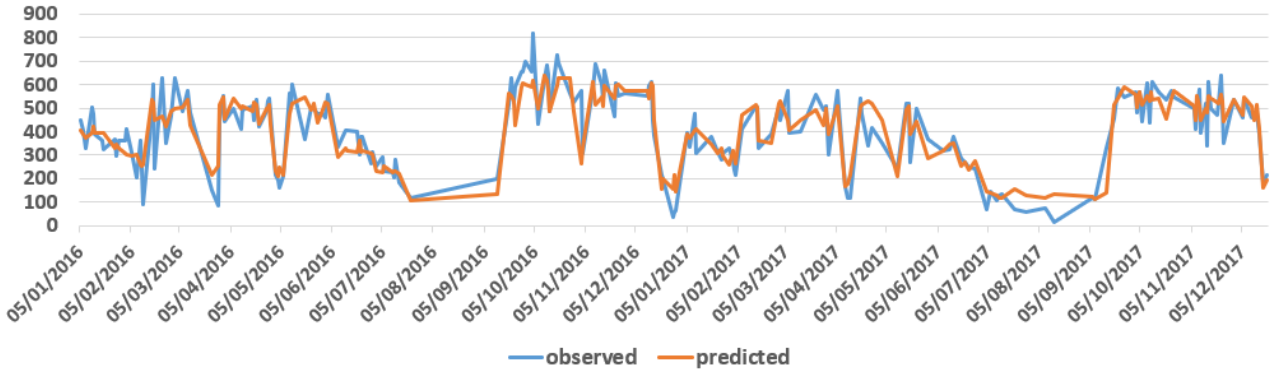

Figure 4.4: Comparison between the observed and the predicted for the model that best predicts daily meat consumptions.

- Ten most important variables

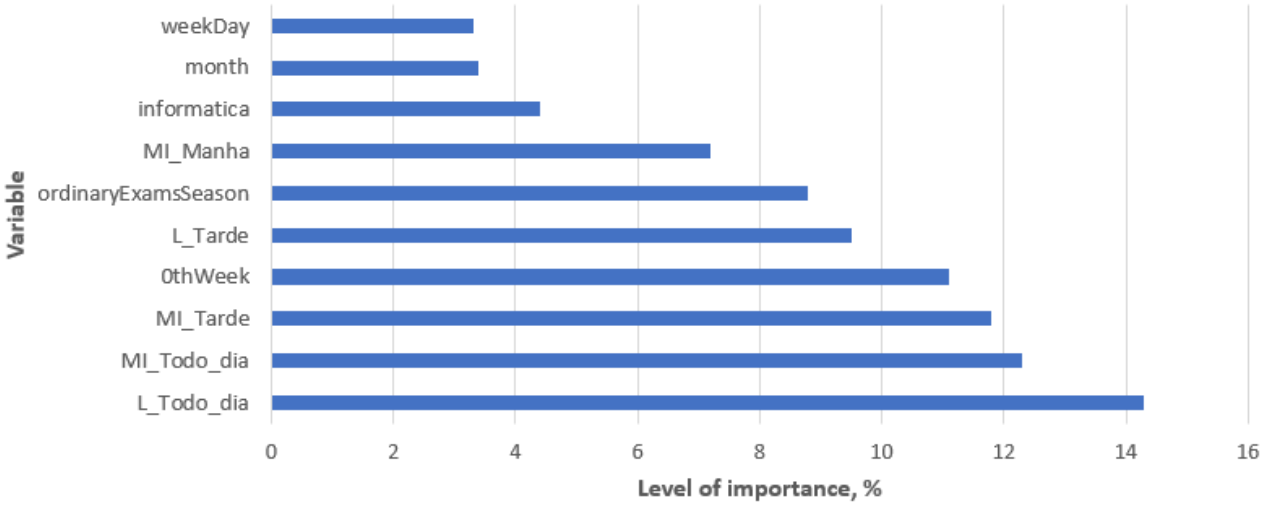

Figure 4.5: Ten most important variables for the model that best predicts daily meat consumptions.

Fish

- **Observed** *versus* **Predicted**

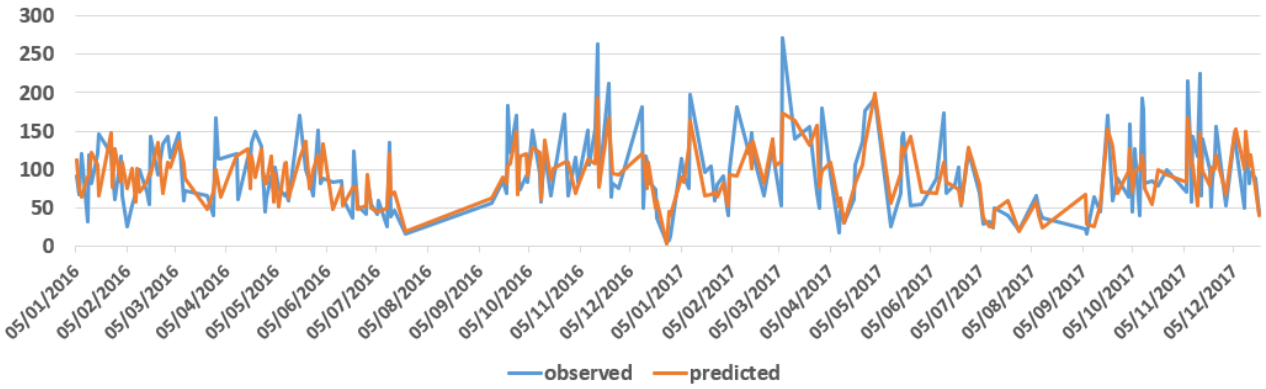

Figure 4.6: Comparison between the observed and the predicted for the model that best predicts daily fish consumptions.

- Ten most important variables

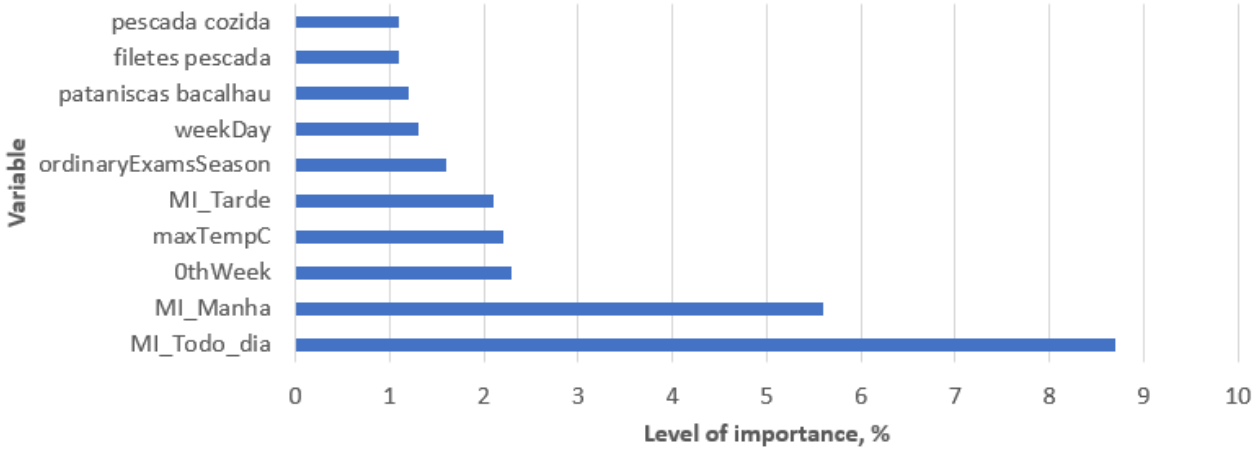

Figure 4.7: Ten most important variables for the model that best predicts daily fish consumptions.

