IOSR Journal of Multidisciplinary Research (IOSR -JMR)
ISSN: 3048-7439 Volume 2, Issue 3, Series 1, May. - June. (2025), PP 47-53
www.iosrjournals.org

Knowledge Creation From Multivariate Domain Data: An Example In Healthcare

Bel G Raggad,
Seidenberg School, Pace U, NY

Sarah Amel Raggad,
Dyson College, Pace U, NY

Abstract
With digital transformation present in most domains of healthcare, we started having an abundance of data wherever we are in healthcare. Furthermore, data keeps changing very rapidly and relying on statistical analysis to create healthcare decision support has become a very challenging task. There is great risk that the hypotheses selected to be tested based on available data would not provide the sought decision support.

We propose an analytical model, limited for categorical data, using evidential reasoning, to create knowledge in a specific domain of healthcare based on available domain data. While only a small dataset on student anxiety obtained from data.gov is used to demonstrate the working of the proposed model, our model also works for larger real-world healthcare data sets even though the model computations become heavier, and a computer is needed to perform them.

Date of Submission: 05-05-2025 Date of Acceptance: 15-05-2025

I. Introduction

Even though the field of information systems has expanded beyond our ability to regulate it, we still can study it in terms of three main conceptual resources: data, information, and knowledge [4]. The literature also organizes information systems into about a dozen computer-based information systems [3].

There is no doubt that data and information play a crucial role in problem solving and decision-making and they are indispensable as they provide the factual basis needed to make informed, rational, and effective choices. Data and information have the power to reduce uncertainty, support strategic planning, and add precision. Data is useful to refine situations by offering objective evidence that clears much ambiguity and reduces guesswork. Information adds accuracy and precision after contextualized and raw data are turned into meaningful interpretation needed for decision support. Management at the strategic level use data trends, forecasts, discovered new knowledge, and data analytics to define clear goals and develop long-term strategies to assure business continuity.

Information and knowledge are crucial in the health domain because they directly impact on the quality of care, patient outcomes, operational efficiency, and even public health policies [1, 2]. They are vital in medical support including for better diagnosis and treatment, for informed decision making, for improved patient care and safety, and for public health and policy. Clinicians can improve the accuracy of their diagnosis by analyzing patient records, lab results, and imaging information. Clinical expertise and medical research can guide the choice of effective treatments and interventions.

Knowledge creation model

Our data-based knowledge creation model (KCM) consists of 5 steps as follows:

1. Data selection
2. Generation of the feasible knowledge space
3. Computation of data support
4. Computation of certainty factors
5. Selection of highly supported hypertuples and rewriting them in natural language.

Our KCM starts with a data selection step that selects the variables/attributes of interest. This model is limited to categorical single-valued data. A second step will generate the feasible space of knowledge, made of all possible hypertuples induced from the selected variables. In a third step, every hypertuple in the feasible space of knowledge is validated by computing the extent to which it is supported by all available data in the selected dataset. The fourth step will compute a certainly factor for each of the hypertuples. In the fifth step, we select all those hypertuples with a higher certainty factor and rewrite them into a natural language.

Data selection

Most often, in any setting, organizations collect data in different ways for different purposes. Independently of those purposes and collection methods, however, knowledge support is always a competitive advantage that organizations considerably invest to sustain. Among available data, we select the data that is most relevant to the knowledge domain of interest. In this study, as shown in Figure 1, we are only considering categorical variables to process in the proposed knowledge creation model. Data types of ordinal and interval variables will be treated in a different study.

Once the dataset is selected, we only select those variables of interest that we will adopt in the knowledge creation process. Let us call the dataset D and select N variables A_1, A_2, …, A_N. We can denote D={A_i, i=1,N}. The variables contain single-valued categorical variables.

