Cloud Enabled Data Analytics and Visualization Framework for Health-Shocks Prediction

Shahid Mahmud [a], Rahat Iqbal [b], and Faiyaz Doctor [c]

[a] *mahmuds4@uni.coventry.ac.uk,* [b] *r.iqbal@coventry.ac.uk,* [c] *faiyaz.doctor@coventry.ac.uk*
Faculty of Engineering and Computing, Coventry University, UK.

Abstract

In this paper, we present a data analytics and visualization framework for health-shocks prediction based on large-scale health informatics dataset. The framework is developed using cloud computing services based on Amazon web services (AWS) integrated with geographical information systems (GIS) to facilitate big data capture, storage, index and visualization of data through smart devices for different stakeholders. In order to develop a predictive model for health-shocks, we have collected a unique data from 1000 households, in rural and remotely accessible regions of Pakistan, focusing on factors like health, social, economic, environment and accessibility to healthcare facilities. We have used the collected data to generate a predictive model of health-shock using a fuzzy rule summarization technique, which can provide stakeholders with interpretable linguistic rules to explain the causal factors affecting health-shocks. The evaluation of the proposed system in terms of the interpret-ability and accuracy of the generated data models for classifying health-shock shows promising results. The prediction accuracy of the fuzzy model based on a k-fold cross-validation of the data samples shows above 89% performance in predicting health-shocks based on the given factors.

Keywords: Technology Integration, Big Data, Data Analytics, Visualization, Cloud Computing, Scientific Overflow of Big Data, Development Process of Big Data Application, and Healthcare Demonstration.

1. Introduction

In the knowledge-driven economies of today, data-driven analytics and decision support harnessing the internet-of-things (IoT) and big data poses unprecedented opportunities for revolutionizing healthcare delivery through the use of cloud computing, machine learning and data mining [1]. Big data in healthcare is concerned with huge and varied sources of meaningful datasets that are too big, too fast, and too complex for healthcare providers to process and interpret using existing tools [2]. It is driven by a continuing need to make health services more efficient and sustainable and the increasing requirement for gearing the delivery of health services toward prevention, early intervention and optimal management [2].

The lack of large population based health datasets acts as one of the greatest obstacles in understanding health-shocks situations, its reasons, and effects in developing countries. Generally, health-shocks can be defined as socio-economic effects on an individual, his family, and society due to the critical illness of suffered by the head of family and/or his family members [3]. From the various studies [4, 5, 6], it is quite evident that the unpredictable timing of health issues and immediate need for large funds for healthcare combined with the travel distance to health facilities could increase the risk of health-shocks in economically less developed areas of world. There is therefore a need to better facilitate the carrying out of large scale population based health studies to collect data pertaining to people's health, social, economic and environmental circumstances as well as their accessibility to healthcare facilities. Additionally there is a need to develop effective and interpretable models from the collected data, which can be used to help decision makers understand the factors associated with health-shocks and predict their occurrences.

Cloud computing plays an important role in facilitating the collection of these datasets as it provides massive amounts of computing and storage power on the internet as a service oriented architecture (SOA) that can be instantly scalable globally [7, 8]. Cloud computing facilitates execution of millions of

commands per second and takes away the technical complexities of hardware and software installation/maintenance and scalability on the go. Cloud computing is becoming more widely adopted by healthcare organizations being a prime big data client for deployment of applications on the cloud. Cloud computing requires major considerations in terms of understanding the unique benefits and the risks factor associated with it. Different models of services such as Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS) [9] in addition to deployment models such as private, community, public and hybrid also need attention in terms of scalability of the system. In regards to data collection surveys based on cloud computing using smart technologies can play a vital role by reducing the time and cost associated with data collection [10, 11]. Furthermore, with the reduction in prices of mobile phones and almost universal coverage, it can make healthcare data capture more feasible at low cost by taking geographical accessibility out of equation [12, 13], thus providing a fast reliable method of capturing accurate data from rural populations.

For the health data collection in Pakistan, we have developed a cloud enabled framework in which all the high performance and computing requirements of GIS system, data gathering and cleaning at large scale and performing data analytics and prediction, have been fulfilled by using a cloud computing infrastructure. Furthermore, GIS helps to provide us highly accurate and interactive maps based on the gathered data. Central to framework is a developed mobile application that enables the healthcare professionals to create and deploy health surveys. As a whole, all data can be collected from mobile application through web-services and stored on the cloud. This can facilitate in maintaining privacy, security, portability, and reliability of the collected data and maximizes the linear scalability, cost effectiveness, deployment and flexibility of data analytics platforms for analyzing the data. Moreover, machine learning based data analytics can be applied at high speed for visualizing and modelling the data to infer useful insights and in supporting stakeholder decisions. Data mining and knowledge discovery techniques are used to automatically model and classify data to identify patterns in the independent data attributes that are associated

with a target dependent attribute and hence be trained to predict the occurrence target outcomes based on new unseen data.

The nature of the collected health data implies that it can contain a number of numerical and categorical parameters that are needed to understand and determine the occurrence of various forms of health-shocks in sampled populations. There are inherent complexities, uncertainties and imprecisions involved in the collection and analysis of this data [14]. These uncertainties are proportionally related to the number and types of variables, the interactions necessary with lay individuals, domain experts and organizations and the sampling methods in collecting the data. It is hard for patients for instance to present their own symptoms and describe how they feel due to language barrier, illiteracy (especially in developing countries), and usage of non-medical vocabulary, which prevents practicing physicians and paramedics from accurately describing and recording their observations [15]. Additionally there are hundreds of inaccuracies due to carelessness of lab technicians and malfunctioning equipment, even under ideal conditions [16]. These uncertainties are profoundly worst in certain developing countries as a result of poor procedures, age-old equipment, untrained paramedics and poorly trained physicians lacking a broad exposure to research and the state-of-the-art in their area of specialization. Hence, there is a need to use qualitative data modeling approaches which both enhance the interpret-ability and accuracy of the classification approach while contending with data uncertainties.

Keeping in view the above, we therefore adopted a fuzzy linguistic summarization (LS) technique based on extracting fuzzy weighted If-Then rules from surveyed population health data. The rules are able to provide a descriptive representation of profiles describing relationship that exist between the independent data variables and the level of health-shocks experienced. The rules are weighted using well-know data mining rule quality measures that enable ranking of the most prominent profile rules for representing the patterns found in the data. The generated rules also specify a classification model which can be used to classify the level of health-shocks experienced based on unseen health

data collected from individual households and villages. The model can be easily retrained on both previous and new data that can be continuously collected via the developed cloud based framework. The system can therefore perpetually update and enhance its classification model and generated profile rules to reflect medium to long term socio-economic, environmental and health effects such as financial crisis, environmental disasters and disease epidemics. We have evaluated the developed cloud analytics framework in terms of the interpret-ability of the generated models for profiling health-shocks and their prediction accuracy using the data obtained form the user study.

The rest of the paper is organized as follows: section 2 presents a literature review on the health scenario in underdeveloped and developing countries, current development in ICT, cloud computing and machine learning techniques in healthcare and health informatics applications; In section 3, we describe the user study which was undertaken to collect data from a population in rural areas of Pakistan using the developed cloud computing framework; Section 4 presents the approach that was used for data preprocessing and predictive modeling of health-shocks that was based on using a fuzzy rule summarization technique; finally conclusions are discussed in section 5.

2. Literature Review

In the following sections, we present a comprehensive review of the existing literature regarding healthcare, health-shocks and the involvement of ICT specifically cloud computing and machine learning in the contemporary healthcare applications.

