A Business Recommender System Based on Zones and Commercial Data

Mai Abusair*, Rania Dameh, Ruba Egbaria, Salsabeel Alzaqa
Department of Computer Science
An-Najah National University, Palestine
Email: mabuseir@najah.edu, raniadameh2014@gmail.com, ruba.egb@gmail.com, salsabeel.alzaqa@gmail.com

Abstract—**In many countries people target different places to open a business and succeed in it. They may choose an unsuccessful business or the location does not need the type of this business. In this paper, we aim to improve the opportunity of choosing a correct business and location. We suggest an approach based on many principles of machine learning. The approach uses a prediction model based on analysing data about zones (areas) and their commercial services. The zones are classified using K-Means clustering method that depends on the number of same businesses and their costs averages in an area. To show the novelty of our work, we developed a system that implements the approach principles for several zones in Nablus city. We evaluate the work by running several test cases to show the system ability in recommending kinds of businesses.**

Index Terms—**Clustering, K-means Algorithm, Business Recommender System.**

I. Introduction

Generally, many people invest money in order to start new businesses. Many problems arise due to their incorrect decisions and their businesses failed after a period of time. This could happen for reasons related to the unexpected quality and income. Further, one of the essential reasons for business failures, is that, many business owners do not know exactly if the area lacks of need for the type of profession they invested in, or if it had a complete sufficiency [1].

Machine learning (ML) in businesses is the emerging trend of the modern world and provides various benefits to firms. It tries to reduce the overall cost of the business operations and save the money of the organization. It also helps the businesses to make smarter decisions in the business processes and also able to provide solutions to the business problems effectively [2].

In [3], the authors proposed a machine learning approach for predicting business success at the early stage, focusing on geographical, demographic, and basic information about the companies. They suggests using gradient boosting classifier for getting good results in their approach. And finally, their model can be used as a decision support system for venture capital funds to help find potentially successful companies.

In [4], the authors developed a system to improve the rate of business success in Nigeria and provided a platform for entrepreneurial improvement in the Nigerian Economy. Their prediction model is based on correlation analysis for the data pre-processing and the combination of Naive Bayes and J48 classification algorithm.

In the use of machine learning in recommender systems, authors in [5] suggested a recommender system that can help business owners in deciding where to start their businesses. They use data from Facebook and urban planning data. They investigated several classification algorithms and they ended up recommending random forest classifier. Their classifier is trained to compute the matching score between a business profile and a zone ID and, thus, recommends a zone for a business. From a business point of view, they didn't consider zone need, capital funds, or more detailed location for a business in a zone.

In this paper, we suggest an approach to predict the suitable business before the business starts. The approach tries to solve the problem of choosing the correct business and zone, and so it can save time, resources and capital funds. The approach uses k-means clustering for classifying the zones. Moreover, it uses Naive Bayes for probabilistic computations for the kind of businesses, streets categories and capitals. In addition, we developed a system that deploys the approach in order to recommend businesses for people who wants to starts their successful investment in selected areas. Moreover, the system is able to recommend a suitable area for a business.

The paper is organized as follows: Section II shows a motivating scenario. Section III shows the data set. Section IV describes the approach. The evaluation is discussed in Section V. Finally, Section VI concludes the paper and suggests future works.

II. Motivating Scenario

In the local area, where we live, we have noticed that many new commercial shops in the surrounding zones closed after a short period of time.

According to our research on this field, the reasons of this phenomenon could be: Firstly, the lack of diversity in professions in a relatively small geographical area. This may cause a reduced sales in frequent stores. Secondly, a business owner does not consult specialized people before starting a business. So, it is possible to choose a business that is not suitable for the area. Thirdly, starting a project with the available money owned by the business owner, with a limited knowledge of the business field and its financial requirements in the future. This will lead to an accumulated debts.

On the other hand, we realized that a business owner, who is interested in opening a project in a specific area and within a certain capital fund, does not have an idea of the commercial shops needed in that area. Accordingly, this paper suggests an approach that helps business owners and entrepreneurs to look for a suitable businesses and areas before staring their investments.