Vegetarian

- **Observed** *versus* **Predicted**

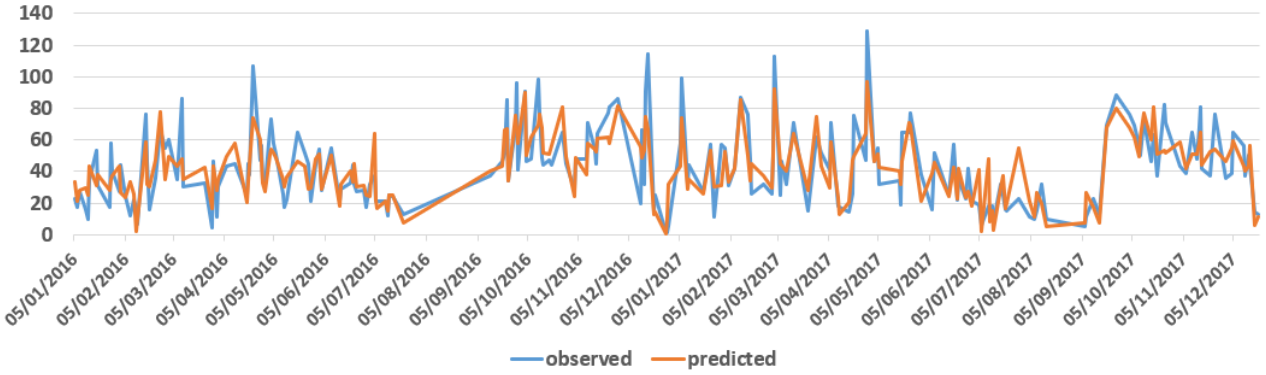

Figure 4.8: Comparison between the observed and the predicted for the model that best predicts daily vegetarian consumptions.

- Ten most important variables

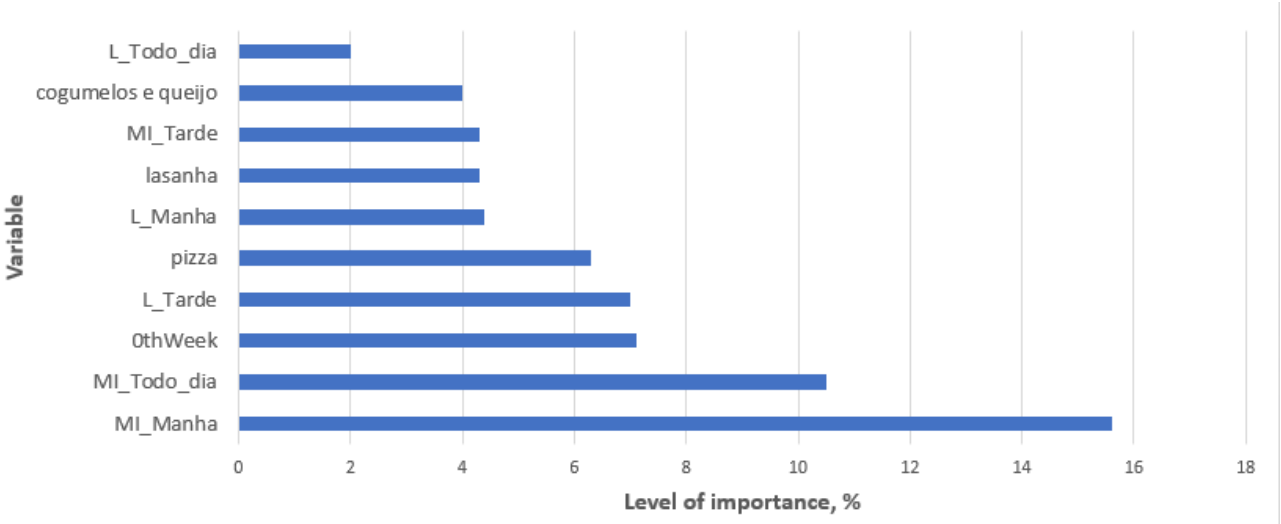

Figure 4.9: Ten most important variables for the model that best predicts daily vegetarian consumptions.

By looking at the charts 4.4, 4.6 and 4.8, one may conclude that all the three models easily capture an abrupt decline or rise in consumptions, since these lags happen every year under the effect of the same variables' values. On the other hand, since the canteen was open in August only in 2017, the models have difficulty in predicting consumptions for that month, given that the existing dataset do not accurately and consistently capture the relationships between the variables in question.

Variables impacting the most were the ones expected to do so upon their collecting. For instance, variables concerning the number of students having classes the whole day. Due to his schedule, it is convenient for the student to stay nearby the faculty by the lunch time, increasing thereby its likelihood to have lunch in the canteen. Therefore, having a look at the charts above (4.5, 4.7 and 4.9), one may observe that variables such as *MI_Todo_dia* and *L_Todo_dia*, have a great impact on the performance of the final models. Also variables such as the ones recording the week of classes in the current semester had a great impact, for instance *0thWeek*, that refers to a day in which there are no classes. Thus, when there are no classes, there are few people in the faculty and as such there is less affluence to the canteen. *Pizza* menu category was also crucial in predicting vegetarian consumption, since when it is present in the daily menu, the chances of people opting for it is much higher than it would be if the vegetarian option were other than *pizza* (see Figure 4.10). The same holds for the *pataniscas bacalhau* menu category, being the fish category, of all, with the most influence in the total daily consumption (see Figure 4.11).