2.1. Overall Health Scenario: Underdeveloped and Developing Countries

To understand the healthcare system of underdeveloped and developing countries, a study was conducted regarding the impact of variables such as age, gender, religion, area of residence, income, education, body mass index,

smoking, employment, marital status, health expenditure, health service quality, working conditions, and living conditions [17]. Here, it was observed that in Bangladesh 63.8% of health expenditure was out-of-pocket, which resulted in financial losses affecting an individual breadwinner. Compounded with other issues such as low wages and inability to find work following illness, this can lead to a spiral of debt affecting both the breadwinners immediate family as well as their wider communities. Generally, unpredictable timing of health issues and the immediate need for large personal funds for healthcare increases the risk of hardship financing healthcare [6, 18, 19, 20] in contrast the possession of assets and having regular income-flow were shown to be predictors of lower expected hardship financing. In [4], "hardship financing" for poor households in an Indian town was studied. Here, it was observed that mostly rural households were subjected to financial hardship due to indirect and long-term costs of healthcare, which resulted in withdrawing a child from school and/or skipping meals.

Health patterns of people belonging to lower socio-economic status (SES) were discussed in [21]. It was observed that new chronic conditions were related to household income, wealth, and education. Furthermore, dimensions of SES such as income, wealth, and education as means of predicting future health outcomes were also discussed. It was observed that higher SES people may have better access to medical care, more information about appropriate medical practice, less strenuous jobs, or access to more material inputs that cumulatively improve health. Moreover, these people may live in more health promoting environments whereas people with lower SES do not get promoted and/or can be more easily expelled from their jobs.

In [6], it was observed that significant economic benefits can be achieved by improving health in developing and developed countries. Generally, health can contribute to economic outcomes through higher productivity from sustained employment, higher labor supply, and improved skills which can increase the financial resources for investment in physical and intellectual capital [22]. Moreover, healthier individuals, with a longer lifespan tend to have greater incentives to invest in education and training as they can harvest the associated

benefits for a longer period, which has a knock effect on wealth generation and economic productivity.

2.2. Healthcare and ICT

Currently, healthcare providers have a seemingly infinite combination of graphic, textual, coded, statistically analyzed, collated, filtered, non-filtered and malleable data [23]. As a result, healthcare organizations endeavour to construct comprehensive summary views of patient's medical records.

In a study undertaken by [23], an exhaustive survey was conducted for the use of ICT in healthcare. Here, definitions of tele-health and tele-medicine along with e-health and an appraisal of mechanisms to weigh the ICT interventions in health were discussed. However, it was too generic to conclude any direct ICT based interventions for the improvement of health systems in a country such as Pakistan. Similarly, a summarized report on HIS-EVALs was presented in [24]. The study provides a historical account of work in this area and highlights the importance of different disciplines for improvement in health information systems such as biostatistics, medical informatics, psychology, computer science, and health economics. Furthermore, Pakistan's health system was defined at the district, provincial and federal levels in addition to key issues of the poor and declining health landscape of Pakistan. In order to reduce health expenditures by adopting health information technology (HIT), a large study was conducted in the U.S. [25]. HIT allows clinics and physicians to manage information providing a secure exchange between healthcare consumers and providers to reduce medical errors. Here, various clinic-level characteristics were combined with geographic location-specific information to create a comprehensive dataset for examining factors that influence HIT adoption decisions.

2.3. Cloud Computing and Healthcare

According to [8, 26], a cornerstone of successful wide spread deployment of data analytics is cloud computing. Generally, cloud computing is built around a series of hardware and software elements that can be remotely accessed through

any web browser. Furthermore, it provides massive amount of computing and storage power on the internet as Service Oriented Architecture (SOA) that can be instantly scalable and globally accessible [8, 26]. The models of cloud computing can help in accelerating the potential of scalable analytics solutions [27]. It offers efficiencies and flexibility for accessing data, performing analysis, delivering insights and extracting value from accumulated data. Regardless of the cloud delivery model, it can be used to unlock the potential of big data though a cloud environment [7]. Moreover, in case of bid data, the storage performance on cloud in comparison to non-cloud environment is far better in addition to data security and integrity [28, 29].

In health-care, the use of big data with predictive analytics has a huge amount of potential, especially, when paired with cloud-based platforms. Cloud computing is significantly facilitating patients, physicians and doctors in terms of data sharing and its availability, regardless of the location of the patient and clinicians [30, 31]. Furthermore, cloud computing in addition to internet-of-things (IoT) and emerging services can play a vital role in conducting healthcare surveys by reducing the time and cost associated with data collection [11, 27]. Moreover, with the reduction in the prices of mobile phones and their near universal coverage, it can make healthcare data collection and delivery more accessible at low cost by irrespective of geographical distance [12, 13].

In general, cloud computing is facilitating the deployment of electronic health records (EHR), data sharing, enhancement and management of data on patients enrollment, revenue cycle and claims processing, just to name a few. However, cloud computing requires a major consideration in terms of understanding the unique benefits as well as the risk factors associated with its usage. In this regard, different models of services are being widely used such as Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS). Similarly, various deployment models such as private, community, public and hybrid need consideration in terms of scalability of the system [9]. Though it provides new and improved patient care capabilities, however data privacy, security, and reliability are still important issues that needs to be resolved in

building cloud oriented system architectures.

2.4. Intelligent Computational Techniques

Generally, various data mining and knowledge discovery techniques can be used for modelling and classifying population based data to identify patterns related to different levels of health-shocks. In healthcare, different techniques such as clustering, classification, regression, association rule mining and decision trees have been widely used [32]. In [33], five classification algorithms namely decision trees, artificial neural networks, logistic regression, Bayesian networks and Nave Bayes have been used for building classification models in order to plan and implement healthcare service programs based on the healthcare demands of local residents. Similarly, a study is presented in [34] for predicting cardiovascular autonomic (CA) dysfunction in the general Chinese population using artificial neural network (ANN) based prediction models. In [35], the authors have applied support vector machines (SVM) for classifying and detecting people with diabetes and pre-diabetes in a cross-sectional representative sample of the U.S. population.

Similarly, rule induction approaches such as associative rule mining, rule based classification and rule based fuzzy logic systems have also been successfully applied within the health informatics domain. A data-adaptive rule-based classification system for Alzheimer's disease classification is described in [36]. Here the system generates relevant rules by finding adaptive partitions using gradient-based partitioning of the data. Here, adaptive partitions are generated from a histogram that is used for analysing the discovered rules, which are used to assist in classifying the new data correctly. In [37], a fuzzy based approach was presented for cholera prediction based on a case study of Southern African. The proposed approach takes into account various factors including preconditions for cholera outbreaks, environmental conditions and socio-economic factors for building the prediction model. The fuzzy model is developed based on historical information, expert knowledge, and climatic and biophysical parameters while cholera outbreak risk was the output parameter. The approach is

aimed to minimize the impact of cholera by informing the policy makers using the developed fuzzy prediction model. Other applications of fuzzy systems have been used to manage malaria [38], automatic control of anesthesia during surgical procedures [39], conduct medical diagnostics [40] model epidemiology to risk factors [1] as well as cancer treatment [41] and the prediction of water shortages[42]. In [43], an adaptive fuzzy rule based linguistic summarization (LS) system is proposed for modelling the behavioral cues of dementia patients based on monitoring their interactions in the home through the use of smart devices and environmental sensors. Here, the aim is to use the generated rules which are also weighted to profile patterns of an individual's behaviours and track behavior changes that can indicate cognitive decline due to disease progression.

In general, Fuzzy logic systems have proved to be an ideal choice to model healthcare systems due to their ability to handle uncertainties, imprecisions, complexity and incompleteness of information [15, 16]. Fuzzy systems provide transparent and flexible rule based models through the use of linguistic quantifiers [44] which can provide a methodology for predictive modeling and classification using approximate reasoning of uncertain information.

3. User study

To understand the health-shocks and its causes in the rural and remote areas of Pakistan, we conducted a user study to collect a dataset of 1000 households from the district Haripur with the help of Begum Mahmuda Welfare Trust hospital (BMWT). Here, one of the main objectives was to identify the shortcomings of the present healthcare system, especially in rural and tribal areas of Pakistan. In the user study, questionnaire was divided into 12 sections with an aim to obtain geographic, demographic and socio-economic data, so that a comprehensive picture regarding the living standards of the participants and effects of health-shocks could be drawn.