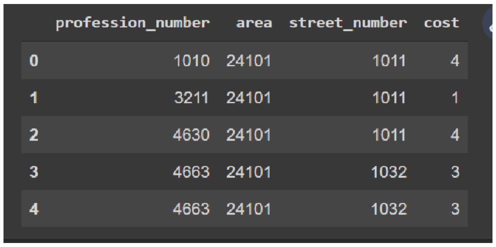
*Corresponding author: mabuseir@najah.edu

III. Preparing data set

At the first place, our work aims to classify local geographical areas and commercial shops. Thus, the data we dealt with are local data collected from the Municipality and the Chamber of Commerce in Nablus city in Palestine. Since we had the data collected from these two resources, we had as a result two files and we merge them in one file. From the data, we gathered all the kinds of commercial shops, the number of each kind of shops in every considered area, the shops locations (area number and street name) and their capital funds. Then, we manipulate the gathered data by adding extra classifications, like categorizing streets into main street, sub main street (for streets branched directly from main streets) and sub street (for internal streets that do not directly access main streets). After all, we deleted those records of incomplete data and we dropped unnecessary fields.

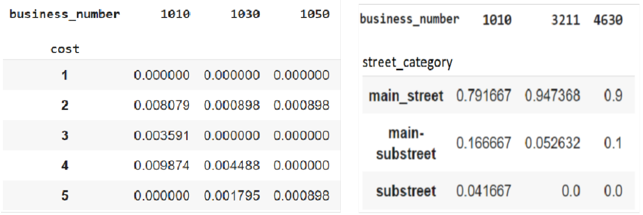
To prepare the data for our suggested approach, we filtered and encoded the data related to the shops and their locations. We relied on 4 columns, which are the area number (as classified in the local municipality), profession number (as classified in the municipality and chamber of commerce), street number (that represents street number and street category) and business cost (registered capital funds provided from Chamber of Commerce), see Figure 1 that shows a screen shot for the filtered data.

To explain the data in a detailed way, we better describe it as follows: (a) The area and profession are provided by Nablus municipality. The geographical areas are split ted into basins that have unique numbers, for simplicity, we call them areas. (for example 24101 is an area in Nablus). The profession number is classified as 4 digits (for example, a shop selling meat, poultry or fish has the number 1010 and so on for each kind of profession). (b) For the street number, we encoded the streets into 4 digits, the first three digits represents a unique serial number, and the last digit is for streets categorization; 1 is for main street, 2 is for main sub street and 3 for sub street. For example: given the street number 1032; 103 represents a street unique number (that has an equivalent name in the municipality) and 2 represents that its a main sub street. (c) For the cost (capital funds) we relied on Chambers of Commerce data. It includes 5 degrees, and each one has a specific range of capital fund. Degree *1* for 1000000+ $, *2* for 50000$ - 999999$, *3* for 25000$ - 49,999$, *4* for 15000$ - 24,999$, *5* for 5000$ - 14,999$.



	profession_number	area	street_number	cost
0	1010	24101	1011	4
1	3211	24101	1011	1
2	4630	24101	1011	4
3	4663	24101	1032	3
4	4663	24101	1032	3

Fig. 1. Filtered data



business_number	1010	1030	1050
cost			
1	0.000000	0.000000	0.000000
2	0.008079	0.000898	0.000898
3	0.003591	0.000000	0.000000
4	0.009874	0.004488	0.000000
5	0.000000	0.001795	0.000898

(a) Probability matrix for profession type with capital funds category

business_number	1010	3211	4630
street_category			
main_street	0.791667	0.947368	0.9
main-substreet	0.166667	0.052632	0.1
substreet	0.041667	0.0	0.0

(b) Probability matrix for profession type with street category

Fig. 2. Naive Bayes probabilities matrices

IV. The suggested approach

It is an approach that helps business owners and entrepreneurs decide the type of business to invest in and the suitable area for it. In this section, we will discuss how machine learning is carried out in the approach and how our business recommender system work. We will show these details in the stages described in the following sections.

A. Building a probabilistic model

In order to build a probabilistic model, first, we reason on our data set values through graphs that show the relationships between them. In addition to building a confusion matrix that represents the predicted and actual values of the data points. Therefore, we find that following Naive Bayes algorithm can fit to our purpose [6].

