UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Economía

Clasificación de la Red Investigativa Ecuatoriana

Marcello Tomas Coletti Anzola

Economía

Trabajo de fin de carrera presentado como requisito
para la obtención del título de
Economista

Quito, 19 de mayo de 2023

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Economía

HOJA DE CALIFICACIÓN
DE TRABAJO DE FIN DE CARRERA

Clasificación de la Red Investigativa Ecuatoriana

Marcello Tomas Coletti Anzola

Pablo Astudillo Estévez,
Doctor en Geografpia Económica

Quito, 19 de mayo de 2023

© DERECHOS DE AUTOR

Nombres y apellidos: Marcello Tomas Coletti Anzola

Código: 00213379

Cédula de identidad: 1757378326

Lugar y fecha: Quit, 19 de mayo de 2023

ACLARACIÓN PARA PUBLICACIÓN

Nota: El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017) Discussion document on best practice for issues around theses publishing, disponible en http://bit.ly/COPETheses.

UNPUBLISHED DOCUMENT

Note: The following capstone project is available through Universidad San Francisco de Quito USFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on http://bit.ly/COPETheses.

RESUMEN

Las nuevas herramientas tecnológicas ayudan a mejorar la eficiencia del trabajo del investigador. Para muchos empresarios, formuladores de políticas, agentes financieros o cualquier persona interesada en estar preparada para el futuro, es importante conocer las nuevas herramientas analíticas. La Economía de la Complejidad está mejorando la forma y la precisión de la economía al introducir nuevos conceptos como el espacio de productos. Este trabajo utiliza técnicas de aprendizaje automático para clasificar el texto de los trabajos de investigación ecuatorianos en clasificaciones industriales. Con esto, este trabajo ofrece una forma para que los investigadores obtengan más información sobre su trabajo y cómo ha impactado, o está impactando, el entorno económico.

Palabras Clave: Teoría de Grafos, Complejidad, Investigación, Economía, Ciencia, Inteligencia Artificial

ABSTRACT

The new technological tools help improve the work of the efficiency of the researcher. For many businessmen, policymakers, financial agents, or everyone interested in being prepared for the future is important to know the new analytics tools. The Complexity Economy is improving the way and the precision of the economics by introducing new concepts as the product space. This work uses machine learning techniques to classify the text from the Ecuadorian research works into industrial classifications. With this, this work offers a way for researchers to get more information about their work and how it impacted, or how it is impacting, the economic environment.

Key words: Graph Theory, Complexity, Research, Economics, Science, Artificial Intelligence

CONTENT REVIEW

TABLES INDEX

FIGURES INDEX

INTRODUCTION

The Complexity Economy is improving the way and the precision of the economics by introducing new concepts as the product space. This work uses machine learning techniques to classify the text from the Ecuadorian research works into industrial classifications. With this, this work offers a way for researchers to get more information about their work and how it impacted, or how it is impacting, the economic environment.

The new tools from the AI and the analytical advances as Complexity Economics and automatic learning, should give an opportunity to innovate methodologies for the development of new research projects. The product space proposed by Hidalgo, Klinger, Barabási, and Hausmann (2007) works as a useful tool to understand the different agents from an economy by knowing their relationship with the other agents. The product spaces have been explored in the context of international commerce.

This work explores a bridge between the Academic and the Industry to know the proximity between them. The idea is to first build a machine learning model to adapt the research data to the industry and test the product space on scientific data but classified as industrial.

The next step consists in the dimensionality reduction of the dataset. For this part the work is based on the reflections or iterations established by the Economic Complexity Literature (Hidalgo & Hausmann, 2009). Those reflections will be explained in appendix A. In a short explanation, the reflections are measures of diversification and specialization, weighted by the information of the products/countries related. Those weighting works like normalizing the data and helping the

researcher reduce the dimensionality and getting more understandable relationships between the variables.