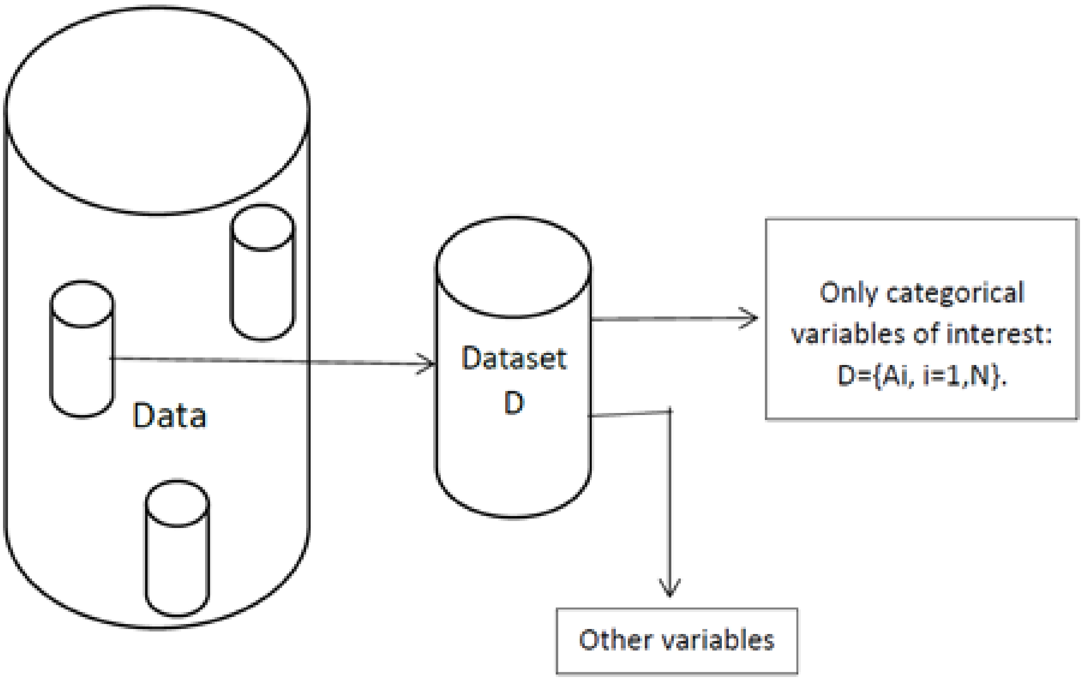


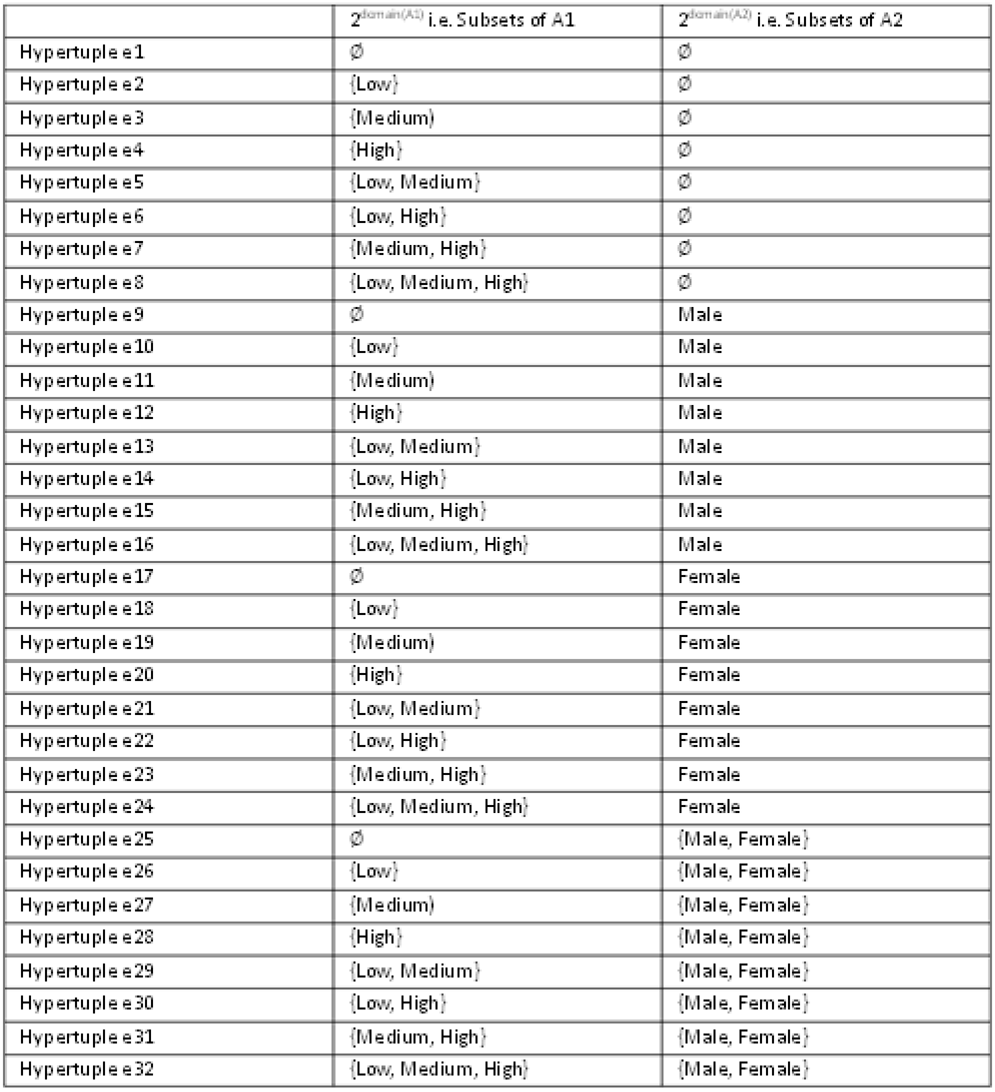
Figure 1: Data selection process

Generation of the feasible knowledge space

In the first step earlier, we selected our dataset of interest D={A_i, i=1,N} that we will process here to create the space of feasible knowledge K. This space of feasible knowledge will consist of all possible combinations of all subsets of the variables. Here is an example: Let A_1 take its values in its domain domain(A_1) ={Low, Medium, High}, and A_2 take its values in its domain domain(A_2) = {Male, Female}. A1 has 2^3=8 subsets, and A2 has 2^2=4 subsets. The set of subsets is called the power set and is