To show the probable cost for every kind of profession, we used Naive Bayes algorithm to get the probabilities for each kind of profession (business_number) and capital fund (cost) of the business in a matrix. See Figure 2a that shows a screen shot for the matrix. In the same way, to show how probable is the kind of profession to be found in particular street type, we got the probabilities between the kind of profession (business_number) and the street category (street_category) in

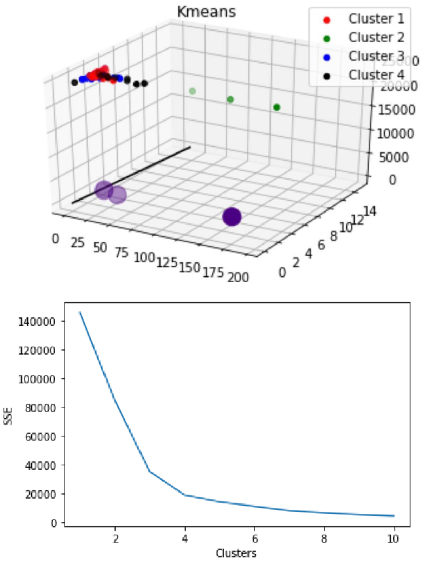


Fig. 3. K-mean clustering

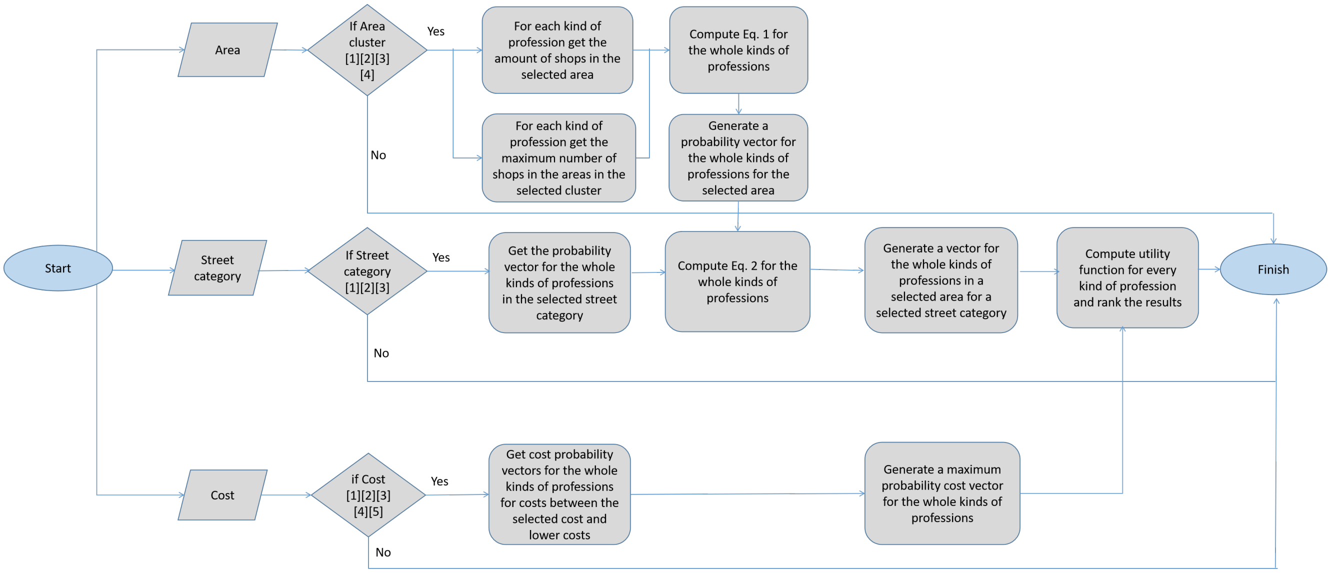

Start
Area
If Area cluster [1][2][3] [4]
Yes
No
For each kind of profession get the amount of shops in the selected area
For each kind of profession get the maximum number of shops in the areas in the selected cluster
Compute Eq. 1 for the whole kinds of professions
Generate a probability vector for the whole kinds of professions for the selected area
Street category
If Street category [1][2][3]
Yes
No
Get the probability vector for the whole kinds of professions in the selected street category
Compute Eq. 2 for the whole kinds of professions
Generate a vector for the whole kinds of professions in a selected area for a selected street category
Compute utility function for every kind of profession and rank the results
Finish
Cost
if Cost [1][2][3] [4][5]
Yes
No
Get cost probability vectors for the whole kinds of professions for costs between the selected cost and lower costs
Generate a maximum probability cost vector for the whole kinds of professions

Fig. 4. The suggested recommender approach

another matrix. See Figure 2b that shows a screen shot for the matrix. Using Naive Bayes for this purpose gives an accuracy of 72%.

B. Classifying areas using K-means clustering

To classify the areas we tried using algorithms that depends on features. We tried using k-nearest neighbour algorithm, but we excluded it since it depends on a preset classification for data. All what we need was having an algorithm that can classify the areas according to some characteristics.