For this purpose, we have developed a cloud based system to capture, store,

index and retrieve the data as shown in Fig. 1. Here, the data was collected through offline questionnaires and an online mobile based system. All data was coded in the online system for analysis and visualization purpose. The developed system has a synchronized GIS web-portal that allows to visualize and monitor the data collection activity in real-time. Amazon web services is integrated with (GIS) to facilitate the capture, storage, index and visualization of data which is collected through the use if smart devices for different stakeholders where in our case smart phones with a dedicated survey app was used for this purpose. The proposed framework further supports role based privileges for both the responders and healthcare professionals. Healthcare professionals can create and deploy different survey forms. Similar dataset can be found in studies carried out by Pakistan Poverty Alleviation Fund (PPAF) and Poverty Scorecard for Pakistan (PSP) [45, 46, 47]. However, PSP [47] focuses only on the poverty dimension and the PPAF questionnaire is intended to gauge the impact of PPAF activities for poverty alleviation [45, 46]. However, the focus of our user study was on geographic, demographic and socio-economic.

3.1. Data Analysis and Visualization

Data analysis and visualization is vital to understand the hidden pattern in data that can lead to policies and recommendations for better planning, cost efficiency, and thus improve quality of life for the target population. In this regard, we created a tool for data visualization as shown in Fig. 2. The tool is linked to the above described architecture (see Fig. 1). Before going into technical details, a brief description of district Haripur is given below:

Haripur district is in the Hazara region of Khyber Pakhtunkhwa province of Pakistan. It is located in a hilly plain area at an altitude of around 610 meters (2,000 ft.) above sea level. According to [48], it has an estimated population of $1,024,497$ with a rural to urban ratio of 88%:12%. In district Haripur, monthly income of the family depends on the number of persons who are involved in skilled labour, unskilled labour and/or child labour. Here, it is worth mentioning that only 13.33% households have monthly income of PKR 15000

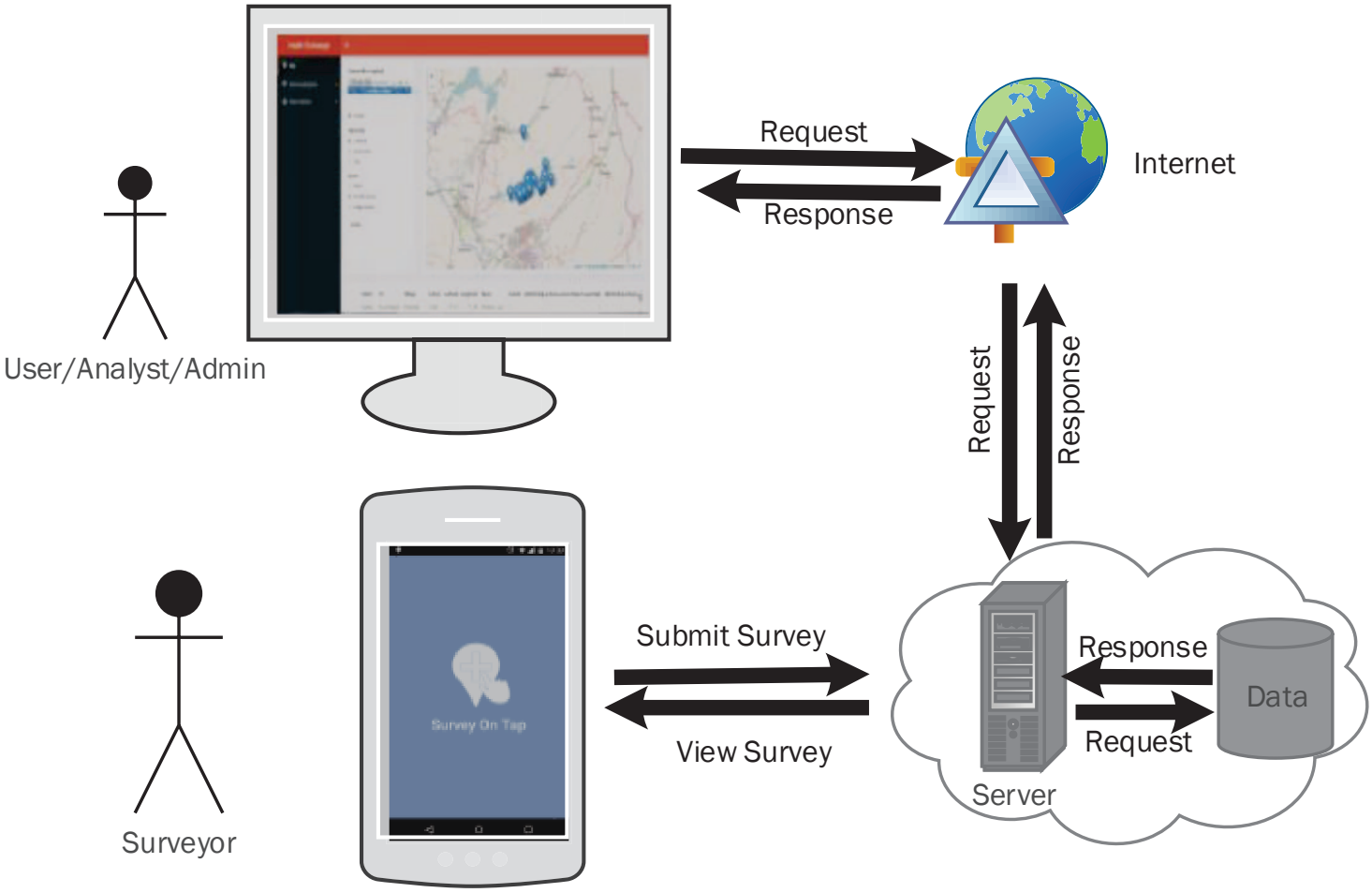


Figure 1: Mobile Survey Architecture.

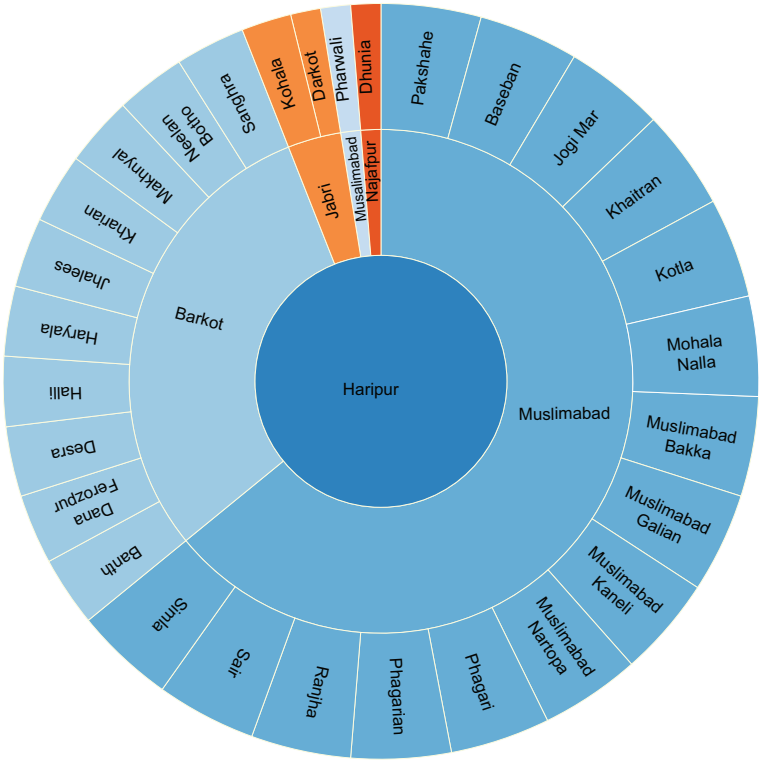


Figure 2: Villages that participated in the survey. Inner most circle represents the district Haripur whereas the outer circle represents the union councils and the outer most circle represents their corresponding villages.

which is approximately 150 USD. Furthermore, only 3% of the households have a monthly income of PKR 35000 which is approximately 350 USD as shown in Fig. 3.