denoted as a power of 2; for example, the set of subsets of the domain of A_1 is denoted $2^{\text{domain(A1)}}$.

The subsets of A_1 denoted $2^{\text{domain(A1)}}$ are the 8 subsets Ø, {Low}, {Medium), {High}, {Low, Medium}, {Low, High}, {Medium, High}, and {Low, Medium, High} and the subsets of A2 denoted $2^{\text{domain(A2)}}$ are the 4 subsets Ø, {Male}, {Female}, and {Male, Female}. The feasible space of knowledge generated by the variables A_1 and A_2 are made of all possible products of subsets of A_1 and subsets of A_2. This product is shown in Table 1.

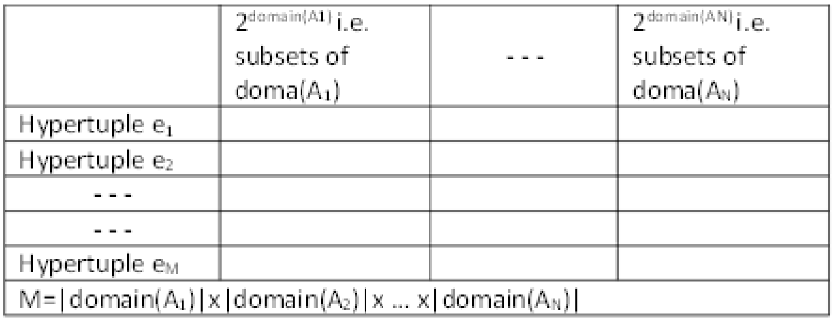
Table 1: Generation of the feasible knowledge space: Case 3x2