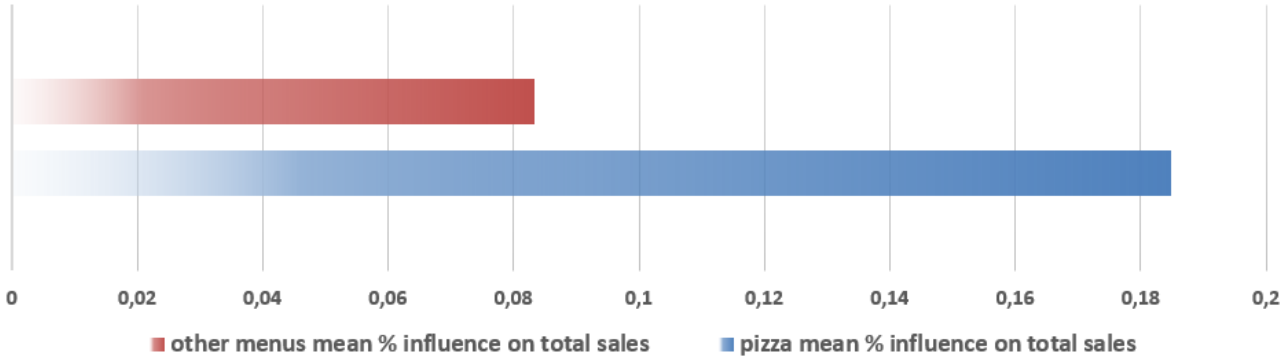

Figure 4.10: Comparison between the average percentage of "pizza" dishes and other menus' sold, in relation to the total.

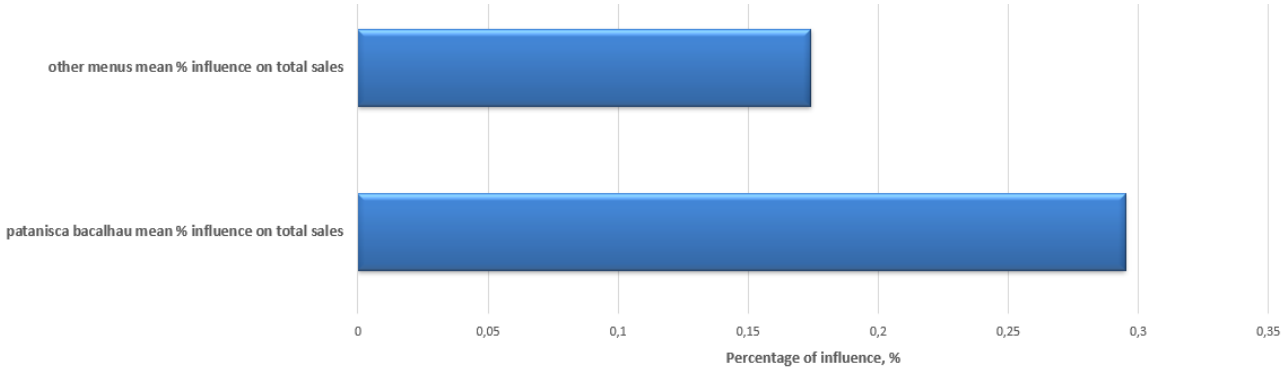

Figure 4.11: Comparison between the average percentage of "pataniscas bacalhau" dishes and other menus' sold, in relation to the total.

Chapter 5

Discussion

5.1 Current canteen food demand management

Upon talks with the main canteen manager, it was known that canteen's food supply for a specific day is based on the sales amount verified in the homologous day of the previous year. Therefore, food waste may arise from such way of managing the food preparation. Although, canteen managers try to mitigate part of the potential deviation observed between the demand and the supply, also termed as the observed and predicted values, respectively. On the one hand reallocating the *non-cooked* food ingredients for the next days, distributing them by the three main lunching services in the canteen building - Snack, Grill and the traditional canteen. When there are leftovers of already cooked food, these, under certain restrictions, are donated to charities, namely institutions operating nearby the surrounding area of FEUP. On the other hand, when the estimates are below the demand, and this is being noticed throughout the operating period of the canteen, other menu (different from those proposed) is prepared to address that unexpected demand. This causes stress and a poor service level, as most of the menus offered are not perceived by the customers as good meals.

5.2 Canteen's management as a benchmark for the predictive models

Since canteen's food supply for a given day is based on the amount of sales verified in the homologous day of the previous year, it is possible to calculate its total monthly waste. Their predicted values are the same as the sales verified in 2016 sales whereas the observed concern to the sales data from 2017. So in order to compare the monthly waste generated between the system developed in this study and the canteen's current prediction system, there were calculated their respective differences between the observed and predicted values for the all months of 2017. As described before in section 4.2, negative differences or deviations translate into food waste, since the predicted value was higher than the observed, hence there were prepared more meals than those necessary. So, for each month and model - meat, fish and vegetarian -, there were summed

all the negative deviations verified in the corresponding days, both for the system developed and for the canteen's system (see chart 5.1).

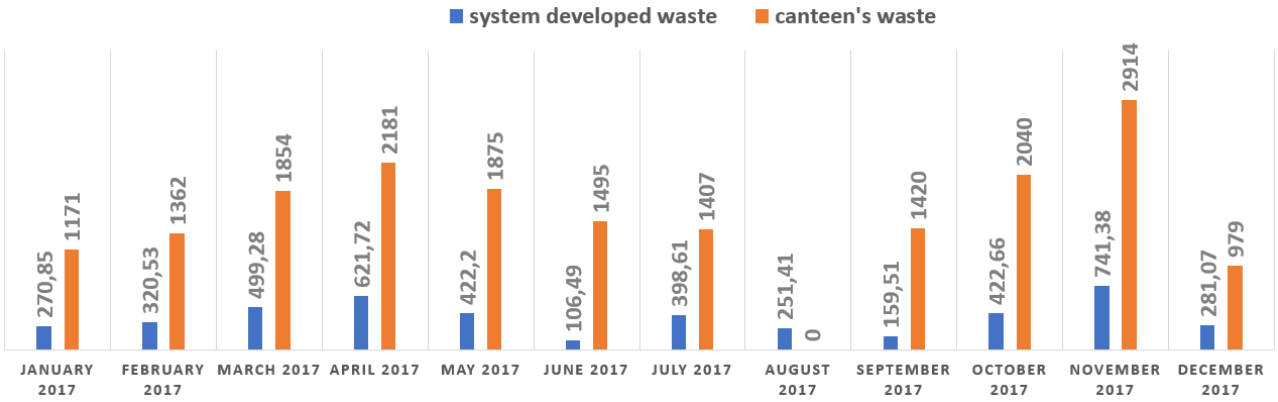

Figure 5.1: Comparison of the monthly waste generated by the system developed in this study and by the canteen forecasting system.

By averaging the ratios between the monthly wastes of 2017 obtained by the prediction method of food demand currently applied by the canteen and the same waste this time obtained by the three proposed models, it was concluded that the resulting waste of the latter is five times lower than the waste resulting from the former, as stated in Table C.1. In the same table it is possible to notice that in the month of August the model made a prediction less sustainable and more erroneous than the prediction made, for the same month, by the method applied by the canteen. As of all the years considered in the dataset, only in 2017 it is verified that the canteen was open in the month of August, it was expected that the models proposed had difficulty in capturing this information and, consequently, to make a more accurate prediction. Such a limitation could have been remedied by the existence of more data, proving once again that the lack of data was a major obstacle to obtaining less significant errors.

To observe how well the models developed for predicting meat, fish and vegetarian consumptions perform in relation to the canteen's same models, please refer to Figure 5.2. A model performs better than other when the former's MAE is lower than the latter's one.