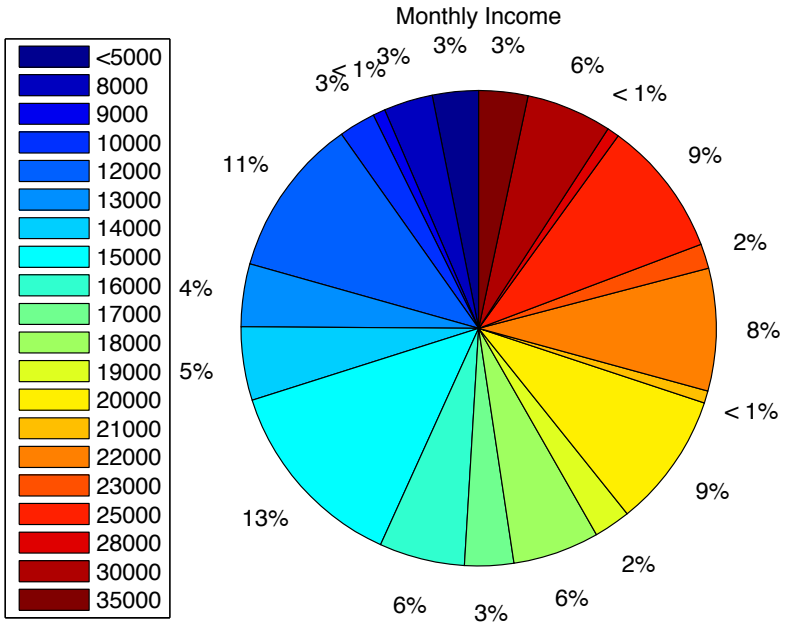


Figure 3: Monthly Income of Households in District Haripur.

In district Haripur, more than 42% of the population is below poverty line; whereas, 31% of the population is in the middle-income group. Few of the main reasons of poverty include: an uneven distribution of resources, poor human resource development and polarization of power. Due to poverty, one of the main sources of income is child labor as shown in Fig. 4.

Moreover, for an estimated population of $1,024,497$, there are about 6 hospitals, 6 rural health centers (RHCs), 6 sub health centers (SHCs), 42 basic health units (BHUs), 2 mother and child health centers (MCHs), 9 dispensaries, 1 tuberculosis (TB) clinic and leprosy clinic. Fig. 5 shows the plot of all the health facilities on the map of district Haripur.

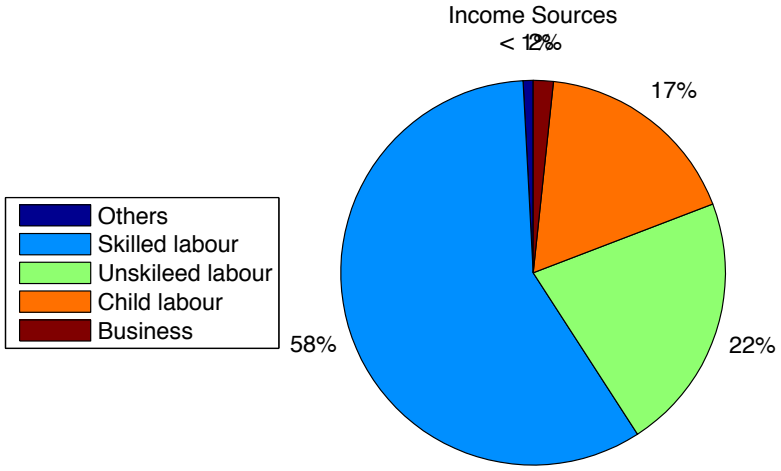


Figure 4: Income Sources of Households in District Haripur [49].

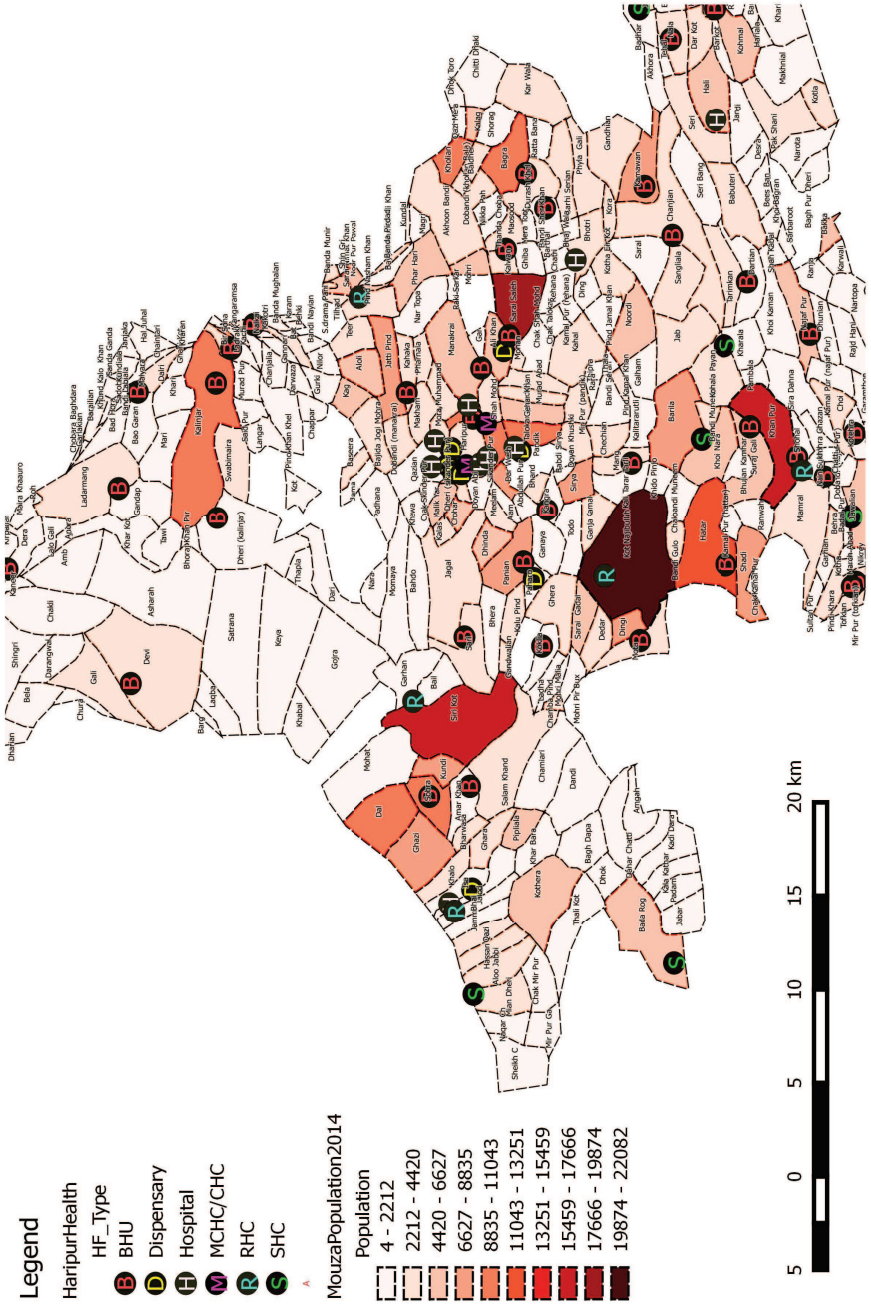


Figure 5: Thematic map representing the population of different Mouzas of district Haripur with different health facilities.