To know if this project was successful, or not, there are a few considerations. The model to classify the papers cannot be validated with a big number of humans. Which means that the model must be validated with more data. Even an unexpected failure in the classification model, the work attempted to replicate the Complexity Economic variables. And even a success situation on this area, that implies that the model cannot be considered validated but opens a door to redesign it.

Explained this scenario this work is not a test to if it is possible or not to merge the research network with the industry. This work is the first experiment on that agenda and searches to answer that was learned on the process.

THE AVAILIBABLE DATA

THE ECUADORIAN RESEARCH DATA

This work uses the data collected from 15,448 papers. Every paper got an identifier which enumerates the university which the paper is affiliated with. Table 1 shows an example of the base paper dataset. Other available variables are the paper publication year, the Ecuadorian province of the university, the city/municipality, a list of the authors, and the text from the title and keywords. The data was got online (Scopus, n.d.). The identifiers of the university is given by this work and represent the initials of the university name.

Table 2: Example of the Papers Dataset

id	year	province	city	authors	text
UPS62	2016	PICHINCHA	QUITO	Garc…	Dema…
UTPL464	2017	LOJA	LOJA	Pesá…	Imple…
USFQ986	2014	PICHINCHA	QUITO	Gam…	Wide…
UCUENCA365	2016	AZUAY	CUENCA	Ocho…	Impact…
USFQ1661	2005	PICHINCHA	QUITO	Aba…	Measure…

The text variable is required to classify the papers by the industries from the ISIC4. The authors will be the links between the papers. The province and the city are useful for future works, but

they are not currently required for the output of this paper. The year of the papers is in the range from 1921 to 2019. Based on the classification a subset of the data will be selected.

THE CORPUS

To classify the model, this work created a vocabulary (corpus) of acceptable words that belong to the intersection of the paper's texts and ISIC4 descriptions. Literature uses more than a billion words in the vocabulary (Mikolov et al., 2013), which implies an input of the same length to the model. However, those models are trained to generate text outputs which the capability to answer with English sentences. As my approach is only to classify, this work limited the corpus to the available words in the Ecuadorian papers. Then, is it expected to get more false negatives than false positives.

Special symbols ($, %, &…) and digits are not in the corpus. Also, the code separates the contractions (*you're* to *you are*)

After applying those changes, the corpus length I 617 words. There exist some corpuses with billion observations. However, working with those corpuses required more computational performance, scpecially in RAM management.

THE SAMPLE DATA

Whereas the initial paper's data contains more than 15 thousand observations, the descriptions from the ISIC4 are 419 observations. Also, the architecture of the classificatory model will not be simple, and it implies a risk of fall in overfitting the model. To solve this issue, this work increases the sample data size with information from other industrial classifications.

The Central Product Classification (United Nations, 2015) and the North American Industry Classification System (United States Census Bureau, 2019) can be related with the ISIC4 thanks to documentations in the official UN website.

After applying the conversion, the observations increased from 429 to 4,261. Of them, 3,409 (80%) were used to fit the model, and 852 (20%) were used to test the model and check for no overfitting. With this, it is possible to build a model which classifies the papers. The table 2 summarizes the Data that will was used in the work.

Table 3: Sample Data Summary.

	Ecuadorian Papers	**Corpus**	**ISIC4**	**SAMPLE DATA**
Observations	15,448	619	419	4,261
Porpouse	To classify	Model structure	Model structure	Model fitting

Figure 1 shows the distribution of the sample dat. It is expected that the output of the model follows that distribution. In the table