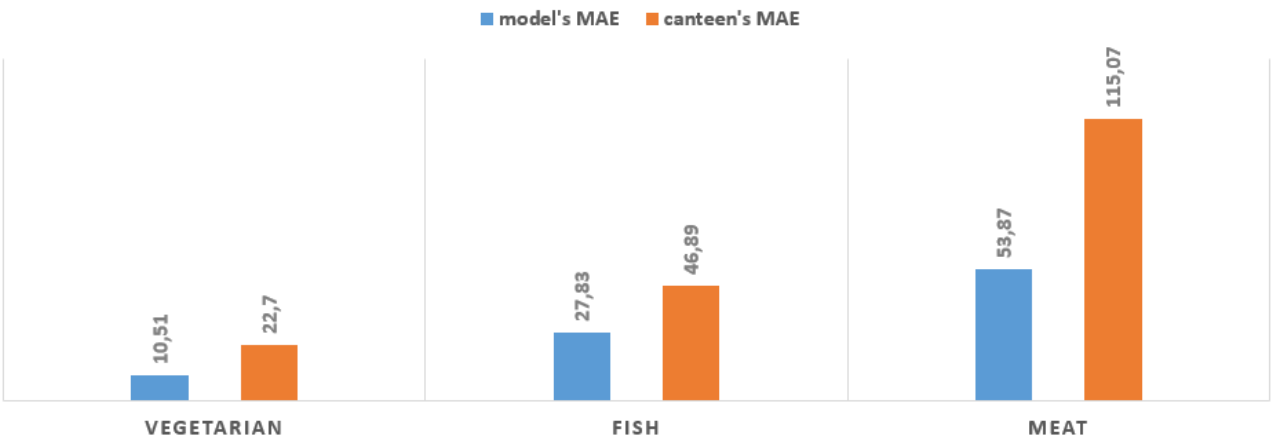

Figure 5.2: MAE for meat, fish and vegetarian consumptions' predictive models for both canteen's forecasting system and the one developed in this study.

SASUP provided an estimate of how much money would be saved for each meal not wasted as well as the amount of CO_2 emitted into the environment that could be avoided. Thus, in order to emphasize the contribution of the present study at environmental level but also at the financial level related to the canteen, a summary (see Tables 5.3 and 5.4), regarding year 2017, of the CO_2 emissions emitted as well as the amount of money wasted, was made. This summary compares the same results with the method currently applied by the canteen for predicting food demand. On the one hand, the more CO_2 emissions are avoided the greater the contribution of this study to reducing the impact of food waste on one of the most serious problems of today's society, global warming. On the other hand, if the financial resources of the canteen are used in the most effective and sustainable way possible, that is to say, in the sense of a balanced daily number of meals prepared compared to the daily demand, these same resources could be used in aspects like the improvement of the infrastructures of the canteen, quality of food as well as in the creation and offer of new menus.

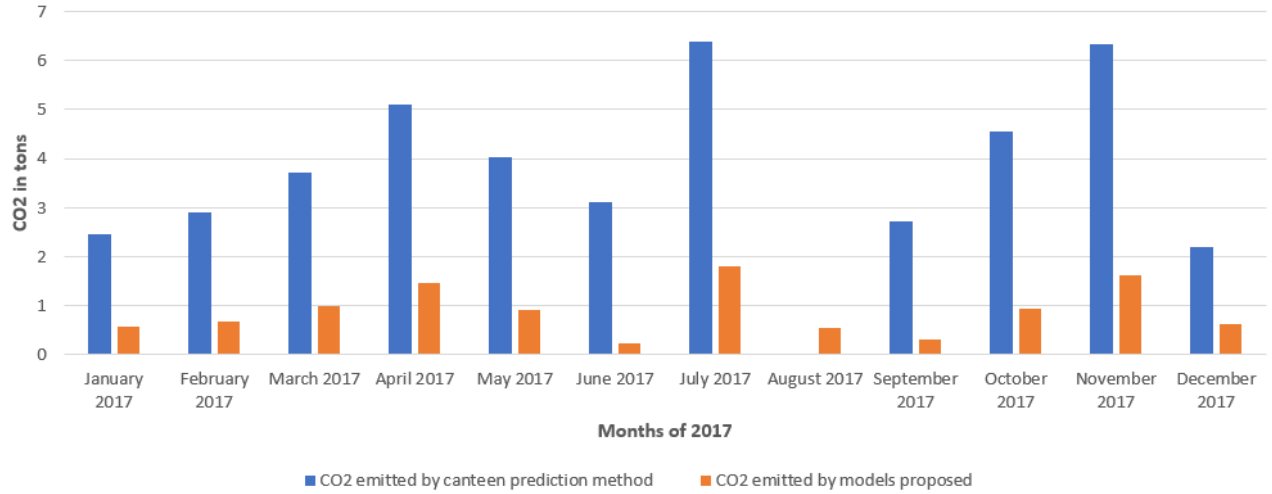

Figure 5.3: Monthly quantities of CO_2 emitted, in tons, by the canteen's prediction model and by the models proposed in this study.

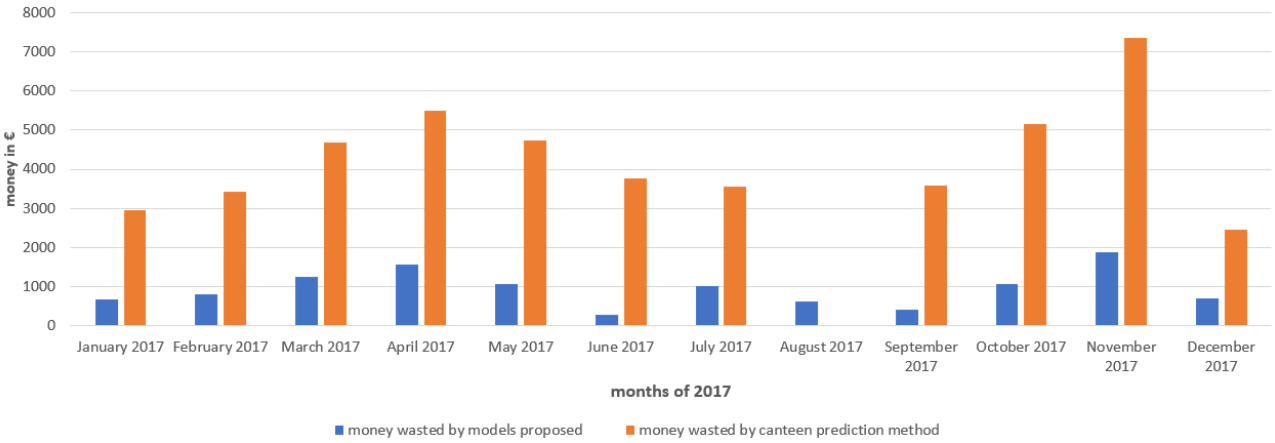

Figure 5.4: Monthly quantities of money *wasted*, in euros, by the canteen's prediction model and by the models proposed in this study.