	$2^{domain(A1)}$ i.e. Subsets of A1	$2^{domain(A2)}$ i.e. Subsets of A2
Hypertuple e1	∅	∅
Hypertuple e2	{Low}	∅
Hypertuple e3	{Medium)	∅
Hypertuple e4	{High}	∅
Hypertuple e5	{Low, Medium}	∅
Hypertuple e6	{Low, High}	∅
Hypertuple e7	{Medium, High}	∅
Hypertuple e8	{Low, Medium, High}	∅
Hypertuple e9	∅	Male
Hypertuple e10	{Low}	Male
Hypertuple e11	{Medium)	Male
Hypertuple e12	{High}	Male
Hypertuple e13	{Low, Medium}	Male
Hypertuple e14	{Low, High}	Male
Hypertuple e15	{Medium, High}	Male
Hypertuple e16	{Low, Medium, High}	Male
Hypertuple e17	∅	Female
Hypertuple e18	{Low}	Female
Hypertuple e19	{Medium)	Female
Hypertuple e20	{High}	Female
Hypertuple e21	{Low, Medium}	Female
Hypertuple e22	{Low, High}	Female
Hypertuple e23	{Medium, High}	Female
Hypertuple e24	{Low, Medium, High}	Female
Hypertuple e25	∅	{Male, Female}
Hypertuple e26	{Low}	{Male, Female}
Hypertuple e27	{Medium)	{Male, Female}
Hypertuple e28	{High}	{Male, Female}
Hypertuple e29	{Low, Medium}	{Male, Female}
Hypertuple e30	{Low, High}	{Male, Female}
Hypertuple e31	{Medium, High}	{Male, Female}
Hypertuple e32	{Low, Medium, High}	{Male, Female}

In a general setting, the space of knowledge generated by the attributes A_1, …, A_N is made of all the products of subsets A_1, …, and subsets of A_N. The rows of the table containing the products of subsets are called hypertuples instead of tuple because the data elements are subsets and not single values as in the dataset D. The knowledge space table, depicted in Table 2, will contain $|domain(A_1)|x|domain(A_2)|x \dots x|domain(A_N)|$ hypertuples.

Table 2: General feasible space of knowledge



	$2^{domain(A1)}$ i.e. subsets of doma(A_1)	- - -	$2^{domain(AN)}$ i.e. subsets of doma(A_N)
Hypertuple e_1			
Hypertuple e_2			
- - -			
- - -			
Hypertuple e_M			
$M=\|domain(A_1)\|x\|domain(A_2)\|x \dots x\|domain(A_N)\|$			

Computation of data support

We will take the hypertuples one by one and run them through the entire data set D to see what rows in the dataset support the hypertuple. Figure 2 depicts the working of the data support process.

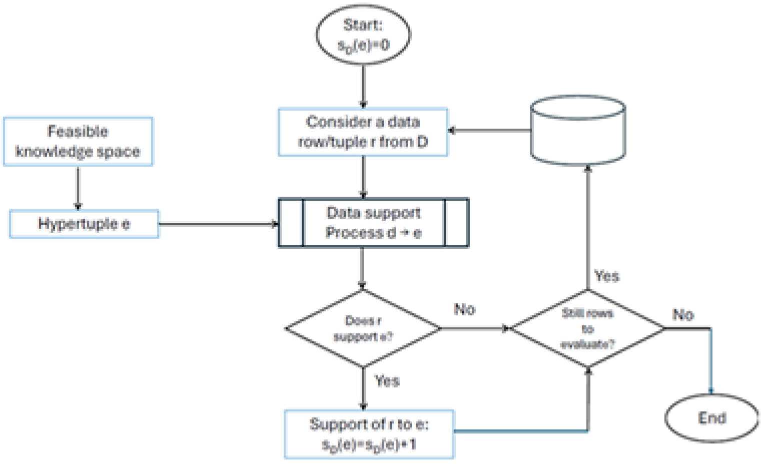


Figure 2: Data support process

A data tuple r is said to provide evidence support to a knowledge hypertuple e if and only if r belongs to e. Also, a dataset D is said to provide partial evidence support to a knowledge hypertuple e if and only if at least one tuple of the dataset provides evidence support to the knowledge hypertuple. Let us denote the set of rows in D that support to the hypertuple e as $P_D(e)$. The data support that a tuple r in the dataset D gives to a hypertuple e in the feasible space of knowledge K is a function S_D from D to the set of real numbers that returns S_D (e) = $|P_D(e)|$.