Here, it is worth mentioning that due to shortage of medical doctors and paramedical staff in addition to different medical facilities, the number of patients per month at different BHUs is not more than 300. However, in comparison to BHUs, the patient load at DHQ hospital Haripur is around 15,000 to 18,000 due to the presence of sufficient number of doctors, paramedical staff and different medical facilities (GoP, 2010). Furthermore, in BHUs, a single doctor is shared by multiple BHUs, for example, one single doctor could be appointed for both BHU Amgah and BHU Bandi Sher Khan. Due to such situations, people of district Haripur have to spend out-of-pocket either to reach hospitals in urban areas or to avail private health facilities, which act as one of the major reasons for the health-shocks.

The district's current health infrastructure does not correspond well with its population. In the district of Haripur, there is only one bed for every $2,247$ people compared to one bed for every 100 people that would be found in the developed countries of the world. Moreover, in district of Haripur, infant mortality rate (IMR) is 66 whereas overall maternal mortality rate (MMR) in the province of Khyber Pakhtunkhwa is 275 [48].

Here, it is worth mentioning that in addition to features above. monetary values were also noted. These monetary values include the transportation cost from the village to nearest health facility. On average, each family has to spend 3634 PKR just to reach the nearest government health units. Moreover, due to limited health facilities at the nearest health units, patients have to travel 38 kilometers (round trip) on average to the main hospital. This round trip costs them 9186 PKR on average. Due to the poor road condition and shortage of transport, it takes from few hours to one day to travel 38 kilometers. The high cost of travelling in comparison to the monthly income of each household is one of the major causes of health-shocks in the district Haripur where more than 42% of the district's population is below poverty line and 31% of the population is in the middle income group that is living hand to mouth. Moreover, the unemployment rate in the district Haripur is almost 30% [48].

To understand the health-shocks and its causes in the rural and remote areas

of Pakistan, we collected a dataset of 1,000 households from the district Haripur with the help of Begum Mahmuda Welfare Trust hospital (BMWT). Haripur district has 146,375 total households with an average family size of 7, 49 to 51% male to female ratio approximately, 2.22 person/acre population density (550/26 people/square km, rural to urban ratio of 88 to 12%, and literacy rate of roughly 65%. The BMWT dataset contains 1,000 households from 29 villages of 5 different union councils of district Haripur. Data analysis and visualization is vital to understand the hidden pattern in the data that can lead to policies and recommendations for better planning, cost efficiency, and thus improve quality of life for the target population. To better understand the collected data we performed various cross sectional analysis. Fig. 6 shows the population split and their reach to schools. Muslimabad is the biggest with 47% population, Barkot has 33.8%, Najafpur 10.6%, Jabri 6.3% and Musalimabad 15%. The residents of Musalimabad had to travel larger distances to reach schools. Schools are most accessible for residents of Jabri. The interquartile range is within 7 KMs for all union councils.

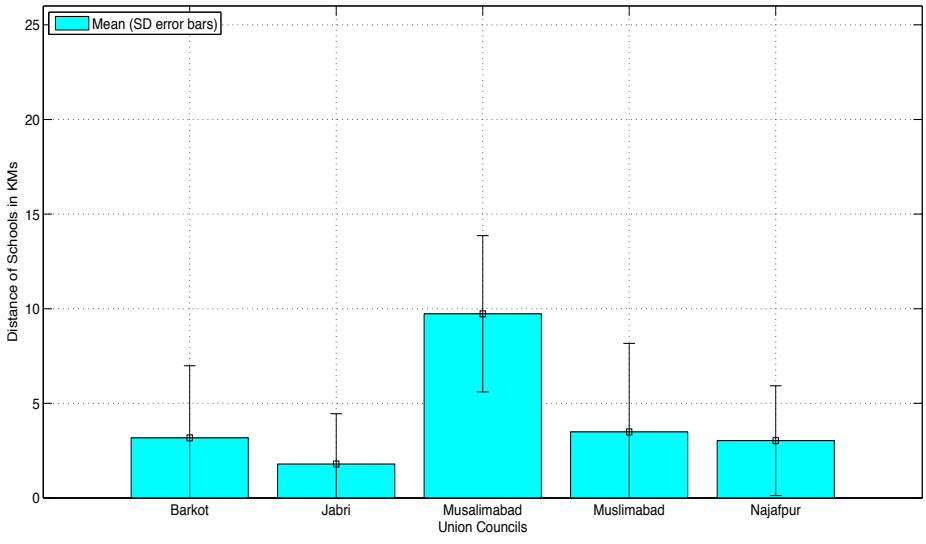


Figure 6: Population Split and Reach to Schools.

Fig. 7 shows the distance to reach to a basic health unit. Barkot is the fartheset, Jabri and Najafpur have more accessibility.

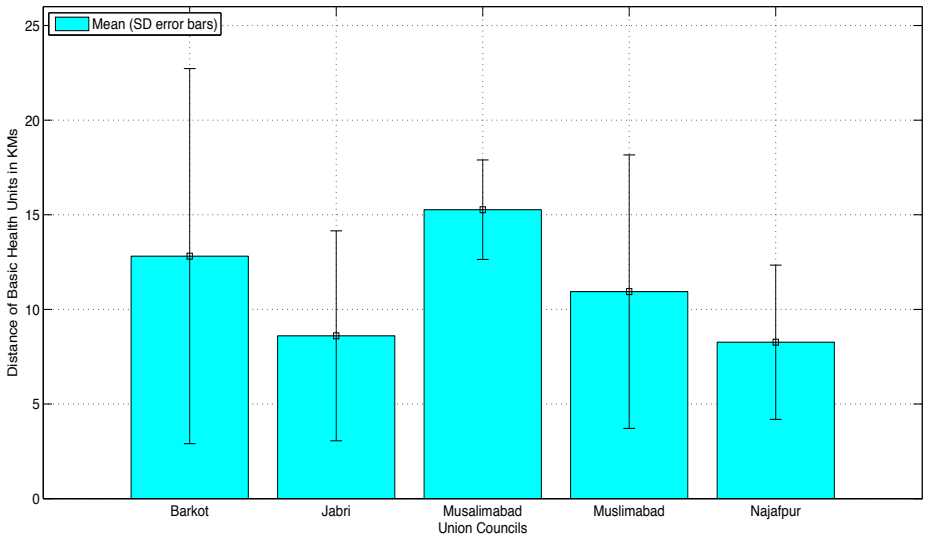


Figure 7: Distance to Basic Health Units (BHUs).

Here, the minimum travelling cost from village to BHU is approximately PKR 1200 and it goes up to PKR 8000. In contrast to travelling cost, 42% of district Haripur is below poverty line, where 13.3% of population has a monthly income of less than PKR 5,000. Moreover, in case of emergencies or major operations, people of district Haripur have to travel to the main district hospital whose travelling cost ranges from PKR 6,200 to PKR 1,4000. For majority of people, the most obvious and first choice to solve the above mentioned problems is to borrow money from their relatives and friends as shown in Fig. 8. Even in case of personal violence problems, residents borrow money to resolve the issues, which indicates that most of the personal violence issues are related to money. Hence, these problems force people to rely on debts, which acts as a major cause of health-shocks as shown in Fig. 9. It also shows that the majority of people are in debt from moderate to large sums of money.

Figure 8: Spiral of Problems and Debts.