By looking at the charts above (5.3 and 5.4), one may conclude that the models proposed have a greater impact in reduction of CO_2 emissions and also in saving money.

Chapter 6

Conclusions and Future Work

This chapter presents some of the conclusions regarding the results achieved across this study, as well as their impact in the main goal of this work - reducing food waste in FEUP's canteen.

6.1 Conclusions

Given that the number of meals sold for a particular menu on a given day is greatly influenced by the human capacity for decision and its associated uncertainties and randomness - since a human being is free to choose and change his mind whenever he wants - it was expected that the predicted values would not be 100% equal to the observed values. However, advanced data mining techniques, never before used in this context, were applied so that the predictive models built could result in a significant food waste reduction.

The datasets through which the models with the lowest generalization error were obtained were those entailing menus' categorization after applying feature selection on it, as described in 3.4.

For the prediction of the daily meat consumption were RFs to obtain the model with the lowest generalization error. It obtained an average error of approximately 54 plates, which means a lower average error in 61 plates when compared to the method of meat demand management currently applied by the canteen, that in turn obtains an average error of 115 dishes (see Figure 5.2). The variables related to the number of students attending integrated masters and undergraduate studies who have classes all day long - *MI_Todo_dia* and *L_Todo_dia* - are the ones that most impact the performance (see Figure 4.5) of the prediction model of daily meat consumption. Combining the fact that customers are mostly students and they usually prefer the meat dish rather than the fish or vegetarian ones, to the fact that students who have classes all day also spend all day nearby the faculty and as such are more prone to lunch in the canteen, it was expected that these variables would be the most striking.

Regarding the prediction of daily fish consumption, SVRs provided the model with the lowest generalization error. This model obtained an average error of approximately 28 dishes, whereas the prediction method of the daily fish consumption currently applied by the canteen obtains an average superior error in 19 plates, that is to say, a total of 47 dishes approximately. One of the most striking variables is once again the one referring to integrated masters students with classes all day. Also the variable referring to the maximum temperature in the lunch hour also has a huge impact, which is reasonable since, normally, when compared to the meat dish, the fish dish is a softer and fresher dish, preferable in days of high temperatures. The variable referring to the "pataniscas bacalhau" menu category also has a great impact on the demand for fish dishes (see Figure 4.11).

Finally, with regard to the daily vegetarian consumption, once again it was the SVRs that resulted in building the least erroneous model. This model obtained an average generalization error of approximately 11 dishes, which corresponds to half of the average error obtained by the prediction method of vegetarian daily consumptions currently applied by the canteen, i.e. 23 dishes. As for the most impacting variables in the performance of the proposed model, once again the variable MI_Todo_dia was one of the most determinant. The variable referring to the menu category "pizza" was also very relevant in obtaining the final performance of the model, since, usually, vegetarian dishes are only chosen more often if this is "pizza", as can be observed in Figure 4.10.

For all proposed models, the impact of the variable referring to weeks in which there are no classes - *0thWeek* - is notorious since, whenever there are no classes, the number of students nearby the faculty is way smaller and as such fewer customers than usual are expected in the canteen.

Based on the models proposed, capable of predicting meat, fish and vegetarian daily consumptions, it is possible to achieve a monthly average reduction of waste of approximately 491% (see Figure C.1).

Although the very promising results, it is important to note that this prediction system is very specific to FEUP's canteens. The results achieved might not be directly applicable to other canteens, due to the fact that the models proposed are based on factors not applicable to other canteens, such as students timetable, weather conditions, etc.

6.2 Future work

Given the nature of this work, some potential future work may be useful to reduce food waste, not only at FEUP but also at other institutions, making its positive impact in problems such as *global warming*, even greater and striking. Below it is possible to find some potential future work points:

- Modularize software so that it can be adapted and integrated into other food facilities, since from one to another the surrounding environment changes and therefore predictor variables might change also.

- Create an user-friendly platform designed for the canteen's managers. In this platform they would be eventually provided with a dialog box where they can insert all the values for the predictor variables in that specific day, hence getting a prediction about the food demand. Such would make predictions more human-readable and interpretable, and thus turning management functions, particularly planning [RS03], way easier.
- Collect more data, since it can help models catching more relationships between the predictor variables and the predicted one. That way models might improve their ability to predict food demand.
- Collecting more data would allow the collection of more variables, for instance, food demand in the homologous day of the previous year and also the mean sales across more than one year, for each menu.
- Adopt unsupervised learning techniques, so that models built are always learning from new data and therefore getting more accurate and real predictions, since new data gives an insight about new trends and patterns of demand.
- Adopt other methods, besides grid search, of tuning SVR and ANN hyper-parameters, such as genetic algorithms - an algorithm with a wide range of application when one wants to perform optimization tasks - as proposed in [CW07].
- Since, normally, the confection form has an impact on the preferences of the customers, it would be worthwhile to categorize the confection of each dish, for example, if it is fried, roasted, cooked, etc. A categorization at the level of allergenic composition would also be expected to impact. For example, if it contains eggs, lactose or even gluten. It would also be relevant to be able to collect for each menu of each type of dish its amount of calories and its own side dish.

References

[AMR17] Muhammad Waseem Ahmad, Monjur Mourshed, and Yacine Rezgui. Trees vs neurons: Comparison between random forest and ann for high-resolution prediction of building energy consumption. *Energy and Buildings*, 147:77 – 89, 2017.

[ANJ16] Lasek A., Cercone N., and Saunders J. *Leon-Garcia A. et al. (eds) Smart City 360*, volume 166, chapter Restaurant Sales and Customer Demand Forecasting: Literature Survey and Categorization of Methods, pages 479–491. Springer, Cham, June 2016.

[Bre01] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, Oct 2001.

[BS09] Ahmet Selman Bozkir and Ebru Akcapinar Sezer. Usage of data mining techniques in discovering the food consumption patterns of students and employees of university. 2009.

[BS11] Ahmet Selman Bozkir and Ebru Akcapinar Sezer. Predicting food demand in food courts by decision tree approaches. *Procedia Computer Science*, 3(Supplement C):759 – 763, 2011. World Conference on Information Technology.

[CW07] Kuan-Yu Chen and Cheng-Hua Wang. Support vector regression with genetic algorithms in forecasting tourism demand. *Tourism Management*, 28(1):215 – 226, 2007.

[GCS+11] Jenny Gustavsson, Christel Cederberg, Ulf Sonesson, Robert van Otterdijk, and Alexandre Meybeck. Fao. global food losses and food waste - extent, causes and prevention. Technical report, FAO and Swedish Institute for Food and Biotechnology (SIK), May 2011.

[GEM98] Zhang G., Patuwo B. E., and Hu M.Y. Forecasting with artificial neural networks: The state of the art. In *International Journal of Forecasting*, pages 35–62, 1998.

