Hospital Recommendation System

Silvanus David, *140953350* and Pratyay Amrit, *140953430*

***Abstract*—Recommendation systems have been flooding the IT industry ever since the concept of information retreival came into existence. Every web application uses one of these recommendation systems, which helps users of said applications to get personalized suggestions on the basis of their past records. The Hospital Recommendation System aims to use this concept for hospitals to get an idea of where to set up health camps around the area. A hospital maintains a record of every patient that visits the hospital with an issue. This record includes details such as disease, symptoms, addresses etc. If this data can be classified, on the basis of the patterns mined, it becomes easy for the hospital to understand which locality is prone to which disease and will allow these hospitals to set up health camps with focus on these diseases, thus eliminating diseases before they become an issue. With the help of doctors, this data can be used to inform the concerned authorities of the reasons why the area is prone to the particular disease. For example, a leaked drainage can invite disease causing insects in the area. If this data is provided and wisely used, health problems can be greatly reduced. This paper aims to find ways to obtain this data.**

***Keywords*—Recommendation Systems, Information Retreival, Medical Recommendation, Hospital Recommendation**

I. Introduction

The outburst of population in recent years have brought all institutions under great pressure. Hospitals, in particular, are flooded with patients everyday. Even then, to the rural citizens, accessibility is, and has always been a major issue. To counter this, hospitals have been periodically setting up health camps for free health checkups and to raise general awareness. It is often noticed that, these health camps diagnose similar kind of issues in their respective regions. Using data mining and classification technologies, we can identify these medical hotspots prior to visiting them. Efficient extraction of valuable information from hospitals can be put to greater use, attacking diseases from many more directions. This can be achieved in many ways, depending on the type of data available in the hospitals. It is interesting to note that some of these goals require nothing more than a better visualization of data. We have come up with the following cases where the hospital would benefit from our efforts:

1. To classify places as prone to diseases and set up health camps.
2. To identify which disease is more common in these prone locations to provide facilities more focussed to the cause.
3. To identify the cause of these common diseases and report the same to concerned authorities.
4. To classify a new patient coming from an identified area as more susceptible to the common disease and have him/her diagnosed for the same.

These cases are but a few ways this system can help the hospitals by providing a recommendation system.

A. Data Set

For the purpose of demonstration of the usability of these systems, we will use a data set acquired from Data World[1]. This dataset is for death of California residents, by cause of death, by Zip Code, for the years 1999-2013. The death counts per cause of death, represents the deaths to California residents. A non-California resident who dies inside of California is not included in the dataset.

year	zip	cod	count	x	y
1999	90009	LIV	1	33.945	-118.383
1999	90068	DIA	1	34.116	-118.330
1999	90704	INJ	1	33.332	-118.344
1999	90813	ALZ	1	33.782	-118.183
1999	91024	INJ	1	34.165	-118.052
1999	91202	HOM	1	34.165	-118.266
1999	91301	PNF	1	34.157	-118.757
1999	91354	ALZ	1	34.447	-118.537
1999	91377	PNF	1	34.136	-118.774
1999	91765	ALZ	1	34.007	-117.810
1999	91905	LIV	1	32.672	-116.320
1999	92041	PNF	1	?	?
1999	92242	ALZ	1	34.149	-114.339
1999	92251	DIA	1	32.847	-115.573
1999	92257	LIV	1	33.378	-115.696
1999	92259	STK	1	32.739	-115.993
1999	92270	HOM	1	33.764	-116.422

Fig. 1. Sample data set. The 'x' and 'y' attributes signify the geographical coordinates of the location of the zip code. The data set is a total of 10000 entries

Cause-of-death were coded using the Tenth Revision of the International Classification of Diseases codes (ICD-10)
HTD Diseases of the Heart
CAN Malignant Neoplasms (Cancers)
STK Cerebrovascular Disease (Stroke)
CLD Chronic Lower Respiratory Disease (CLRD)
INJ Unintentional Injuries
PNF Pneumonia and Influenza
DIA Diabetes Mellitus
ALZ Alzheimer's Disease
LIV Chronic Liver Disease and Cirrhosis
SUI Intentional Self Harm (Suicide)

HYP Essential Hypertension and Hypertensive Renal Disease
HOM Homicide
NEP Nephritis, Nephrotic Syndrome and Nephrosis
OTH All Other Causes of Death

B. Methodology

For each of the cases mentioned earlier, we will be suggesting ways to achieve the goal, and apply them on the data set. Some of them just involve a better visualization of the dataset. We will be comparing the different ways used in each of these methods. For example, for the first case, to identify the places prone to diseases in general would require a simple place vs disease graph. However, clustering the data would provide a better idea of the geographical boundaries prone to the diseases. Pros and cons will be discussed in each of the cases in their respective sections. To generate results we will be using RapidMiner Studio for the purpose of demonstration. If a hospital recommendation system is to be developed, it can be developed using any of the programming technologies that the developer feels comfortable in.