Accordingly, we choose using K-means clustering [7]; an iterative unsupervised algorithm in which each observation belongs to the cluster with the nearest mean belonging to the same group share some key characteristics [8]. We applied K-means clustering algorithm that depends on two features as an input for the algorithm; the range number of business shops and the average of capital funds for shops in an area. In the outcome, the algorithm shows that the best fit is four clusters of areas as shown in Figure 3.

C. Preparing equations

For the sake of finding a way that can determine the shortage of businesses in an area, we benefit from the outputs of the previous sections in suggesting two main equations. As follows:

$$Y_S^A = (X_S^{AC} - N_S^A)/X_S^{AC} \quad (1)$$

$$U_S^{C_A} = Y_S^A * Z_S^C \quad (2)$$

Where

- Y is the shortage probability for a kind of profession S in a given area number A
- U is the shortage probability for a kind of profession S in a specific street category C in a given area A
- X is the maximum number of shops of a particular profession S in an area cluster AC
- N is the number of shops of a particular profession S in a given area A
- Z is the probability of the kind of profession S existence in a specific street category C

In Equation 1, to compute the shortage probability for a kind of profession in a given area number, we rely on two factors; The first factor is the maximum amount of a kind of profession (represented by business_number) existed among all the areas that belong to the given area's cluster. The second factor is the amount of the same kind of profession in the given area. We subtract the first factor from the second one and we divide them by the first factor. The output of Equation 1 is an input for Equation 2.

In Equation 2, to compute the shortage probability for a kind of profession in a particular street, we use the probability of a particular profession existence in a specific street category, and we multiply it with the shortage probability for the kind of profession existence in a given area (the outcome of Equation 1).

D. Operating the recommender system

To operate our work, we created a function that accepts 3 variables from the user (area number, street category, cost degree). See Figure 4 that shows the recommender system approach.

The first step is classifying the entered area. We searched in all the clusters to get in which cluster the entered area is. Then, we apply the Equation 1 for the whole kinds of professions in a cluster.

The second step is to use the entered street category to get the probability of which kind of professions are allowed to be in this street category. Then we apply Equation 2 for all the kinds of profession in a street category.

About the entered capital fund degree (cost degree), we compare it with the capital degrees in the data. If the degree is 1, we will consider the whole probabilities from degree 1 until degree 5 (note that degree 1 represents the highest capital fund). Thus, in order to present more options to the user, we consider probabilities for the entered degree by the user and all the degrees below it (Ex. if the user entered 1 so we will consider in the computations 1, 2, 3, 4 and 5, and if the user entered 3 we will consider in the computations 3,4 and 5 and so on). Eventually, we filtered out all the zero probabilities because they are useless in our study.

For the sake of ranking the recommended results, we use the following Utility Function $UF$to be computed for every kind of profession in given street category in a given area:

$$UF_S^{C_A} = W_{need}.U_S^{C_A} + W_{cost}.MaxCost_S \qquad (3)$$

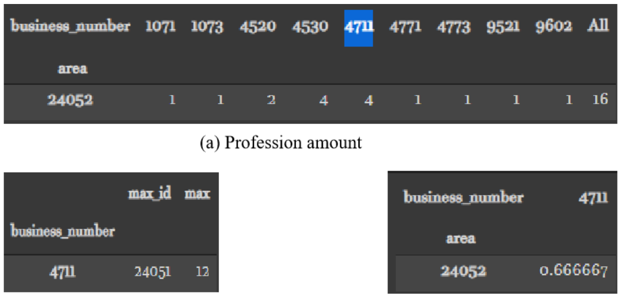
Where

- W_{need} represents the weight of importance for the shortage probability for professions in an area computed in Equation 2
- W_{cost} represents the weight of importance for the cost probability for a profession
- $MaxCost_S$ represents the maximum probability cost extracted for a given profession within the degrees selected

The formula in Equation 3 aims to rank the results by giving a weight of importance for the shortage probability of a profession in an area (here we will consider it 0.7), and a weight of importance for the cost probability of a profession (here we will consider it 0.3). The maximum cost considered in the equation is related to user selection for the cost degree (for example, if the user selects degree 3 that means we will consider the maximum value of the probabilities among the probability costs computed for degree 3, 4 and 5). See Figure 4 that summarizes the recommender system approach. To clarify these computations, we will show a clear example in the next section.