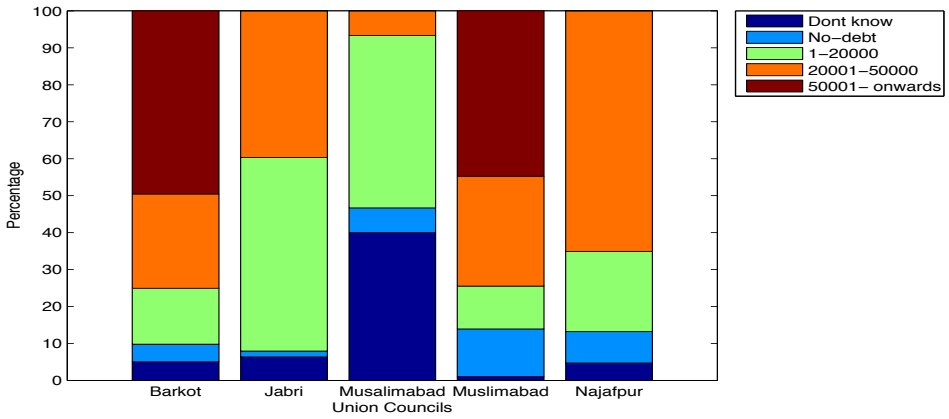


Figure 9: Debt Scenario in District Haripur.

Moreover, majority of houses have stone-and-mortar walls along with thick wood ceiling. Metal sheeting is mainly found in Barkot (28% residents) and Muslimabad (13% residents). Houses will get damaged due to harsh weather with equal chances of major and minor damages. Most of the reconstruction will be completed within 2 years of damage. No significant relation was found between house strength and ownership of house. The Haripur Villagers gender distribution shows that Jabri has reported 100% Males and Barkot, Muslimabad, Najafpur has majority of males. Only Musalimabad reported majority of females, i.e., 66.67%. Fig. 10 shows number of permanent adult members of the household. Majority of households have 2 to 4 adults. The household size is right-skewed and there are a small number of families with much larger number of adults.

Here, it is worth mentioning that in BMWT dataset, there is a positive relationship between distance to basic health units (BHUs), percentage of debts, toilet facilities, and frequency of major illnesses. Due to debts and distance to BHUs, minor illness turns into major. Similarly, dependency ratio, which is defined as age-population ratio of those typically not in the labor force (the dependent part) and those typically in the labor force (the productive part),

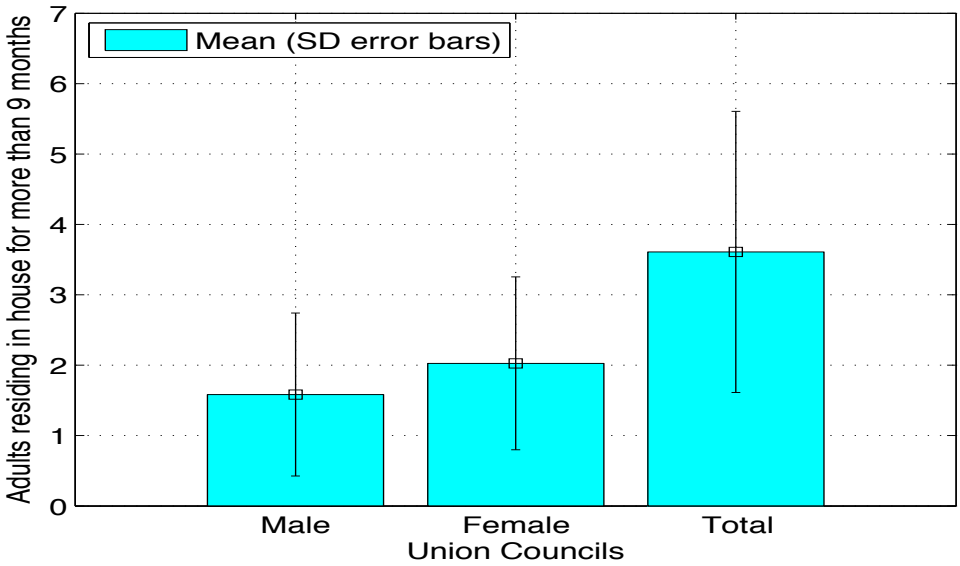


Figure 10: Adults Living in the Household.

ranges from 3.4 4.9 for villages of district Haripur. Fig. 11 shows earning adults, total adults and the dependency ratio. Barkot has the highest dependency ratio, while Jabri has the lowest.

In Haripur, water is disposed mostly within 75 meters of home or piped down the drain. Very little is used for irrigation. Garbage is discarded within 75 meters of the house or else burnt. Food is fed to the livestock if not disposed within 75 meters as shown in Fig. 12.

Fig. 13(a) presents the situation concerning the land ownership. Muslimabad has the highest number of private owners. However, majority respondents refused to share the acreage details of their land. Furthermore, no relation between size of land ownership and type of land ownership was found during the survey. Similarly, Fig. 13(b) presents the food shortage. Based on BMWT survey, 47% households have never faced a food shortage, while 4% face bi-weekly food shortage.

Furthermore, it was shocking that houses with toilet facilities had higher

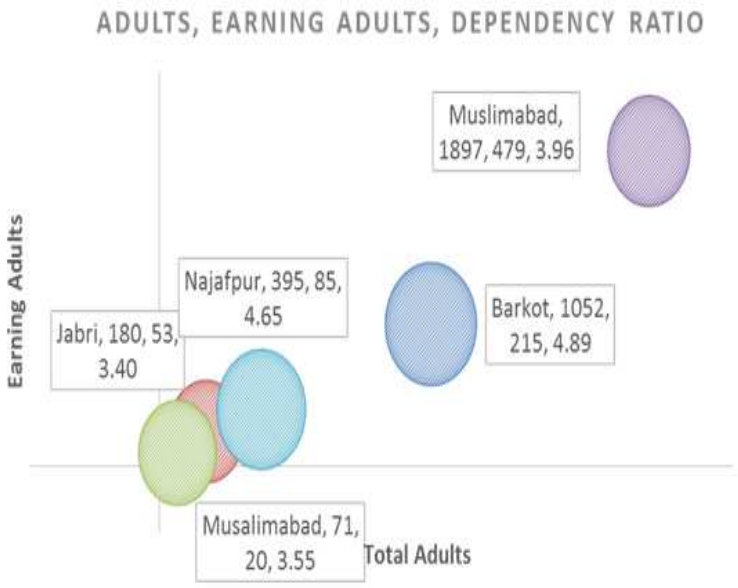


Figure 11: Dependency Ratio.

rates of diseases in comparison to houses with no toilet facility. One of the main reasons for this was access to water resource and poor sewerage system. Here, it is worth mentioning that the time required for one person of a household to collect water for one day's usage is almost 4 hours as in some cases, it takes 1.5 to 2 hours to reach the water source. The same amount of time is required to carry that water back home. In the case of a family with two to three children, it requires more than 5 or 6 buckets of water at least for a day.

A similar relationship has been observed between frequency of minor/major diseases and toilet facilities as shown in Fig. 14. Here, minor disease is defined as any normal disease or injury which doesn't require bed rest where as any disease or injury which require 2 or more days of bed rest or hospital admission including disability is represented by major disease.

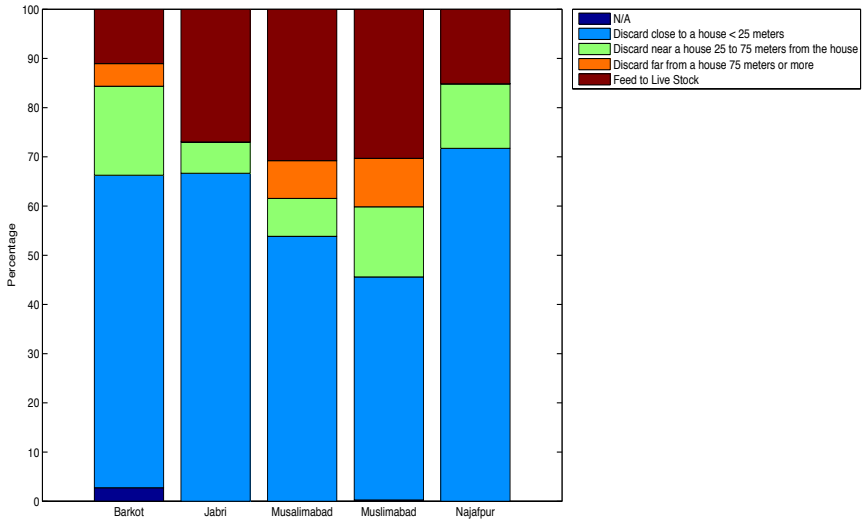


Figure 12: Food Disposal.