II. Hospital Recommendation Models

We will be considering the following cases as models for designing a hospital recommendation system:

1. To classify places as prone to diseases and set up health camps.
2. To identify which disease is more common in these prone locations to provide facilities more focussed to the cause.
3. To identify the cause of these common diseases and report the same to concerned authorities.
4. To classify a new patient coming from an identified area as more susceptible to the common disease and have him/her diagnosed for the same.

These models, however, should not be confused as independent methods as they are all interrelated.

A. To classify places as prone to diseases and set up health camps

The simplest method would be to group the zip codes in our data set and plot it against the number of deaths and get a clear picture of where the maximum casualties occured. Three peaks can be seen, and we get an idea that there is something abnormal about these three areas. (Fig. 2)

A problem with this method is that, there is no way of knowing the trend around these peak areas, in case a certain disease has affected the regions around it and is spreading.

Another method is to plot the entire on the basis of its coordinates and look at the count in the third axis or using color codes. (Fig. 3)

This will give a better idea in the sense that it will represent the number of casualties on a geographical map, which would help in identifying spreading epidemics, and plan a better setup for health camps.

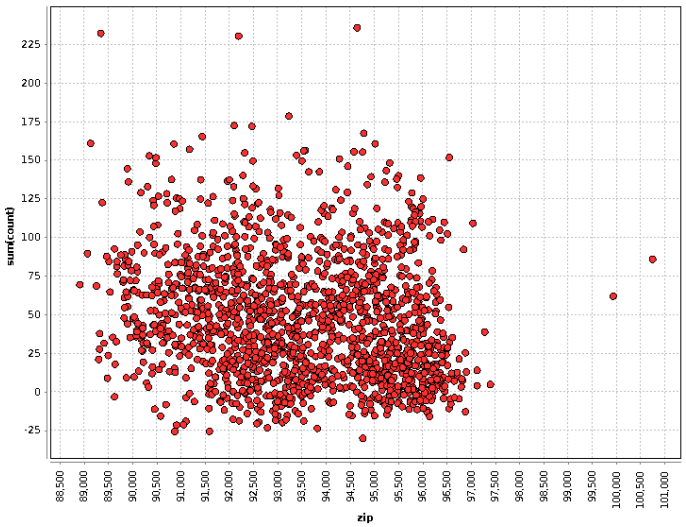

Fig. 2. Zip as x-axis and sum(count) as y-axis. Three peaks can be seen.

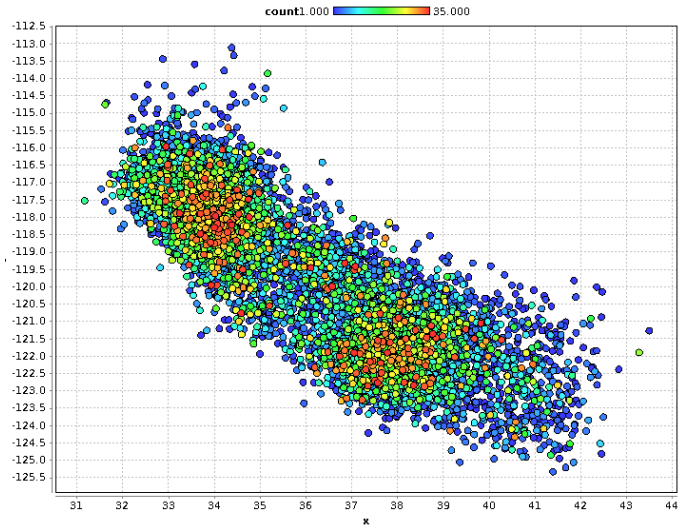

Fig. 3. 'x' vs 'y', color coded count. Redder is higher. Regions around the red areas are yellow, meaning that areas around those peaks are affected as well.

B. To identify which disease is more common in these prone locations

If we follow the first method as proposed in finding the plases more prone to diseases in general, we will need to plot the zip against the causes of death. The result, however, will be of no use towards achieving our goal. (Fig. 4)

With the second method, however, if we map x and y to the plot and color code with causes of disease, the results come out to be quite interesting. (Fig. 5)

By combining information obtained from Fig. 3 and Fig. 5, we can infer that the two locations having most casualties are

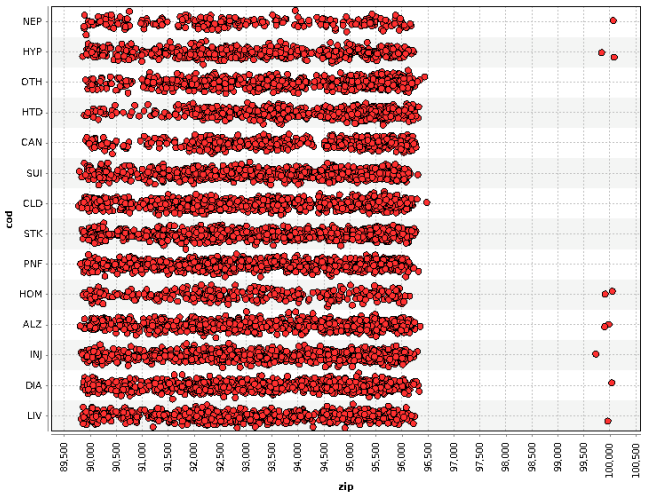

Fig. 4. Zip on x-axis and COD on y-axis. All places have almost all causes of death.