Computation of certainty factors

Given the information on hand so far, we can, as in [5], compute the basic probability assignment [5, 6] on the space K. For any hypertuple e, the probability mass on e is given by $m(e) = s_D(e)/\sum_{x \text{ in } K} s_D(x)$. The certainty factor of a hypertuple is computed as $m(e)/ \sum_{i=1,N} |ei|$.

When the probability mass on a hypertuple e is computed, all those hypertuples that cover e will enjoy a higher mass and they will also enjoy a higher belief value. But the higher we go in covering e the more precision and specificity we lose, because when the size of the subset increases we cannot determine specifically which elements in it gets the probability mass. That said, our objective should not only be aiming to have a higher probability mass but also a more precise hypertuple. The precision is higher when the size of the hypertuple is smaller. The certainty factor should therefore take into account of a higher probability mass and a smaller size, at the same time, as proposed above.

Selection of highly-supported hypertuples and rewriting them in a natural language

Once all the masses of hypertuples are computed and also their certainty factors, we can now rank the hypertuples in a descending order of the certainty factor. This will take care of the balance between higher mass values and a higher precision. We can truncate at an arbitrary rank c_0 judged giving sufficient evidence support and only accept the hypertuples with certainty factors higher than c_0.

Case of Knowledge creation from Anxiety data

We borrowed data on student anxiety from data.gov. A dataset of 48 rows of data is used to demonstrate the working of our proposed data-based knowledge creation model. The selected variables are the following:
A1: Gender: Domain(A1) = {M, F}
A2: GPA: Domain(A2) = {1, 2, 3}
A3: History of Anxiety: Domain(A3) = {Y, N}
A4: Smoker: Domain(A4) = {Y, N}
A5: Stress Lavel: Domain(A5) = {L, M, H}

The space of feasible knowledge has |K| hypertuples where

|K| = $2^{|\text{domain(A1)}|}$ x $2^{|\text{domain(A2)}|}$ x $2^{|\text{dmain(A3)}|}$ x $2^{|\text{domain(A4)}|}$ x $2^{|\text{domain(A5)}|}$
= 2^2 x 2^3 x 2^2 x 2^2 x 2^3
= 4 x 8 x 4 x 4 x 8
= 4096 hypertuples.

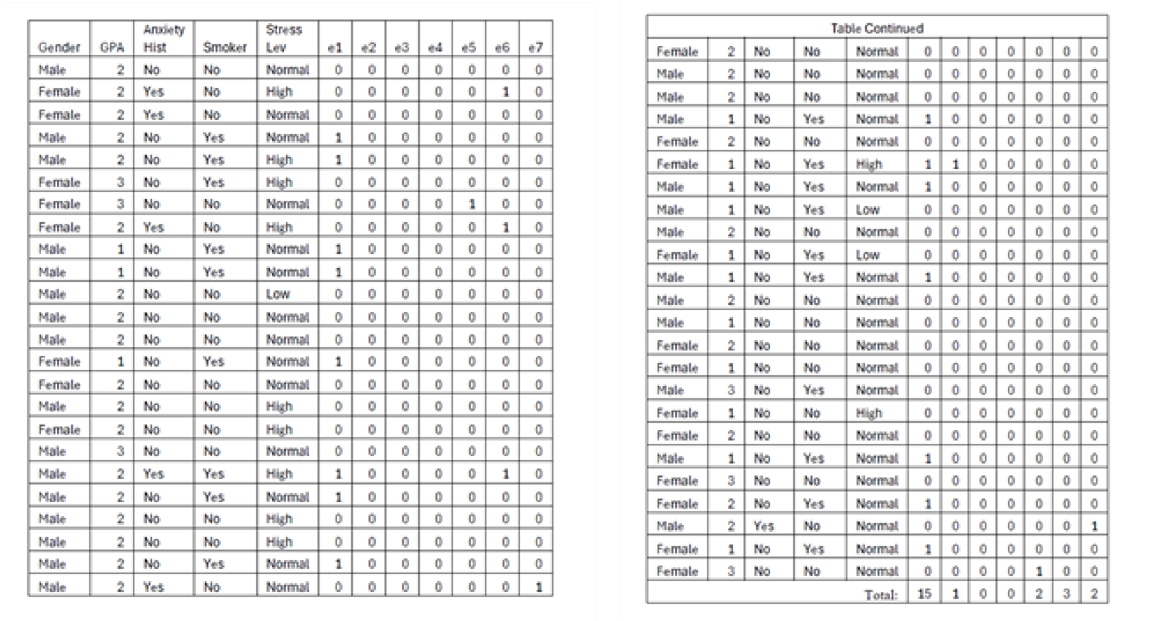
As you can see above there is an exponential expansion of hypertuples when the number of variables gets higher and when the size of the domains of variables gets higher. Most often, however, the human being can only visually and cognitively examine pieces of knowledge with a limited number of variables and with limited values taken by these variables. Computation, on the other hand, is not a problem because mathematical approximation can be applied to assure a fast and cost effective computation [8].