V. Evaluation

To evaluate our work, we developed a web application that deploys the suggested approach. We mainly rely on Python and Pandas to implement the system functionality. We included data for 72 areas belongs to Nablus city with their shops. The system offers a main service in which the user can enter the area number, street category, and the capital fund (cost) he/she wishes to invest. Then, the system will suggest a list of recommended business shops suitable for investment. Moreover, the system is able to recommend suitable areas for a selected business shop.



business_number	1071	1073	4520	4530	4711	4771	4773	9521	9602	All
area										
24052	1	1	2	4	4	1	1	1	1	16

(a) Profession amount

	max_id	max
business_number		
4711	24051	12

(b) Maximum number of shops

business_number	4711
area	
24052	0.666667

(c) The profession's need in an area

Fig. 5. Results of performing Equation 1

We examined the system randomly using 4 different areas (24052, 24013, 24032, 24111) belong to 4 different area clusters with the highest capital fund (cost degree 1) selection, and we tried them once with a main street selection, once with a main sub-street selection and once with a sub-street selection. This result in 12 test cases with the original data provided by the system. Then, to show the ability of the approach to detect the shortages of kind of shops in an area, we manipulated randomly the shops amounts in the original data for the selected areas, and we repeat running the test cases. Afterwards, we track a kind of profession related to supermarkets (has number 4711) and we reason on the differences in the computations and its index in the recommended list results.

To clarify the process we will show the computations by running an example on area 24052 that belongs to area cluster 1 in Section V-A and Section V-B.

A. Running an example for area 24052 before manipulating the data

The area 24052 belongs to the first area cluster. This cluster includes the areas that has similarity in the number of shops and the average of their costs. See Figure 5a that shows a screen for kinds of professions amounts in area 24052.

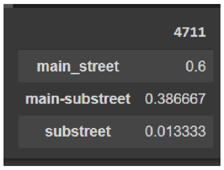
Recalling Equation 1, we need the maximum number of shops for a particular profession (4711) in the area cluster (area cluster 1). As it is shown in Figure 5b, the maximum number is 12. In other words $N_{4711}^{cluster1} = 12$. Additionally, as shown in Figure 5a, the area number 24052 has 4 super market shops (business_number 4711). This represents $N_{4711}^{24052} = 4$.

To apply Equation 1, which reflects the probability for the area (24052) need (shortage) of a specific profession (4711). $Y_{4711}^{24052} = (12 - 4)/12 = 0.667$

Where zero value for y means there is no shortage in this kind of profession and 1 means there is an insistent need for this kind of profession in the selected area. Therefore, in this example, the need of super market shops (4711) in area 24052 is 0.667. See Figure 5c.

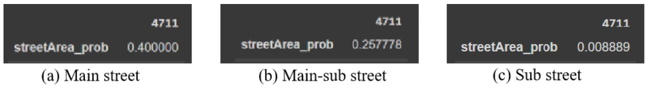
To apply Equation 2, which reflects the probability of the profession (4711) existence in a specific street category in an area (24052), see Figure 6 that shows the needed probabilities of existence of profession 4711 in the different street categories $Z_{4711}^{1} = 0.6$, $Z_{4711}^{2} = 0.386$ and $Z_{4711}^{3} = 0.013$.

Recalling Equation 2, and recalling Y_{4711}^{24052} value result from performing Equation 1, we compute the following for profession 4711:



	4711
main_street	0.6
main-substreet	0.386667
substreet	0.013333

Fig. 6. The probability of the profession existence in a specific street



	4711
streetArea_prob	0.400000

(a) Main street

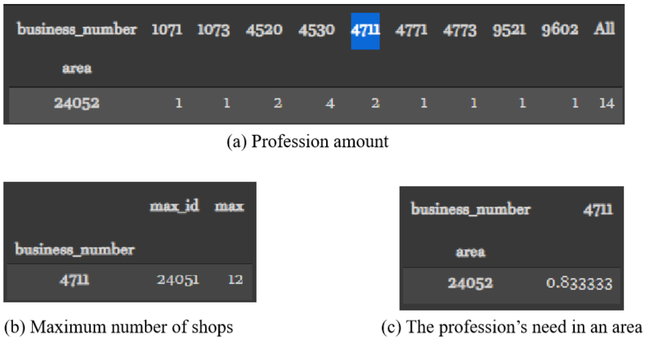
	4711
streetArea_prob	0.257778

(b) Main-sub street

	4711
streetArea_prob	0.008889

(c) Sub street

Fig. 7. The probability of the profession need in every street type after performing Equation 2



business_number	1071	1073	4520	4530	4711	4771	4773	9521	9602	All
area										
24052	1	1	2	4	2	1	1	1	1	14