4. Proposed Data Modeling and Analytics Approach

4.1. Preprocessing

This section discusses different steps taken to preprocess the data, which was acquired in the user study (described on section 3) in order to apply data modeling and analytics approach. The preprocessing approach consisted of the following steps.

4.2. Data Normalisation

In order to minimize the noise in the dataset, the data instances with values out of the predefined data ranges were removed. After the removal of outliers, the data categories of categorical variables were ranked by giving each category an ordinal number based on expert opinion. Furthermore, these ranked categories were assigned data instances related to the categorical variables. Finally, in order to unify the range of the data for the all considered variables, values

were normalised between 0 and 1, i.e., by dividing each data instance by the maximum value of the variable which it represents.

4.2.1. **Integrating Variables**

Here, the variables in the dataset were combined into four main factors (derived variables) namely: Living Standard, Health Risk, Access and Income Allocation. Following is a description of how these factors were calculated based on information derived from the tool.

4.2.1.1. **Living Standard.** Living standard is derived using the variables in the dataset which are related to the household's standard of living which include: Nature of Ceiling (NC), Resistance of House against Severe Weather (RW), Disposing-off of food (DF), Disposing-off of garbage (DG), Disposing-off of water (DW) and Water Source (WS). Furthermore, theses variables were then combined using the following equation:

$$LivingStandard = \frac{NC + RW + DF + DG + DW + WS}{6} \tag{1}$$

4.2.1.2. **Health Risk.** This factor was derived using the variables, which are related to the current health situation of the person. Theses variables include: Minor disease frequency (MD), severe disease frequency (SD), Dental hygiene level (DL) general hygiene level (GH) and Toilet Facility (TF). The influence of these variables contribution to health risk are not equal, hence in calculating the health risk, these variables were given different weights based on the following equation:

$$HealthRisk = \frac{2 \times MD + 4 \times SD + DH + GH + TF}{9} \tag{2}$$

4.2.1.3. **Access.** This factor presents the cost of access to the health treatments. It is influenced by; cost of travel to the local health unit which is generally visited to treat the minor diseases (CM), the cost of travel to the hospital which is normally visited to treat the severe diseases (CS) (which was found to be a fixed approximate cost of 50 rupees per check in that area), Local health unit doctor charges for minor disease treatment (LC) and finally the Hospital

doctor charges for the sever diseases (HC) (which was found to be a fixed approximate cost of 250 rupees per check in that area). The following equation therefore describes the calculation of Access:

$$Access = (CM \times MD \times 50) + (CS \times SD \times 250). \quad (3)$$

4.2.1.4. **Income Allocation.** This factor refers to financial situation of the family. Here, it is derived from: Annual Income of the family (AI), Owned Land size (LZ), Cost of Food (CF) (which was found to be 10% of family income per person), Cost of Maintenance of the house (CM) (which was found to be 20% of the annual income of the family) and Debt (DT). This factor was calculated based on the following equation:

$$IncomeAllocation = AI + LZ - CF - DT, \quad (4)$$

where the Annual income was calculated by multiplying the number of earning adults (ED) in the family by 15000 rupees, (which was the average monthly income per working person) which was then multiplied by 12 to calculate the annual income based on the equation below:

$$AI = ED \times 15000 \times 12. \quad (5)$$

4.2.2. **Data Labelling**

The data was automatically labelled by calculating the estimated risk of health-shocks (EHS) as a result of the following equation:

$$EHS = HealthRisk + Access - LivingStandard - IncomeAllocation. \quad (6)$$

where Health Risk and Access were both considered to have positive relationship with the health-shocks while Living Standard and Income Allocation were considered to have an inverse relationship with the health-shocks.

Here, values of EHS were normalised by dividing each data instance with the maximum estimated health-shock found in the data. Once the data was automatically labelled, it was presented to a field expert to check and make adjustments, if needed. The purpose of labelling the data was to facilitate the expert by providing an initial labelling, which they could amend as needed.

4.3. **Proposed Fuzzy Linguistic Summarization Approach**

Fuzzy Logic Systems (FLSs) provide transparent and flexible model for the handling of real world information imprecision through the use of linguistic quantifiers such as Poor or High [44]. FLSs represent a methodology for computing with words where linguistic quantifiers described using fuzzy sets are combined with human interpretable If-Then rules [44]. The fuzzy rules convey richer and more easily understandable linguistic summarization (LS) of patterns associating the independent input variables with the dependent target output decisions or states found in the data [50]. Additionally the extracted fuzzy classification rules rule have quality measures associated with each rule that can be used measure the strength of patterns found in the data and provide the ability to rank the top rules associated with particular output conditions. Here, we have used a Fuzzy Linguistic Summarisation approach [51] consisting of four phases as shown in Fig. 15, which is described below:

4.3.1. **Definition of Linguistic Quantifiers from Data**

In phase 1, the input/output data comprising of the four independent variables and the single dependent variable representing severity of health-shocks are mapped to predefined linguistic quantifiers where these derived variables were acquired from the pre-processed data collected during the user study. In numerical and continuous valued data attributes, uncertainties relating to the linguistic quantification over different data values of the attribute suggest the need to use fuzzy sets. This is a generalisation of a crisp set that allows the gradual assessment of the membership of an element belonging to a set by using a fuzzy membership function (MF) as follows [52]: Given a domain of discourse $\mathbf{X}$, a fuzzy set $\mathbf{A}$ on $\mathbf{X}$ is a set expressed by a characteristic function $mu_A : \mathbf{X} \rightarrow [0, 1]$ that measures the membership grade of the elements in $\mathbf{X}$ belonging to the set $\mathbf{A}$:

$$\mathbf{A} = \{(x, \mu_A(x))|\forall\, x \in \mathbf{X}, \mu_A(x) \in [0, 1]\}, \tag{7}$$

where $\mu_A(x)$ is called the fuzzy MF of the fuzzy set $\mathbf{A}$.

Here, we divided the preprocessed data into a set of MFs which quantify the values of the data attributes into linguistic labels that partitions the data space into fuzzy regions. Each variable's space is partitioned into five overlapping triangular MFs (Low, Medium to Low, Medium, Medium to High and High) covering the range of the independent and dependent variables, an example of which is shown in Fig. 16. This was achieved and further optimised using expert knowledge pertaining to the ranges for each of the variables.

4.3.2. **Fuzzy Rule Extraction from Data**

In phase 2, fuzzy rule extraction was carried out based on the approach described in [53]. It is a single-pass method for extracting fuzzy rules from sampled data. The data is mapped to the fuzzy sets for the antecedents and consequents of the rules generated in phase 1. We use the approach in [53] to extract multi-input antecedents (for each independent variable) and a single-output consequent (for each dependent/target variable). Combinations of these describe the relationship between y^t and $x^t = (x_1, \cdots, x_n)^t$, that take the following form:

$$\text{IF } x^t \text{ is } \mathbf{A}^q \text{ and } \cdots \text{ and } x_n^t \text{ is } \mathbf{A}_n^q, \text{ THEN } y^t \text{ is } \mathbf{B}^q,$$

where $s = 1, 2, \cdots n$ and n is the number of inputs, $t = 1, 2, \cdots, N$, where N is the number of data instances [54], q is the value of one of the predefined linguistic labels associated to the input or antecedent fuzzy sets $\mathbf{A}$, and the output or consequent fuzzy set $\mathbf{B}$. Using the process described we generate an If-Then profile rule for each data instance. This will result in a profile rule-base comprising of duplicate and contradictory rules.