[LBS+01] Lon Mu Liu, Siddhartha Bhattacharyya, Stanley L. Sclove, Rong Chen, and William J. Lattyak. Data mining on time series: An illustration using fast-food restaurant franchise data. *Computational Statistics and Data Analysis*, 37(4):455–476, 10 2001.

[LP05] A.A. Levis and L.G. Papageorgiou. Customer demand forecasting via support vector regression analysis. *Chemical Engineering Research and Design*, 83(8):1009 – 1018, 2005.

[LS17] Xinliang L. and Dandan S. *Tan Y., Takagi H., Shi Y., Niu B. (eds) Advances in Swarm Intelligence*, volume 10386, chapter University Restaurant Sales Forecast Based on BP Neural Network - In Shanghai Jiao Tong University Case., pages 338–347. Springer, Cham, June 2017.

REFERENCES

[MMB91] Judy L. Miller, Cynthla S. McCahon, and Brenda K. Bloss. Food production forecasting with simple time series models. *Hospitality Research Journal*, 14(3):9–21, 1991.

[PA09] S.P. Cowpertwait Paul and V. Metcalfe Andrew. *Introductory Time Series with R*. Springer, 2009.

[PM96] Michelle J. Pickert and Judy Miller. Food production forecasting in six commercial foodservice operations: a pilot study. *Hospitality Research Journal*, 20(2):137–144, 1996.

[RS03] Kisang Ryu and Alfonso Sanchez. The evaluation of forecasting methods at an institutional foodservice dining facility. *The Journal of Hospitality Financial Management*, 11(1):27–45, 2003.

[SBZH07] Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(1):25, Jan 2007.

[SCJ98] Makridakis S., Wheelwright S. C., and Hyndman R. J. *Forecasting methods and applications*. Wiley, 1998.

[SJH13] Benkachcha S., Benhra J., and El Hassani H. Causal method and time series forecasting model based on artificial neural network. In *International Journal of Computer Applications*, pages 0974–8887, 2013.

[SM95] A. Sanchez and J. L Miller. The evolution of an expert system in foodservice forecasting. *National Association of College and University Foodservice*, 19:61–71, 1995.

[SS04] Alex J. Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, August 2004.

[TSK05] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2005.

[Vap95] V. Vapnik. *The nature of statistical learning theory*. Springer, New York, 1995.

[Wor] World weather online api. https://developer.worldweatheronline.com/api/. Accessed 26 Feb. 2018.

[ZH] Patuwo B. Zhang, G. and M. Hu. Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting*, 14:35–62.

Appendix A

Variables

A.1 Example of a dataset observation

Table A.1: Example of a dataset observation.

variable	**value**
day	28
month	4
rojoes a moda do minho (meat)	1
carapau grelhado com arroz de tomate (fish)	1
folhada de seitan e alho frances com arroz primavera (vegetarian)	1
remaining_menus[(1)]	0
weekOfClasses	eleventh
enrollmentWeek	0
1stYear1stWeek	0
1stYear2ndWeek	0
valentinesDay	0
superMegaFeupCaffe	1
carnivalHolidays	0
queimadasfitasWeek	0
engineeringWeek	0
christmasHolidaysWeek	0
easterHolidaysWeek	0
supplementaryExamsSeason	0
ordinaryExamsSeason	0
wasHolidayThreeDaysAgo	1
wasHolidayTwoDaysAgo	0
wasHolidayOneDayAgo	0
isHoliday_In_A_Day	0
isHoliday_In_Two_Days	0
isHoliday_In_Three_Days	0
supplementaryExamsSeason	0
1stYearSupplementarySeasonPreparation	0
D_Tarde	7
L_Manha	8
L_Tarde	71
L_Todo_dia	16
MI_Manha	1392
MI_Tarde	1175
MI_Todo_dia	1574
M_Manha	12
M_Tarde	0
M_Todo_dia	0
queimaWednesday	0
praxe	eletro, mecanica
weekDay	4
maxTempC	18
precipitation	0
weatherDescription	sunny
sales	542

Notes

[1] Field *remaining menus* comprises a total of *N* variables set to 0 and proportional to the number of meat, fish and vegetarian menus offered by the canteen across the observations present in the respective dataset, over 2016 and 2017.

For each of the binary variables referring to special weeks or seasons (see table 3.1), if a given observation - day - underlies the condition implied by such variable, it is marked with a 1 in the respective field or with a 0 otherwise.

Appendix B

Results

B.1 Models' performances

Meat

- Menus dataset

  - Without feature selection

Table B.1: Results on meat dataset with menus descriptions, without feature selection.

	Random Forest	**Support Vector Regression**	**Artificial Neural Network**
R^2	0,81	0,77	0,63
MAE	54,96	60,01	82,12
Positive deviation	27,25	30,16	48,64
Negative deviation	-27,71	-29,85	-33,48
Hyper-parameters	*max_features* = 79	C = 8192 γ = 0.03125	*nodes* = 568 *learning_rate* = 0.215

Note: *nodes* refers to the number of nodes in the hidden layer.

– With feature selection

Table B.2: Results on meat dataset with menus descriptions, with feature selection.

	Random Forest	**Support Vector Regression**	**Artificial Neural Network**
R^2	0,82	0,80	0,68
MAE	53,99	56,47	75,62
Positive deviation	25,90	29,60	44,04
Negative deviation	-28,09	-26,87	-31,58
Hyper-parameters	*max_features* = 27	C = 8192 γ = 0.00195	*nodes* = 140 *learning_rate* = 0.464

Note: *nodes* refers to the number of nodes in the hidden layer.

- Types dataset

– Without feature selection

Table B.3: Results on meat dataset with menus categories, without feature selection.

	Random Forest	**Support Vector Regression**	**Artificial Neural Network**
R^2	0,81	0,78	0,55
MAE	54,03	60,47	92,65
Positive deviation	26,78	30,84	54,50
Negative deviation	-27,25	-29,63	-38,15
Hyper-parameters	*max_features* = 61	C = 2048 γ = 0,00781	*nodes* = 120 *learning_rate* = 0.464

Note: *nodes* refers to the number of nodes in the hidden layer.

Fish

- Menus dataset

  - Without feature selection

Table B.4: Results on fish dataset with menus descriptions, without feature selection.

	Random Forest	**Support Vector Regression**	**Artificial Neural Network**
R^2	0,33	0,26	0,35
MAE	32,06	33,49	31,99
Positive deviation	17,26	18,00	20,39
Negative deviation	-14,80	-15,49	-11,60
Hyper-parameters	*max_features* = 165	C = 8192 γ = 0,00012	*nodes* = 452 *learning_rate* = 0.100

Note: *nodes* refers to the number of nodes in the hidden layer.

  - With feature selection

Table B.5: Results on fish dataset with menus descriptions, with feature selection.

	Random Forest	**Support Vector Regression**	**Artificial Neural Network**
R^2	0,35	0,43	0,37
MAE	31,30	28,47	30,83
Positive deviation	16,82	15,03	16,44
Negative deviation	-14,48	-13,44	-14,39
Hyper-parameters	*max_features* = 87	C = 2048 γ = 0,00195	*nodes* = 152 *learning_rate* = 0.215

Note: *nodes* refers to the number of nodes in the hidden layer.