(a) Profession amount

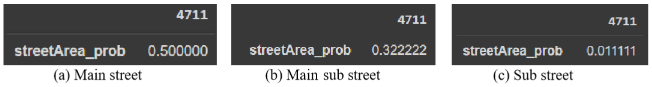
	max_id	max
business_number		
4711	24051	12

(b) Maximum number of shops

business_number	4711
area	
24052	0.833333

(c) The profession's need in an area

Fig. 8. Results of performing Equation 1 after manipulating data



	4711
streetArea_prob	0.500000

(a) Main street

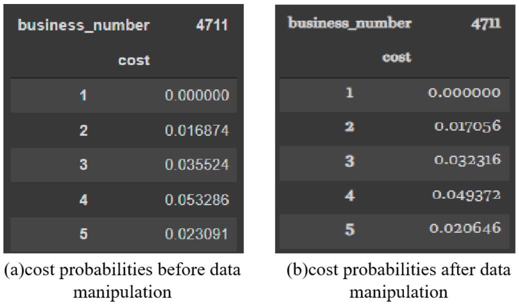
	4711
streetArea_prob	0.322222

(b) Main sub street

	4711
streetArea_prob	0.011111

(c) Sub street

Fig. 9. The probability of the profession need in every street type after performing Equation 2 on the manipulated data



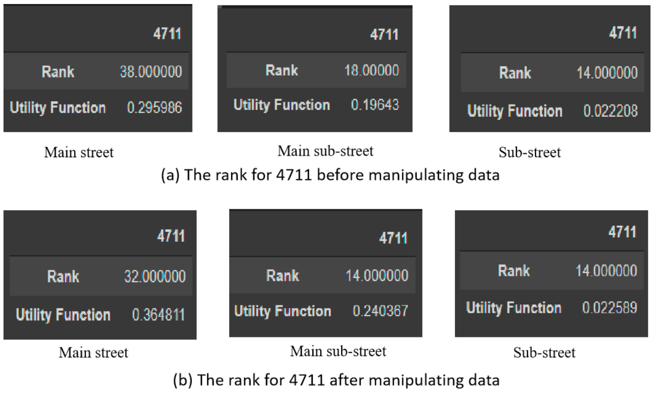
business_number	4711
cost	
1	0.000000
2	0.016874
3	0.035524
4	0.053286
5	0.023091

(a)cost probabilities before data manipulation

business_number	4711
cost	
1	0.000000
2	0.017056
3	0.032316
4	0.049372
5	0.020646

(b)cost probabilities after data manipulation

Fig. 10. Cost probabilities for profession 4711



	4711
Rank	38.000000
Utility Function	0.295986

Main street

	4711
Rank	18.00000
Utility Function	0.19643

Main sub-street

	4711
Rank	14.000000
Utility Function	0.022208

Sub-street

(a) The rank for 4711 before manipulating data

	4711
Rank	32.000000
Utility Function	0.364811

Main street

	4711
Rank	14.000000
Utility Function	0.240367

Main sub-street

	4711
Rank	14.000000
Utility Function	0.022589

Sub-street

(b) The rank for 4711 after manipulating data

Fig. 11. Utility function results and ranks in the different street categories

$U_{4711}^{1^{24052}} = 0.667 * 0.6 = 0.40$
$U_{4711}^{2^{24052}} = 0.667 * 0.386 = 0.257$
$U_{4711}^{3^{24052}} = 0.667 * 0.013 = 0.008$

See Figures 7a, 7b, and 7c that show computations for main street, main sub-street, and sub-street, respectively.

After all, in order to show the profession (4711) rank in the list of recommendations. We calculate the utility function for the whole professions and we rank the results according to utility function computed values.

Recalling Equation 3, and with considering a weight for the profession need probability $W_{need} = 0.7$ and a weight for the cost probability $W_{cost} = 0.3$ (these weights give good results and can be adjusted). For the sake of testing, we compute the utility function for profession 4711 in every street category in area 24052, as follows:

$UF_{4711}^{1^{24052}} = 0.7 * 0.4 + 0.3 * 0.053 = 0.295$
$UF_{4711}^{2^{24052}} = 0.7 * 0.257 + 0.3 * 0.053 = 0.195$
$UF_{4711}^{3^{24052}} = 0.7 * 0.008 + 0.3 * 0.053 = 0.022$

Since the computations are considered here for a user who selected the highest cost (degree 1). That means, we will consider degree 1 and the whole degrees below it (2,3,4 and 5) and we take the maximum value among them. Referring to Figure 10a, that shows the maximum probability value among the 5 cost degrees $MaxCost_{4711} = 0.053$.