Table 4: Distribution of the Expected Classification

Group	**Amounts**
Equals to Zero	1676
Equals to 1	2493
Outliers (>1)	92

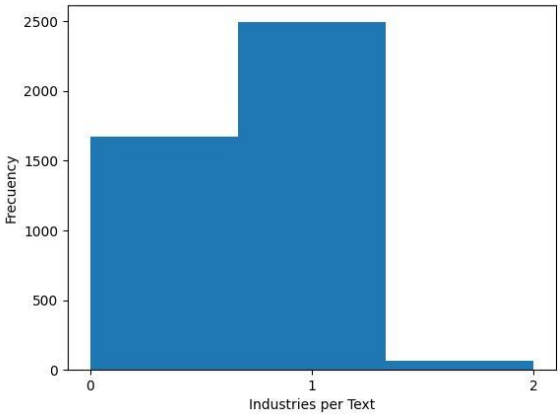


Figure 1: Frequency of Papers Classifications

CLASSIFYING THE PAPER TEXTS

A manual classification of the Ecuadorian paper's dataset would take a lot of time. Also, doing it manually implies classifying the papers according to my criteria. Classifying by using my criteria implies that I need to know every topic of science that has an implication in the industry that I want to classify the paper. This working paper applies a method that classifies the paper based on a literature review of papers that can be classified as the ISIC4 description.

The classification process will be based on the review of the word vectorizing (w2v) literature (Mikolov et al., 2013). The methodology assig a vector to the text of the input data, then it is possible to set weights from the word. The original literature searches to establish an output in text form, but this work is only centered on classification and not in generating a natural language output. The detailed explanation of the model used is in appendix B.

This approach is the best to do this classification process. Doing it manually requires too much time to complete the task and uses my criteria, however I'm not a person with the complete knowledge of the whole ISIC4. Using a set of people to classify the papers would be a good alternative. Then, the final classification is set based on the answers of the participants. However, this method needs a big number of people with the necessary time to answer all the papers. Using sets of representative words may avoid those words that are also related to the industry. So, for w2v being the best approach to solve the problem, it must solve the problem the human bias and the amount of participant in the model fitting.

W2v representation fits the model with a dataset of text that is already set as the ISIC4. Then, w2v is the best option because of its established associations based on vocabulary. The problem of the w2v is the dimensionality.

MODEL FITTING

Based on the appendix B, the model is and $\widehat{Y_{419}} = F_{419}(X_{617} * W_{617,419}) + \epsilon_{419}$in f represents a sigmoid function to get the final output vector in values between 0 and 1. The W matrix represents linear transformation. The linear transformation contains the biggest amount of params in the model. The extension of this model is due to the necessity to transform a vector input into another form of a different dimension.

Thanks to the little corpus length, the model was trained in less than a hour, and the coding required more time. The summarize are showing in the table 3:

Table 5: Results from the Model Fitting

	Input dimension	Output dimension	Params
Linea Layer	617	419	258,523
Sigmoid	419	419	419

	Observations	MSE (30 Epochs)
Fitting Sample	3,409	5.18
Testing Sample	852	1.38

And the model was fitting using the Gradient Descent methodology, with the Adam optimizer (Appendix B).

CLASSIFICATION OUTPUTS

The model output converged quickly to the sixth decimal near zero. It was expected because it would be weird to see papers related to many industries. The output was a similarity matrix optimized by SDG, as seen in the figure 1. Decimals are important because are in they where the cutoff can be found.

Table 6: Example of the Similarity Output

A0210	A0220	A0230	A0240
0.00010062262	0.00017768075	0.00014200524	0.000113689224
6.0282255e-09	2.3156158e-08	1.0108635e-08	8.663112e-09
1.0628098e-05	2.3125955e-05	1.5888789e-05	1.2057848e-05

0.00010062262	0.00017768075	0.00014200524	0.000113689224
5.304037e-05	9.8650344e-05	7.899223e-05	5.701993e-05

Given the sample data it is possible to show the next graph:

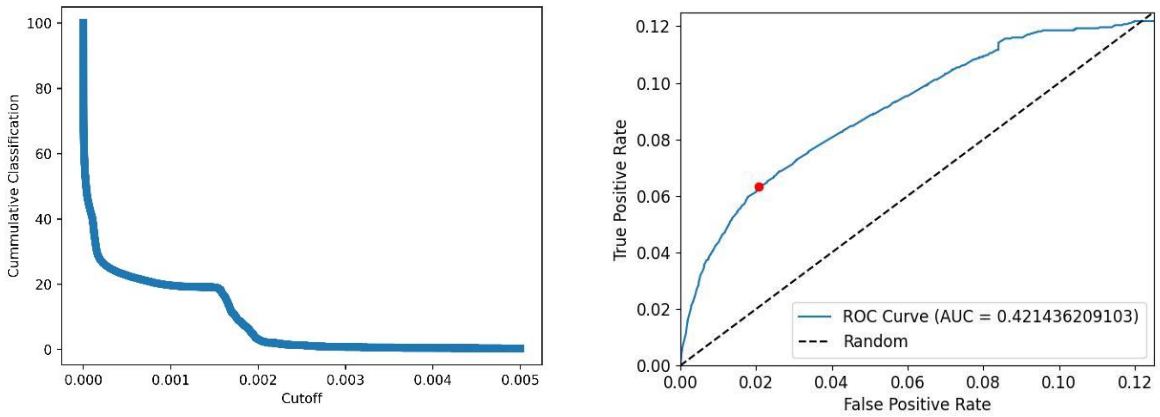


Figure 2: The cumulative classification, and the ROC curve on the left

Figure 2 shows that the number of observations with a similarity bigger to 0.003 is insufficient. This is why the ROC curve is in the range 0 to 0.12. However, it shows the traditional curve from the ROC curve. This work founded the best cutoff at 0.002.

CLASSIFYING THE ECUADORIAN DATA

After classifying the papers, the network was built by connecting those papers in one author was author in both papers. Even if this is not a transaction as the traditional complexity data, this work assumes that the authors share their knowledge when working on different projects. Then,

the research network can be built without classifying the papers, but having the network classified, the researcher can connect with the industrial network in future projects.

Table 7: Dataset Related to this Project.

Dataset	Description	Observations
Papers Texts	List the Ecuadorian papers published in Scopus from the last century to 2019. Is classified by the ISIC4 codes	15,448
Pairs Dataset	This dataset contains combine the information to papers that had been liked by a common author. For every product classified of every paper is a different observation. This is the dataset required to build the product space.	2,483,000
Dark Research Network	Is the research network that wasn't classified by the model. Those observations were not estimated because it required much more RAM memory consumption	Unknown due the technical limitations
The Product Space	Is the final output searched by this working paper. The product spaces can be used as tool fro prognosticating future sectors in the research or industry.	406

In the economic complexity literature, it is necessary to use the relative comparative advantage to first determine if the exporter country is a significative exporter (Hidalgo & Hausmann, 2009).

On this work that is not necessary. By construction, there is not an amount as it is in the trade data. Here the degree distribution of every paper is counted by the other papers that the author has in common, after cleaning eventual outliers. After that, it is possible to start the reflections method. So, from the formula proposed:

$\phi_{i,j,t} = min = \{P\left(RCA_{x_{ij}} \middle| RCA_{x_{jt}}\right), P\left(RCA_{j,t} \middle| RCA_{x_{ii'}}\right)\}$(Hidalgo et al, 2007), my modification is that as every researcher acts as a binary variable, the RCA will the products that shares to a same paper.

Figure 3 shows the frequencies of the share. The general shares count how much pairs a pair of paper and product (ISIC) share the knowledge. The Products sharing shows how many papers are linked to a product, and the paper shares count the number of products linked to it. The Ecuadorian papers shares does show a similar distribution as the expected. On the other hand, the product shares changed the magnitude. The General Share distribution shows a completely different distribution, maybe caused by the different dimensions of the separated papers and products components.

Figure 5 (appendix 3) shows the scatter plots for the first two reflections for the papers share and the product respectively. The paper graph shows the expected output based on the literature. However, the industrial graph shows an unexpected curve.