- Types dataset

  - Without feature selection

Table B.6: Results on fish dataset menus categories, without feature selection.

	Random Forest	**Support Vector Regression**	**Artificial Neural Network**
R^2	0,36	0,27	0,35
MAE	31,36	33,08	33,43
Positive deviation	16,32	18,10	26,90
Negative deviation	-15,04	-14,98	-6,53
Hyper-parameters	*max_features* = 129	C = 2048 γ = 0,00049	*nodes* = 336 *learning_rate* = 0.100

Note: *nodes* refers to the number of nodes in the hidden layer.

Vegetarian

- Menus dataset

  – Without feature selection

Table B.7: Results on vegetarian dataset with menus descriptions, without feature selection.

	Random Forest	**Support Vector Regression**	**Artificial Neural Network**
R^2	0,46	0,54	0,54
MAE	13,88	12,96	12,99
Positive deviation	7,39	6,82	8,12
Negative deviation	-6,49	-6,14	-4,87
Hyper-parameters	*max_features* = 20	C = 512 γ = 0,00781	*nodes* = 172 *learning_rate* = 0.100

Note: *nodes* refers to the number of nodes in the hidden layer.

  – With feature selection

Table B.8: Results on vegetarian dataset with menus descriptions, with feature selection.

	Random Forest	**Support Vector Regression**	**Artificial Neural Network**
R^2	0,54	0,63	0,67
MAE	12,94	11,36	11,07
Positive deviation	6,64	6,24	6,50
Negative deviation	-6,30	-5,12	-4,56
Hyper-parameters	*max_features* = 4	C = 8192 γ = 0,00049	*nodes* = 260 *learning_rate* = 0.100

Note: *nodes* refers to the number of nodes in the hidden layer.

- Types dataset

  – Without feature selection

Table B.9: Results on vegetarian dataset menus categories, without feature selection.

	Random Forest	**Support Vector Regression**	**Artificial Neural Network**
R^2	0,49	0,65	0,56
MAE	13,68	11,45	13,35
Positive deviation	6,86	6,05	9,82
Negative deviation	-6,82	-5,40	-3,54
Hyper-parameters	*max_features* = 8	C = 2048 γ = 0,00195	*nodes* = 36 *learning_rate* = 0.215

Note: *nodes* refers to the number of nodes in the hidden layer.

Appendix C

Discussion

C.1 Reduction of waste by the three models proposed comparatively to the prediction method of food demand currently applied by the canteen

Table C.1: Waste reduction by the three models proposed.

month	**system developed waste**[1]	**canteen's waste**[2]	**times of reduction**	**% of reduction**
January	270,85	1171	4,32	432,3
February	320,53	1362	4,25	424,9
March	499,28	1854	3,71	371,3
April	621,72	2181	3,51	350,8
May	422,2	1875	4,44	444,1
June	106,49	1495	14,04	1403,9
July	398,61	1407	3,53	353
August	251,41	0	0	0
September	159,51	1420	8,9	890,2
October	422,66	2040	4,83	482,7
November	741,38	2914	3,93	393,1
December	281,07	979	3,48	348,3
mean	374,6425	1558,166667	4,911666667	491,2166667
	374,6	**1558,2**	**5**	**491**

Notes
[1] System developed waste is obtained through the sum of the negative deviations (see section 4.2 for further details in what it is a negative deviation) verified on each day and dish type, for all the months of 2017.
[2] The same holds for the canteen's waste.

Appendix D

Menus and categorization of menus

This section gives an example of the description of the menus offered by the canteen during the years 2016 and 2017, in this case relating to the vegetarian dish, but also exemplifies the categorization of the same menus. In the original dataset, the descriptions of the different menus, whether meat, fish or vegetarian, were used as the possible values of a variable *menu*. For the second dataset, which includes the categorization of these same menus, their categories compose the set of possible values for a variable *menu category*.

D.1 Menus descriptions

Table D.1: Statistics summary vegetarian menus.

menu	count	frequency	frequency in %	sales mean	total sales
feijoada de seitan e espinafres com arroz	2	0,005	0,5	143,5	287
bolonhesa de tofu com esparguete	3	0,007	0,7	124	372
tarte de queijo e cogumelos	9	0,021	2,1	93,2	839
pizza com ovo	6	0,014	1,4	88,2	529
pizza calzone com ovo	3	0,007	0,7	79,7	239
lasanha com legumes, seitan e coco	1	0,002	0,2	76	76
crepe de legumes e seitan com batata salteada	1	0,002	0,2	75	75
pizza de legumes com ovo	11	0,025	2,5	74,4	818
carbonara de cogumelos e queijo	7	0,016	1,6	71,3	499
alho frances salteado com seitan	1	0,002	0,2	68	68
lasanha de brocolos e queijo	7	0,016	1,6	66,3	464
guisado de feijao preto com legumes e arroz branco	1	0,002	0,2	65	65
lasanha de soja	7	0,016	1,6	64,1	449
trouxa de seitan e cogumelos com arroz	1	0,002	0,2	62	62
massa gratinada com soja	7	0,016	1,6	61,4	430
folhados de cogumelos e soja com arroz	1	0,002	0,2	60	60
patanisca de legumes com arroz de cenoura	7	0,016	1,6	59,6	417
lasanha de legumes com soja	5	0,011	1,1	59,4	297
lasanha de legumes	7	0,016	1,6	58,9	412

Menus and categorization of menus

menu	count	frequency	frequency in %	sales mean	total sales
gratinado de legumes e cogumelos	8	0,018	1,8	58,4	467
massa farfalle salteada c/ soja e legumes chineses	1	0,002	0,2	58	58
crepes recheados c/queijo e espinafres	7	0,016	1,6	57,3	401
pataniscas de courgette e lentilhas	7	0,016	1,6	57	399
strudel com espinafres e soja com arroz	1	0,002	0,2	56	56
massa gratinada com soja e cogumelos	6	0,014	1,4	55	330
lasanha de lentilhas	6	0,014	1,4	54,8	329
lentilhas estufadas e legumes grelhados	2	0,005	0,5	51,5	103
crepe de legumes e ovo	7	0,016	1,6	51,1	358
bolonhesa de seitan com esparguete	9	0,021	2,1	50,9	458
fusilli com seitan e legumes	1	0,002	0,2	50	50
bolonhesa de seitan com cogumelos	1	0,002	0,2	49	49
massa fusilli gratinada com legumes e queijo	9	0,021	2,1	48,2	434
bolonhesa de soja com massa	6	0,014	1,4	48,2	289
massa tricolor com tofu e ervilhas	1	0,002	0,2	48	48
lasanha de brocolos	1	0,002	0,2	48	48
bolonhesa de legumes	5	0,011	1,1	47,8	239
bolonhesa de legumes com massa fusilli	4	0,009	0,9	46	184
empadao com legumes chineses	5	0,011	1,1	45	225
rancho vegetariano (soja e grao)	1	0,002	0,2	45	45
crepe de feijao, legumes e ovo com arroz	2	0,005	0,5	44	88
tomates recheados com legumes e soja	1	0,002	0,2	44	44
fusilli salteado com seitan e ananas	2	0,005	0,5	43,5	87
folhados de tofu com legumes com arroz de tomate	1	0,002	0,2	43	43
jardineira de legumes e feijao	1	0,002	0,2	43	43
pataniscas de legumes com arroz de feijao seco	3	0,007	0,7	42,7	128
crepe de seitan e alho frances com lentilhas	5	0,011	1,1	42,4	212
tortilha de tofu	10	0,023	2,3	42,3	423
pimentos recheados com tofu e arroz	3	0,007	0,7	42	126
seitan assado com pera e batata assada	1	0,002	0,2	42	42