After computing the UF for the whole kinds of profession, we got the ranks represented in Figure 11a for profession 4711.

B. Running an example for area 24052 after manipulating the data

After performing some shortages in the number of shops in the selected areas in the test cases. We recompute the same equations as in the previous section, so as to see the differences occurred in the probabilities and the final rank.

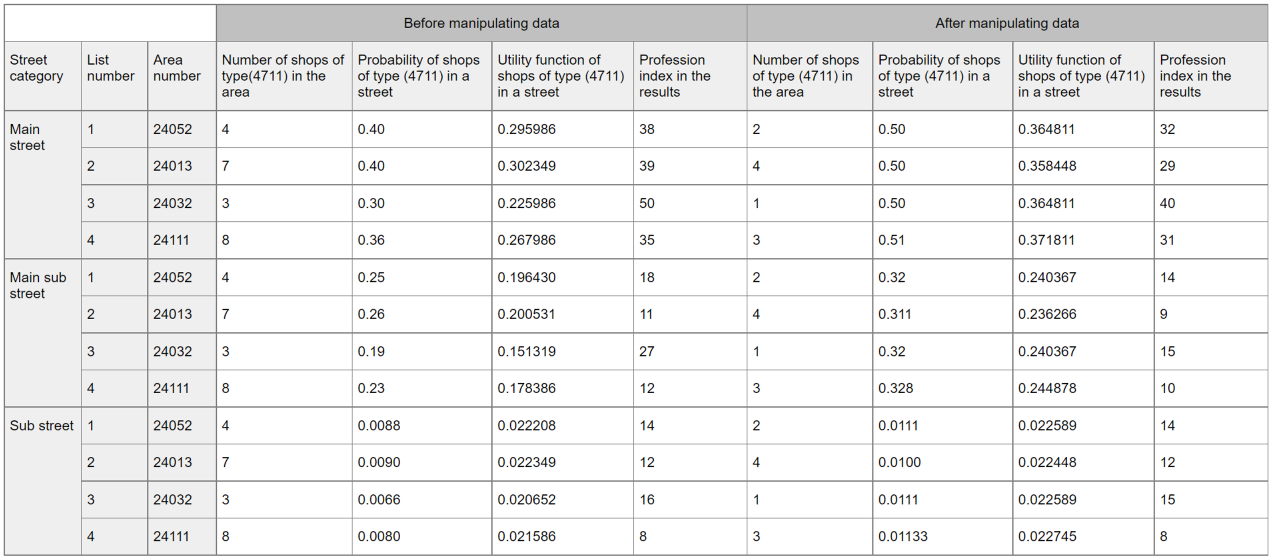
As shown in the data frame below, the amount of the shops in profession 4711 becomes 2. See Figure 8a that presents a screen shot for the profession amounts. As its shown in Figure 8b, the maximum number of of 4711 profession in area cluster 1 is 12. The need of super market shops (4711) in area 24052 is 0.8333. As shown in 8c.

To apply Equation 2, which reflects the probability of the profession (4711) existence in a specific street category in an area (24052). Again, see Figure 6 that shows the probabilities of existence of profession 4711 in the different street categories $Z_{4711}^{1} = 0.6$, $Z_{4711}^{2} = 0.386$ and $Z_{4711}^{3} = 0.013$.

Recalling Equation 2, and recalling Y_{4711}^{24052} value result from performing Equation 1, we compute the following for profession 4711:

$U_{4711}^{1^{24052}} = 0.833 * 0.6 = 0.5$
$U_{4711}^{2^{24052}} = 0.833 * 0.386 = 0.322$
$U_{4711}^{3^{24052}} = 0.833 * 0.013 = 0.011$

TABLE I
TEST CASES RESULTS



			Before manipulating data				After manipulating data			
Street category	List number	Area number	Number of shops of type(4711) in the area	Probability of shops of type (4711) in a street	Utility function of shops of type (4711) in a street	Profession index in the results	Number of shops of type (4711) in the area	Probability of shops of type (4711) in a street	Utility function of shops of type (4711) in a street	Profession index in the results
Main street	1	24052	4	0.40	0.295986	38	2	0.50	0.364811	32
	2	24013	7	0.40	0.302349	39	4	0.50	0.358448	29
	3	24032	3	0.30	0.225986	50	1	0.50	0.364811	40
	4	24111	8	0.36	0.267986	35	3	0.51	0.371811	31
Main sub street	1	24052	4	0.25	0.196430	18	2	0.32	0.240367	14
	2	24013	7	0.26	0.200531	11	4	0.311	0.236266	9
	3	24032	3	0.19	0.151319	27	1	0.32	0.240367	15
	4	24111	8	0.23	0.178386	12	3	0.328	0.244878	10
Sub street	1	24052	4	0.0088	0.022208	14	2	0.0111	0.022589	14
	2	24013	7	0.0090	0.022349	12	4	0.0100	0.022448	12
	3	24032	3	0.0066	0.020652	16	1	0.0111	0.022589	15
	4	24111	8	0.0080	0.021586	8	3	0.01133	0.022745	8