Figure 6 (appendix 3) shows the Product space vs the Euclidean distance between the products. It shows pretty like the outputs from the literature. The output from this work reduces the amount of data available during the reflections. However, the form is exactly the expected.

CONCLUSIONS

This work was the first step in a research agenda for the development of a tool to prognosticate the growth of economic sectors, starting from the academical sector. To fulfill those objectives, and include the academical sector, it was necessary to experiment with techniques that are not common in economics. At the end of the journey, there are two important considerations for future projects.

The first of them is related to the use of the techniques of artificial intelligence. Those tools take too much time to pass from a status of development to operations. It could be an improvement to start working with pretrained algorithms like ChatGPT or Bard. But there exists a scenario in which those tools were not successful, and the authors must build the tools for them. The technological capabilities can be a limit. Also, this work didn't use every recommended step from this the development area.

Thanks to the god results in the product space, this work opens a door to continue with this research line.

REFERENCES

Diederik , P., & Lei. (2015). DAM: A METHOD FOR STOCHASTIC OPTIMIZATION. International Conference on Learning Representations, 1–15. https://doi.org/10.48550/arXiv.1412.6980

Google. (2023, March 21). *About Keras*. KerasGitHub. Retrieved May 15, 2023, from https://github.com/keras-team/kerasHidalgo, & Hausmann. (2009, June 30). The building blocks of economic complexity. Proceedings of the National Academy of Sciences, 106, 10570–10575. https://doi.org/10.1073/pnas.0900943106

Hidalgo, C., & Hausmann, R. (2009, June 30). The building blocks of economics complexity. (P. Sarathi, Ed.) Proceedings of the National Academy of Sciences, 106(26), 10570-10575. doi:https://doi.org/10.1073/pnas.0900943106

Hidalgo, C. A., Klinger, B., Barabási, A. L., & Hausmann, R. (2007). The product space conditions the development of nations. Science, 317(5837), 482-487.

Mikolov, Chen, Corrado, & Dean. (2013, September 7). Efficient Estimation of Word Representations in Vector Space. In *Arxiv* (https://doi.org/10.48550/arXiv.1301.3781). Cornell University. Retrieved May 14, 2023, from https://arxiv.org/abs/1301.3781

Scopus. (n.d.). *Scopus Preview*. Retrieved May 15, 2023, from https://www.scopus.com/sources.uri?zone=TopNavBar&origin=

United Nations. (2008). Introduction to ISIC. United Nations Statistics Division. Retrieved May 14, 2023, from https://unstats.un.org/unsd/classifications/Econ/isic

United Nations. (2015). *Introduction to CPC*. United Nations Statistics Division. Retrieved May 15, 2023, from https://unstats.un.org/unsd/classifications/Econ/cpc

United States Census Bureau. (2019). *North American Industry Classification System - NAICS*. Retrieved May 15, 2023, from https://www.census.gov/naics/?48967

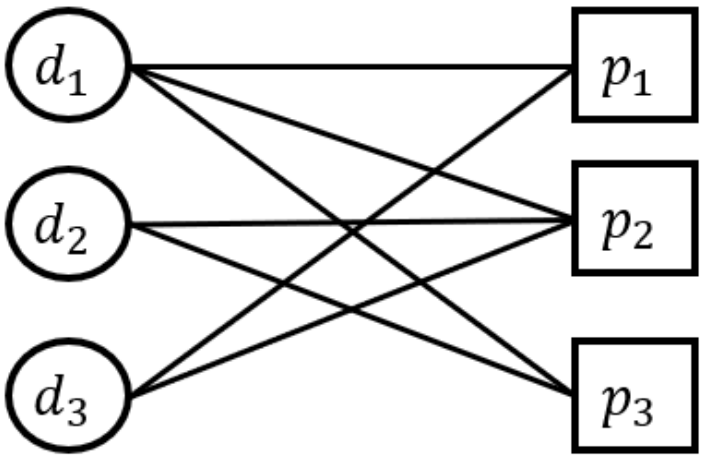
APENDIX A:

REFLECTIONS

As said before, measuring the diversification or specialization by counting the number of products related to a node is wrong. Given only that measure, how many products make a node significative diversified? The reflections work as a method to get the data normalized.