menu	count	frequency	frequency in %	sales mean	total sales
esparguete a bolonhesa com soja	8	0,018	1,8	41,6	333
empadao de legumes chineses e soja	6	0,014	1,4	41,2	247
legumes salteados com soja e massa fusilli	4	0,009	0,9	41	164
chili vegetariano de soja com arroz	1	0,002	0,2	41	41
tarte de queijo e seitan	1	0,002	0,2	40	40
strogonoff de cogumelos e seitan com esparguete	3	0,007	0,7	39,7	119
estufado de tofu com arroz de ervilhas	2	0,005	0,5	39,5	79
empadao com brocolos e feijao vermelho	1	0,002	0,2	39	39
arroz xau xau com lentilhas	1	0,002	0,2	39	39
seitan de cebolada	8	0,018	1,8	39	312
folhados de cogumelos e soja com pure de batata	1	0,002	0,2	39	39
caril de tofu, grao e maca com arroz de ervilhas	1	0,002	0,2	37	37
espetada de tofu e legumes	8	0,018	1,8	36,9	295
strogonoff de seitan	7	0,016	1,6	36	252
estufado de seitan com legumes e salada russa	1	0,002	0,2	36	36
arroz a valenciana com soja	9	0,021	2,1	35,8	322
tomate recheado com cuscuz de legumes	4	0,009	0,9	35,8	143
trouxa de ovo e feijao com arroz	4	0,009	0,9	34,8	139
folhado de seitan e alho frances com fusilli	4	0,009	0,9	34,8	139
ratatouille com seitan e arroz	1	0,002	0,2	34	34
pudim de espargos com arroz	5	0,011	1,1	33,2	166
legumes salteados c/soja em tosta de pao integral	6	0,014	1,4	33	198
massa a carbonara com tofu	3	0,007	0,7	33	99
tortilha de legumes	4	0,009	0,9	32,8	131
massa salteada com soja e legumes	1	0,002	0,2	32	32

Menus and categorization of menus

menu	count	frequency	frequency in %	sales mean	total sales
tofu assado c/pimentos	8	0,018	1,8	31,6	253
ratatouille com tofu e arroz	6	0,014	1,4	31,3	188
bolonhesa de seitan com cogumelos e legumes	2	0,005	0,5	30	60
hamburguer de soja com arroz de ervilhas	3	0,007	0,7	29,7	89
folhado de seitan e alho frances com arroz primavera	4	0,009	0,9	29	116
chili de feijao branco com arroz	2	0,005	0,5	29	58
pudim de espargos	3	0,007	0,7	28,7	86
caril de tofu e ervilhas	7	0,016	1,6	28,6	200
empadao de legumes e seitan	11	0,025	2,5	28,4	312
arroz xau xau com tofu	10	0,023	2,3	28	280
estufado de grao e legumes com polenta de soja	6	0,014	1,4	26,2	157
legumes com ovo e feijao em folha de brick	5	0,011	1,1	26	130
jardineira de tofu	11	0,025	2,5	25,7	283
crepe de legumes e ovo com arroz	4	0,009	0,9	25,3	101
massa cotovelo salteada c/ soja e legumes chineses	3	0,007	0,7	24,7	74
rancho vegetariano com grao-de-bico e soja	1	0,002	0,2	24	24
favas guisadas com alho frances e arroz	9	0,021	2,1	23,4	211
peixinhos da horta com arroz de feijao vermelho	4	0,009	0,9	23,3	93
moqueca de tofu	6	0,014	1,4	23,2	139
cuscus de soja e legumes grelhados	4	0,009	0,9	23	92
empadao de soja com legumes	4	0,009	0,9	22	88
feijoada de tofu e farofa	5	0,011	1,1	21,6	108
caril de tofu com fusilli	3	0,007	0,7	19	57
tarte de abobora com grao	1	0,002	0,2	19	19
caril de maca e soja com arroz	4	0,009	0,9	17,5	70
seitan no forno	1	0,002	0,2	16	16
brocolos gratinados com soja	1	0,002	0,2	16	16
crepe de cogumelos	1	0,002	0,2	15	15
quiche de brocolos e soja	1	0,002	0,2	14	14
empadao de soja e brocolos	1	0,002	0,2	14	14
salada de grao-de-bico com tofu	3	0,007	0,7	13	39
penne gratinado com seitan, feijao e legumes	1	0,002	0,2	11	11
caril de tofu com maca	1	0,002	0,2	10	10
empadao de legumes	1	0,002	0,2	10	10
feijoada de farofa com tofu	1	0,002	0,2	8	8

D.2 Menus categorization

Table D.2: Statistics summary of vegetarian menu categories.

menu category	frequency	frequency in %
alho frances	0,023	2,3
arroz	0,046	4,6
bolonhesa	0,08	8
bolonhesa de tofu	0,007	0,7
caril de tofu e ervilhas	0,018	1,8
caril de tofu e maca	0,018	1,8
chilli	0,007	0,7
cogumelos e queijo	0,037	3,7
combinados	0,05	5
crepe	0,064	6,4
empadao	0,067	6,7
estufado	0,021	2,1
feijoada de seitan	0,005	0,5
feijoada de tofu	0,014	1,4
folhado de seitan	0,018	1,8
folhados de cogumelos	0,005	0,5
fusilli com seitan	0,007	0,7
gratinado de legumes	0,018	1,8
guisado de legumes	0,007	0,7
jardineira	0,028	2,8
lasanha	0,078	7,8
legumes salteados	0,023	2,3
massa com tofu	0,009	0,9
massa gratinada	0,05	5
massa salteada com soja e legumes	0,011	1,1
patanisca	0,039	3,9
pizza	0,046	4,6
pudim de espargos	0,018	1,8
rancho	0,005	0,5
ratatouille	0,016	1,6
saladas	0,018	1,8
seitan	0,023	2,3
strogonoff	0,023	2,3
tofu e legumes	0,021	2,1
tofu e pimentos	0,025	2,5
tomates recheados	0,011	1,1
tortilha	0,032	3,2
trouxa	0,011	1,1