We are herein limiting ourselves to the evaluation of a small number of hypertuples for which we will seek data support:

e1 = ({M, F}, {1, 2}, {Y, N}, {Y}, {M, H})
e2 = ({F}, {1, 2}, {N}, {Y}, {M})
e3 = ({F}, {3}, {N}, {N}, {L})
e4 = ({M}, {3}, {N}, {N}, {H})
e5 = ({F}, {3}, {N}, {N}, {L, M, H})
e6 = ({M, F}, {1, 2}, {Y}, {Y, N}, {H})
e7 = ({M}, {1, 2}, {Y}, {Y, N}, {L, M})

The computation of data support to the above hypertuples is are provided in Table 3. One can see that, except for e_1, only a few rows of the dataset provided evidence support for the proposed hypertuples. For the hypertuple e_1, there are 15 tuples/rows from the dataset D that provided evidence support.

Table 3: Data support process

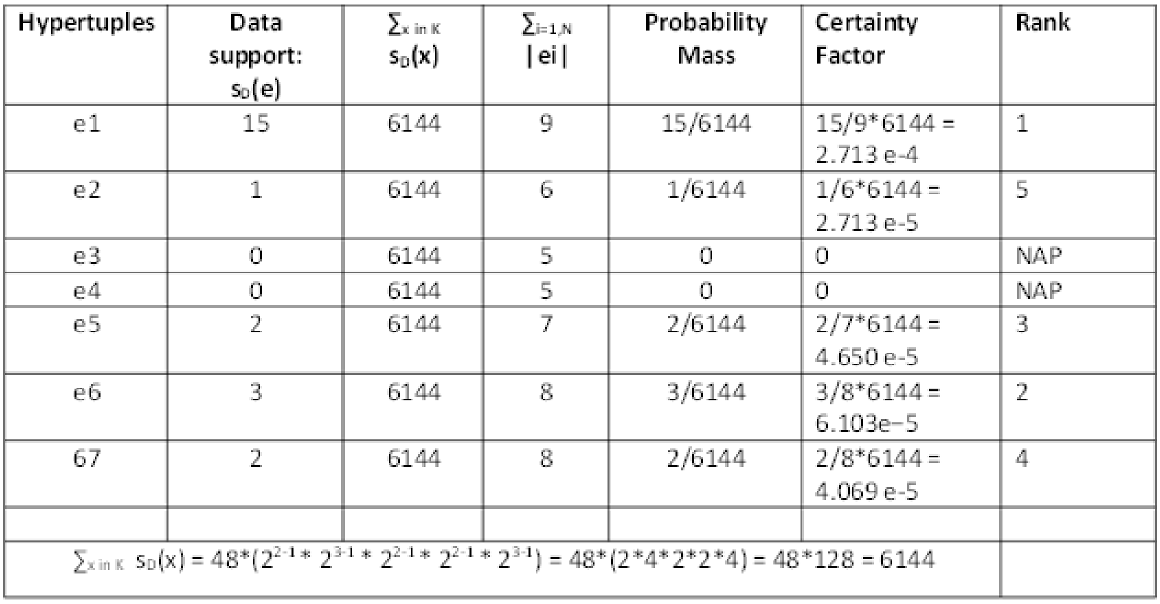


Gender	GPA	Anxiety Hist	Smoker	Stress Lev	e1	e2	e3	e4	e5	e6	e7
Male	2	No	No	Normal	0	0	0	0	0	0	0
Female	2	Yes	No	High	0	0	0	0	0	1	0
Female	2	Yes	No	Normal	0	0	0	0	0	0	0
Male	2	No	Yes	Normal	1	0	0	0	0	0	0
Male	2	No	Yes	High	1	0	0	0	0	0	0
Female	3	No	Yes	High	0	0	0	0	0	0	0
Female	3	No	No	Normal	0	0	0	0	1	0	0
Female	2	Yes	No	High	0	0	0	0	0	1	0
Male	1	No	Yes	Normal	1	0	0	0	0	0	0
Male	1	No	Yes	Normal	1	0	0	0	0	0	0
Male	2	No	No	Low	0	0	0	0	0	0	0
Male	2	No	No	Normal	0	0	0	0	0	0	0
Male	2	No	No	Normal	0	0	0	0	0	0	0
Female	1	No	Yes	Normal	1	0	0	0	0	0	0
Female	2	No	No	Normal	0	0	0	0	0	0	0
Male	2	No	No	High	0	0	0	0	0	0	0
Female	2	No	No	High	0	0	0	0	0	0	0
Male	3	No	No	Normal	0	0	0	0	0	0	0
Male	2	Yes	Yes	High	1	0	0	0	0	1	0