Whereas in the complexity literature, the reflections are used as k_{α_N}. Then, the first reflection is only the count of industries that are assigned to a paper where $\alpha = 1$, and the number of papers assigned to a industry where $\alpha = 0$.

For example. Assume a set of papers $\{d_1, d_2, d_3\}$. Also assume a set of industries or products $\{p_1, p_2, p_3\}$. We can relate them as follows:



Give, that, the initials reflections are $k_{10,1} = 3, k_{10,2} = 2, k_{10,3} = 2$ for the papers case. The values for the products are $k_{00,1} = 2, k_{00,2} = 3, k_{00,3} = 2$. To get the values of the reflections 1 and 2, it is necessary to weight based on the products related to a paper. The first reflection is the

average of the ubiquity of the products related to a peper, or the average of the diversification of the papers related to a product.

$$k_{11,1} = \frac{1}{3} * (2 + 3 + 2) = 2.33 \qquad k_{01,1} = \frac{1}{2} * (3 + 2) = 2.5$$

$$k_{11,2} = \frac{1}{2} * (3 + 2) = 2.5 \qquad k_{01,2} = \frac{1}{3} * (3 + 2 + 2) = 2.33$$

$$k_{11,3} = \frac{1}{2} * (2 + 3) = 2.5 \qquad k_{01,3} = \frac{1}{2} * (3 + 2) = 2.5$$

Finally, the second reflection is the average of the first reflection:

$$k_{12,1} = \frac{1}{2.33} * (2.5 + 2.33 + 2.5) = 3.15 \qquad k_{02,1} = \frac{1}{2.5} * (2.33 + 2.5) = 1.19$$

$$k_{12,2} = \frac{1}{2.5} * (2.33 + 2.5) = 1.19 \qquad k_{01,2} = \frac{1}{2.33} * (2.33 + 2.5 + 2.5) = 3.15$$

$$k_{12,3} = \frac{1}{2.5} * (2.5 + 2.33) = 1.19 \qquad k_{01,3} = \frac{1}{2.5} * (2.33 + 2.5) = 1.19$$

Given those reflections, we can write them with the formula:

$$k_{1N+1,i} = \frac{1}{r_{1N,i}} \Sigma r_{0N,j} \; ; \; k_{0N+1,i} = \frac{1}{k_{0N,i}} \Sigma k_{1N,j}$$

Where j identifies those nodes for which the node is related.

APENDIX B:

Words In Vector Representation

The data does not contain the information the set a direct relation with the paper and some industry from the ISIC4. However, it contains text from the title, keywords, and indexed keywords. With those strings of characters, (or strings as is said in the programming language) it is possible to set a vocabulary of words and use it to classify the papers.

Assuming the strings set is cleaned, the researcher must build a set of unique words to accept as vocabulary. $\{word_0, word_1, word_2, \dots, word_n\}$. For this example, I'll use:

$$\{ argentina, france, dragon, seed, ball, apple, water, \dots \}$$

Now, for every input it is necessary to create a vector of zero values with the length of the vocabulary. After that, the values will be changed to one if the words are in the string input, depending on the words index in the vocabulary. Given the string: "*After the last penalty, Argentina won, and France goes second*", this work set the string to lowercase, then keep only those words in the vocabulary, and finally transform to one hot encoding:

[After the last penalty, Argentina won, and France goes second]

[after the last penalty, argentina won, and france goes second]

[argentina france]

[1 1 0 0 0 0 0 …]

Given this transformation, it is possible to build a model that helps the classification of the papers into ISIC4 codes. The first step (layer) is to transform the dimensionality of the input to the ISIC4 length. I want to set an output in which the values are one when the paper input can be classified as the industry, and zero when not. So, the second layer of the model uses a sigmoid function to normalize the values between zero and one.