Thus, the probability of $U^{1_{24052}}_{4711}$ increased from 0.4 (before manipulating data) to 0.5. Moreover, the probability of $U^{2_{24052}}_{4711}$ increased from 0.257 to 0.322. Finally, the probability of $U^{3_{24052}}_{4711}$ increased from 0.008 to 0.011. Thus, the shortage is detected by the suggested algorithm.

See Figures 9a, 9b, and 9c that show computations for main street, main sub-street and sub-street, respectively.

Again, in order to show the profession (4711) rank in the list of recommendations; First of all, we refer to Figure 10b, that shows the maximum probability value among the 5 cost degrees $MaxCost_{4711} = 0.049$ (after manipulating data). Then, we calculate the utility function for the whole professions and we rank the results according to utility function computed values. Recalling Equation 3, the utility function for profession 4711 in every street category in area 24052 becomes as follows:

$$UF^{1_{24052}}_{4711} = 0.7 * 0.5 + 0.3 * 0.049 = 0.364$$
$$UF^{2_{24052}}_{4711} = 0.7 * 0.322 + 0.3 * 0.049 = 0.240$$
$$UF^{3_{24052}}_{4711} = 0.7 * 0.011 + 0.3 * 0.049 = 0.0225$$

After computing the UF for the whole kinds of profession, we got the ranks represented in Figure 11b. This shows an improvements in utility function computed, for the manipulated data, and the rank as well.

Table I shows the probabilities results and ranks for the test cases before and after manipulating the data with all the possibilities. Consequently, by reasoning on the results, we realized that the probabilities computed shows an increase in the profession need by an average of 5%, and an improvements in the profession rank by an average of 4.25 progressing positions, and, thus, a success for the approach.

VI. Conclusion and Future Work

We presented a new way of classifying the areas and business shops using machine Learning. We suggested an approach that tries to minimize the reasons of why some businesses fails. It is able to recommend options for investments to business owners. It focuses on the area need for kinds of professions along with their capital funds.

The challenge in this work is summarized in being able to build a correct data set, especially, if we want to consider large amount of zones. Data must be collected from different resources that do not share the same format and kind of classifications.

As a future work, we are willing to let customers give suggestions for new professions in the areas. Moreover, we wish to consider the population as a new factor in areas classification.

References

1 Lussier, R. N., "Reasons why small businesses fail: and how to avoid failure," *The Entrepreneurial Executive*, vol. 1, no. 2, pp. 10–17, 1996.

2 Sriram, V., Lakshmi, K., Podile, V., Naved, M., and Kumar, K., "Role of machine learning and their effect on business management in the world today," *International Virtual Conference on Innovation in Multidisciplinary Studies-IVCIMS 2021*.

3 Żbikowski, K. and Antosiuk, P., "A machine learning, bias-free approach for predicting business success using crunchbase data," *Information Processing & Management*, vol. 58, no. 4, p. 102555, 2021.

4 Afolabi, I., Ifunaya, T. C., Ojo, F. G., and Moses, C., "A model for business success prediction using machine learning algorithms," in *Journal of Physics: Conference Series*, vol. 1299, no. 1. IOP Publishing, 2019, p. 012050.

5 Lin, J., Oentaryo, R. J., Lim, E.-P., Vu, C., Vu, A., Kwee, A. T., and Prasetyo, P. K., "A business zone recommender system based on facebook and urban planning data," in *European Conference on Information Retrieval*. Springer, 2016, pp. 641–647.

6 Lowd, D. and Domingos, P., "Naive bayes models for probability estimation," in *Proceedings of the 22nd international conference on Machine learning*, 2005, pp. 529–536.

7 Li, Y. and Wu, H., "A clustering method based on k-means algorithm," *Physics Procedia*, vol. 25, pp. 1104–1109, 2012.

8 Teknomo, K., "K-means clustering tutorial," *Medicine*, vol. 100, no. 4, p. 3, 2006.