Male	2	No	Yes	Normal	1	0	0	0	0	0	0
Male	2	No	No	High	0	0	0	0	0	0	0
Male	2	No	No	High	0	0	0	0	0	0	0
Male	2	No	Yes	Normal	1	0	0	0	0	0	0
Male	2	Yes	No	Normal	0	0	0	0	0	0	1

Table Continued											
Female	2	No	No	Normal	0	0	0	0	0	0	0
Male	2	No	No	Normal	0	0	0	0	0	0	0
Male	2	No	No	Normal	0	0	0	0	0	0	0
Male	1	No	Yes	Normal	1	0	0	0	0	0	0
Female	2	No	No	Normal	0	0	0	0	0	0	0
Female	1	No	Yes	High	1	1	0	0	0	0	0
Male	1	No	Yes	Normal	1	0	0	0	0	0	0
Male	1	No	Yes	Low	0	0	0	0	0	0	0
Male	2	No	No	Normal	0	0	0	0	0	0	0
Female	1	No	Yes	Low	0	0	0	0	0	0	0
Male	1	No	Yes	Normal	1	0	0	0	0	0	0
Male	2	No	No	Normal	0	0	0	0	0	0	0
Male	1	No	No	Normal	0	0	0	0	0	0	0
Female	2	No	No	Normal	0	0	0	0	0	0	0
Female	1	No	No	Normal	0	0	0	0	0	0	0
Male	3	No	Yes	Normal	0	0	0	0	0	0	0
Female	1	No	No	High	0	0	0	0	0	0	0
Female	2	No	No	Normal	0	0	0	0	0	0	0
Male	1	No	Yes	Normal	1	0	0	0	0	0	0
Female	3	No	No	Normal	0	0	0	0	0	0	0
Female	2	No	Yes	Normal	1	0	0	0	0	0	0
Male	2	Yes	No	Normal	0	0	0	0	0	0	1
Female	1	No	Yes	Normal	1	0	0	0	0	0	0
Female	3	No	No	Normal	0	0	0	0	1	0	0
				Total:	15	1	0	0	2	3	2

As indicated above, there are 4096 hypertuples to validate by computing their data support from the subset D. We certainly cannot do this manually and a computer program is needed for that. But let’s compute the data support provided by the subset D to the 7 hypertuples we arbitrarily select for demonstration purposes. After applying the formulas above, we compute the certainty factors as shown in Table 4.:

Table 4: Computation of certainty



Hypertuples	Data support: $s_D(e)$	$\sum_{x \text{ in } K}$ $s_D(x)$	$\sum_{i=1,N}$ $\lvert ei \rvert$	Probability Mass	Certainty Factor	Rank
e1	15	6144	9	15/6144	15/9*6144 = 2.713 e-4	1
e2	1	6144	6	1/6144	1/6*6144 = 2.713 e-5	5
e3	0	6144	5	0	0	NAP
e4	0	6144	5	0	0	NAP
e5	2	6144	7	2/6144	2/7*6144 = 4.650 e-5	3
e6	3	6144	8	3/6144	3/8*6144 = 6.103e–5	2
67	2	6144	8	2/6144	2/8*6144 = 4.069 e-5	4
$\sum_{x \text{ in } K}$ $s_D(x) = 48*(2^{2-1} * 2^{3-1} * 2^{2-1} * 2^{2-1} * 2^{3-1}) = 48*(2*4*2*2*4) = 48*128 = 6144$						

Our knowledge creation model puts more confidence in the hypertuples that have higher confidence factors as shown in Table ? above. Hypertuple e1 is ranked first followed by e6, e5, e7, and e2, in this order. Hypertuples e3 and e4 have no ranks because they are not supported by the dataset D.

The last step is the accepting of ranked hypertuples and their rewriting into a natural language. We obtain the following knowledge pieces:
Rank 1: e1 = “Everybody with low gpa who smokes has some stress.”
Rank 2: e6 = “Students with an anxiety history are getting a high stress level.”
Rank 3: e5 = “Smoking female students are not getting high gpa.”
Rank 4: e7 = “Male students with a history of anxiety do not show a high stress level.”
Rank 5: e2 = “Smoking female students with a low gpa, even though with no anxiety history, still get high stress conditions.”

Managerial implications

While the example used above to demonstrate the working of our proposed model is applied to healthcare data, this model still applies to any other domain where knowledge is valued in a managerial decision process. The created knowledge may be very useful in understanding available data and reduce a great deal of the uncertainty tainting the problem-solving effort. The creation of knowledge is also an important step to conduct before starting a more serious statistical analysis and in discussing obtained findings.

II. Conclusion

The paper proposed an analytical model, using evidential reasoning, to create knowledge in a specific domain based on available domain data. While the literature proposed ways to treat ordinal and interval data, the proposed model only treated categorical data. A small dataset on student anxiety was obtained from data.gov and used to demonstrate the working of the proposed model. With a larger real-world data set, the same computations apply but a computer program is needed to complete all required computations to produce the certainty factors of the feasible knowledge hypertuples and their rewriting into a natural language.

References

[1] Bossen Et Al. (2019), Data Work In Healthcare: An Introduction Health Informatics Journal, SAGE Publications Inc, 25(3):465-474.

[2] Dalrymple, P.W., (2018), Data, Information, Knowledge: The Emerging Field Of Health Informatics, Bulletin Of The American Society For Information Science And Technology, 37(5):41-44

[3] Raggad, B.G. (1997), Information Systems Concepts: A Guide For Executives, Logistics Information Management, 10(4), 146-153.

[4] Schell, G. And R. Mcleod, (2010), Management Information Systems, 10th Edition, Prentice Hall.

[5] Shafer, G. A Mathematical Theory Of Evidence, Princeton Univ. Press, Princeton, NJ, 1976. Shell G. And R. Mcleod (2010), Management Information Systems: 10th Edition, Mcgraw Hill.

[6] Smets, P. And R. Kennes, The Transferable Belief Model, Artificial Intelligence, 1994;66,191-234.

[7] Wang, H., Liu, J. And J.C. Augusto, (2010), Mass Function Derivation And Combination In Multivariate Data Space, Information Sciences, 180(6), 813-819.

[8] Wang, H. And S. Mcclean, (2008) Deriving Evidence Theoretical Functions In Multivariate Data Space, IEEE Transactions On Systems, Man, And Cybernetics, 38(2), 455-465.