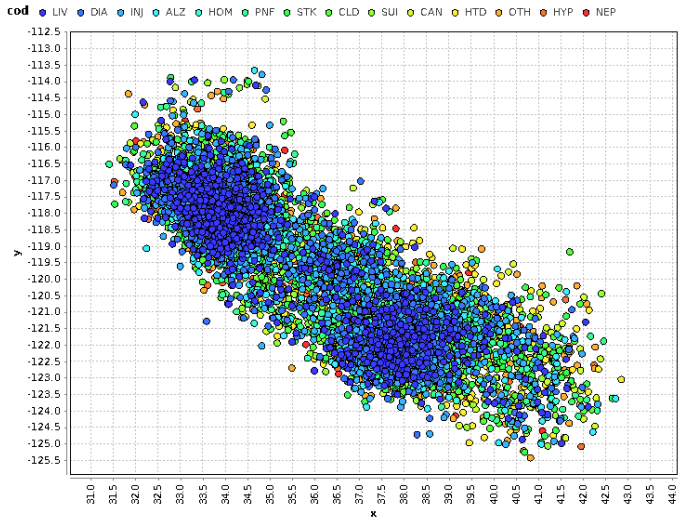

Fig. 5. 'x' vs 'y', color coded COD, the causes of death 'LIV' and 'DIA' seem to be most common, and are in the same area as Fig. 3

suffering from the same disease. Thus, it would be logical that these two places get specialized health camps focussing on the most common diseases of that area.

C. To identify the cause of these common diseases

The data set provided is insufficient to identify the cuase of these diseases. Since we can identify which diseases are causing the most trouble in what area, we need features other than geographical location, such as, local industries, drainage system etc. These can be used in combination to the hospital records to infer the causes of diseases. This would be out of the scope of this paper. However, it must be noted as an important usage of the hospital recommendation system.
With just the help of the hospital recommendation system, the doctors who visit these locations along with the health camps can identify the cause of the disease by looking around for abnormalities. This method, though more independant, is unreliable as it will depend entirely on the knowledge of the doctors.

D. To classify a new patient coming from an identified area as more susceptible to the common disease

A simple, but unreliable method would go as follows: the location of a new patient can be directly mapped on the map and, using Fig 5, a disease can be predicted. A better solution can be given by assuming a number of regions to divide the entire geography in. Assuming that the hospital has the capacity to deploy 20 health camps at a time, we can make a 20 cluster map of the geographical coordinatese using k-means or k-medoids to calculate the region the new patient belongs to. The clusters can be classified on the basis of the most common diseases in them. This method works better because the clustering can be done prior to classification of the patient. It works on the same principle as the first method, just that the classification part is supported more by the computer than the human brain. A decision tree can also provide such information in a much more intuitive way.

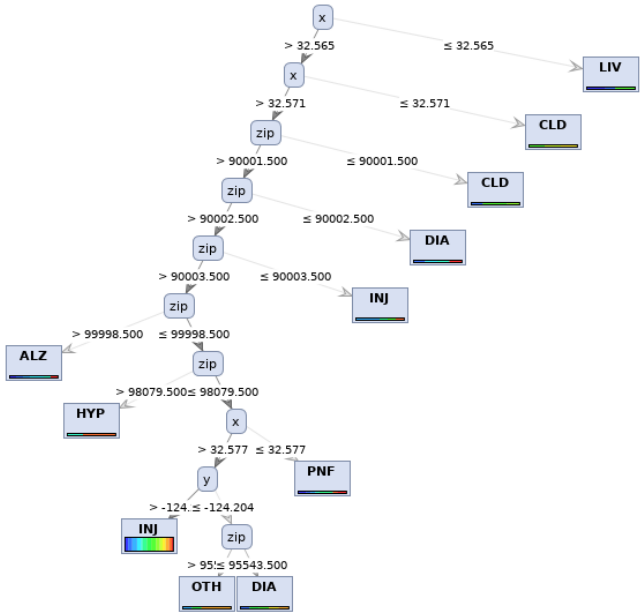

Fig. 6. A decision tree to decide the disease on the basis of x-y coordinates and zip code

III. Conclusion

In this paper, we discussed different ways a Hospital Recommendation System can be modelled. We also discussed

scenarios where the Hospital Recommendation System will be used. It can be concluded that, just the data of diseases and the locations where patients come from can be converted into great deal of information and be used for the people's benefit as seen in the 4 cases.

REFERENCES

[1] Data World, for the Data Set (https://data.world/health/causes-of-death-by-zip-code)