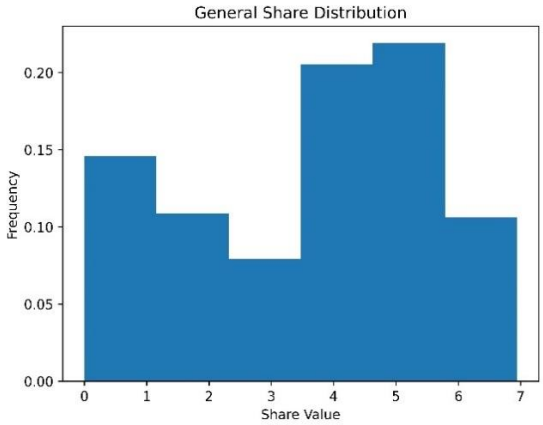
$$X_{1n} * W_{nm} = Z_{1m}$$

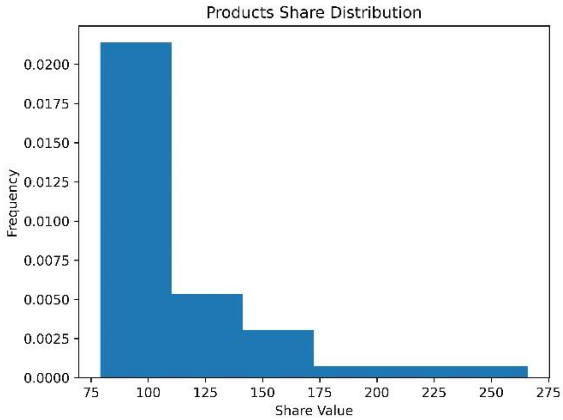
$$\widehat{Y_{1m}} = f(Z_{1m}) = f(X_{1n} * W_{nm})$$

This shows an important input for the model, because based on the vocabulary, the dimension of the input will be big. This work will train the model by using the ADAM optimizer, because it is good with large-dimensional parameters spaces (Diederik & Lei, 2015). This optimizer is based on a stochastic gradient descent optimizer. In other words, the model fitting process minimize the error between the prediction and the validation data by changing the weights of the layers.

This step is made in Python3.10, using the library Keras library which is based on the TensorFlow (Google, 2023). The library only requires setting a first layer with an input using the dimension of the vocabulary, and output with the dimension of the ISIC4, using a linear activation. Then, the second layer uses an input of the ISIC4 length (same output dimension) and uses a sigmoid activation function.

APENDIX 3: Outputs





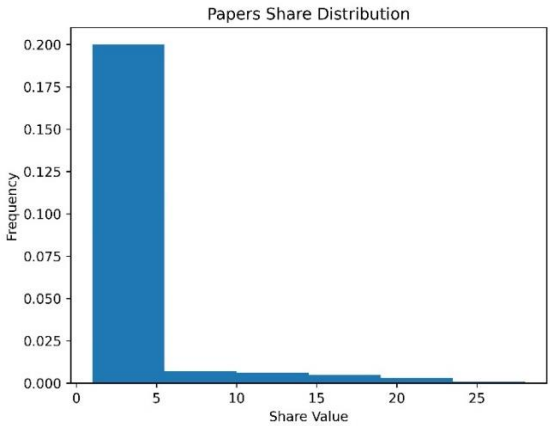


Figure 3: Frequency of Share

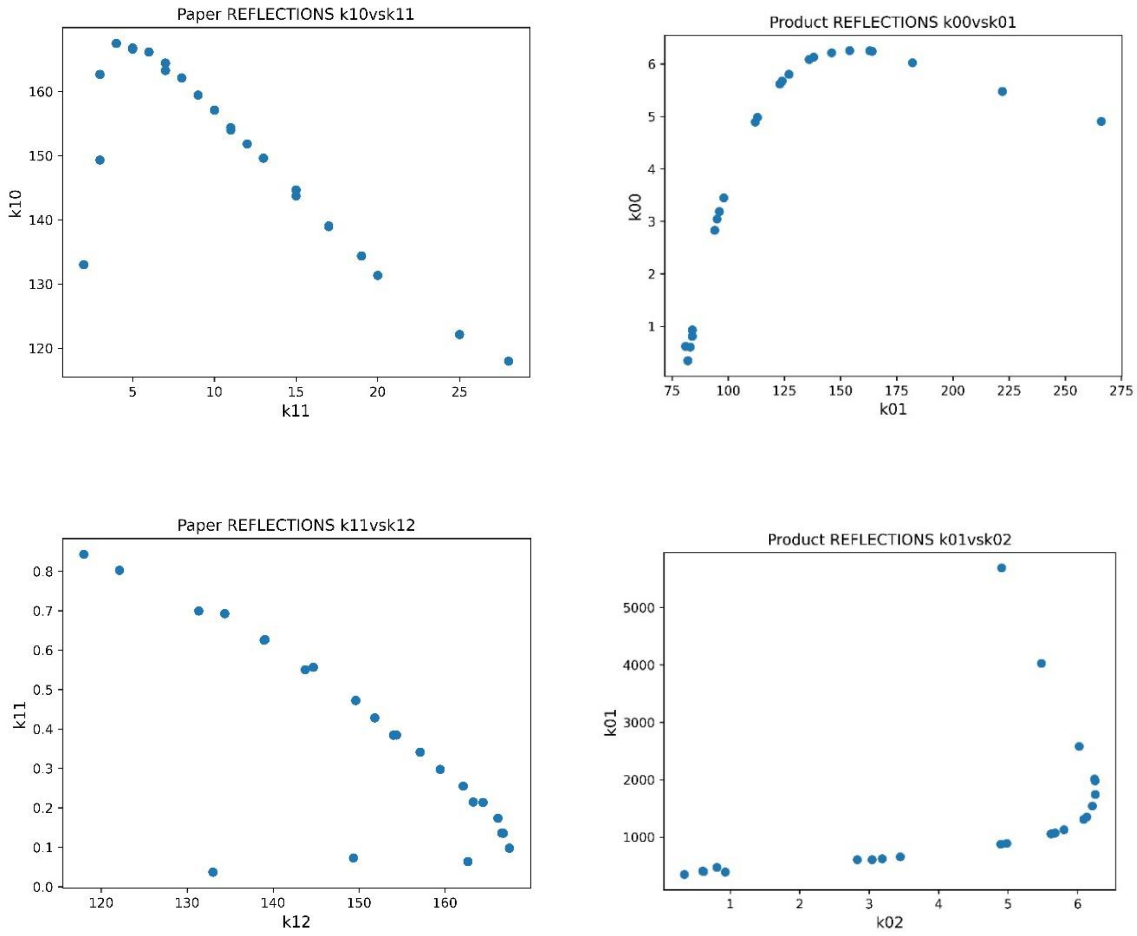


Figure 4: Reflections

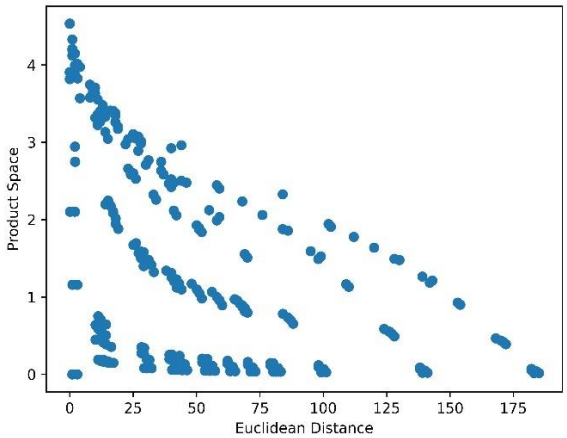


Figure 5: Product Space for Products Pairs