4.3.3. **Compression of Fuzzy Rules**

In phase 3, the data instance based profile rules are then compressed in order to summarize the data instances into unique end rules. This process involves a modified calculation of two rule quality measures from which we then derive the scaled weight of each unique summarization rule. The quality measures are based on generality (measuring the number of data instances supporting each

rule [50]) and reliability (measuring the confidence level in the data instances that support each rule [50]). In our approach, the rule generality is measured using fuzzy rule support and the reliability of the rule is based on calculating its confidence.

The fuzzy rule support of a rule is computed as the product of the rule's support and firing strength. Here, the support of a rule refers to coverage of input data instances that map to it [55], while its firing strength measures the degree to which the rule matches those input data instances [44]. The rule's fuzzy support can be used to identify the unique rules with the most frequent occurrences of data instances associated with them, where the data instances most closely map to those rules. The fuzzy support of each rule is scaled based on the total data instances mapping to each output (consequent) set so that the frequencies are scaled in proportion to the number data instances found in each consequent set. The calculation of the scaled fuzzy support for a given uniquely occurring rule is based on the calculation described in [55] and [51]. Following the calculation of the rule support, duplicate instance based profile rules can be identified and eliminated to compress the rule base into a set of M unique and contradictory rules for modelling the data points.

The confidence of a rule measures the rule's validity in describing how tightly data instances are associated to a specific output set. The confidence value range is between 0 and 1. A confidence of 1 implies that the pattern which the rule describes is completely unique to a single output (consequent) set. A confidence of less than 1 implies that the pattern described in the rule occurs in the data associated with more than one output (consequent) set. In this case it should then be interpreted as being best associated with the output set having the highest confidence. The rule scaled confidence is based on the calculation described in [55] and used in [51].

4.3.4. Calculation of Scaled Rule Weights

In phase 4, each rule's scaled fuzzy weight is calculated as the product of the scaled fuzzy support and confidence of a rule as shown below:

$$scW_i = FuzzSup \times Conf \tag{8}$$

where $FuzzSup$ is the scaled fuzzy support and $Conf$ is the scaled fuzzy confidence. Each of the generated M rules is assigned the scaled fuzzy weight measure scW_i and takes the following form:

$$\text{IF } x^t \text{ is } \mathbf{A}^q \text{ and } \cdots \text{ and } x_n^t \text{ is } \mathbf{A}_n^q, \text{ THEN } y^t \text{ is } \mathbf{B}^q,$$

The scaled fuzzy weight measures the quality of each rule in its ability to model the data. It can be used to rank the top rules associated to each output set and choose a single winner rule among compatible rules based on methods for rule weight specification described in [55] and used in [51].

5. System Evaluation

In order to achieve an efficient fuzzy rule based system, two quality aspects should be considered; interpretability and accuracy. Interpretability refers to ability of the model to generate understandable and sensible system in terms of the real world systems rules. Accuracy refers to ability of the system to produce a similar response to the real world system [56].

5.1. Model interpretability

Interpretability is a subjective property which depends on the expert opinion and could be influenced by different factors such as structure of the fuzzy model, the number of the input variables, the number of the linguistic labels and shape of the fuzzy sets [56]. In order to assess the interpretability of the produced fuzzy rules, the rules were reviewed by field experts to give their judgement on the understandability of the rules in terms of whether they made sense in terms of the health-shocks risk estimation based on the conducted user study or not.

Table 1 shows a sample of the produced rules where each rule consists of the four antecedents (input variables): living standard, health risk, access and income allocation and one consequent (output variable), which is the heath-shock risk estimation. The rules in the table are sorted in descending order by their scaled weight, which expresses its firing strength. The scaled weight was added to increase the interpretability of the rule by providing additional information about how dominant that rule was in terms of representing and modeling the patterns found in the data set.

Considering the first three rules as examples to explain the interpretability of the generated rules, rule one states that if Living Standard is Medium (M), Health Risk is Medium to Low (ML), Access is Medium (M) and Income Allocation is Medium (M) then Heath Shock is Medium (M) which makes sense as only Health Risk is ML while the three other factors are M making the estimated health-shocks risk to be M. Rule two is also understandable suggesting that when Income is MH this decreases the health-shocks risk possibly as a result of ML income allocation. Finally rule three, represents the case where health risk is relatively high (MH) and the Access cost to the medical treatment is also high (MH) which is associated with poor Income Allocation (ML) leading to a high (H) health-shocks risk probability.

5.2. **Model classification accuracy**

In contrast to interpretability, accuracy is a more objective measure of model performance. In general, there are a number of well defined methods and measures to evaluate the accuracy of the model such as classification and regression, which assess the accuracy based on the percentage of the correctly classified data instances in the dataset. Here, the accuracy of the developed fuzzy model was evaluated using the classification measure in two ways; the ability of the system to accurately classify the health-shock risk on the full dataset (model accuracy) and the ability of the system to correctly classify the heath-shock risk on unseen data (prediction accuracy).

In order to evaluate the modeling accuracy of the fuzzy model on the full

Table 1: Weighted Fuzzy Rules

	Living Standard	***Health Risk***	***Access Access***	***Income Allocation***		***Heath Shock***	***Scaled Weight***
1.	M	ML	M	M	→	M	0.19463
2.	M	ML	M	MH	→	ML	0.18145
3.	M	MH	MH	ML	→	H	0.12755
4.	M	MH	H	M	→	H	0.09921
5.	MH	ML	M	M	→	ML	0.09677
6.	M	MH	MH	M	→	MH	0.08592
7.	ML	MH	MH	ML	→	H	0.08036
8.	M	MH	M	ML	→	MH	0.07852
9.	M	L	M	M	→	ML	0.0672
10.	M	M	M	M	→	M	0.06486
11.	M	MH	M	M	→	MH	0.05609
12.	M	ML	MH	M	→	M	0.05428
13.	ML	ML	M	M	→	M	0.05399
14.	MH	ML	M	MH	→	ML	0.05161
15.	M	M	MH	L	→	H	0.04762
16.	MH	MH	MH	ML	→	H	0.04762
17.	M	MH	M	L	→	H	0.04762
18.	M	M	MH	M	→	M	0.03874
19.	M	M	MH	ML	→	MH	0.03435
20.	MH	MH	MH	M	→	MH	0.03053
21.	M	MH	MH	MH	→	MH	0.03053

dataset, the generated fuzzy rules were applied on the data instances in the dataset and the estimated health-shock was compared with the actual instance labels as shown below:

$$acc_j = \frac{1}{h} \sum_{v_i, y_i \in D_k} \sigma(v_i, y_i) \tag{9}$$

where D is the full data set of size h, $\sigma(v, y) = 1$ if $v = y$ and 0 otherwise, v_i is the predicted value of the instance i and y_i is the actual value of the instance i, where $i = 1toh$. Fig. 17 shows that the fuzzy rule based system achieved 97% modeling accuracy in classifying health-shocks risk correctly on the full data set.

Similarly, in order to evaluate the prediction accuracy of the fuzzy based system on unseen data, k-fold cross-validation was used [57]. In k-fold cross-validation, the dataset D is divided into k equal size (of size h items) subsets called folds. The validation process is then carried out for k iterations and in each iteration $j : 1tok$, the subset k_j is held out and called hold-out set D_h. The rest of the subsets are grouped in a training set $D_t = D - k_j$. The accuracy of the model for each fold k_j was calculated as:

$$acc_j = \frac{1}{h} \sum_{v_i, y_i \in D_h} \sigma(v_i, y_i), \tag{10}$$

where $\sigma(v, y) = 1$ if $v = y$ and 0 otherwise, v_i is the predicted value of the instance i and y_i is the actual value of the instance i. The final accuracy of the model is calculated by taking the average of the resulting accuracy values for the all iterations as shown below:

$$ACC = \frac{1}{h} \sum_{j=1}^{k} acc_j, \tag{11}$$

For the evaluating the fuzzy rule based system, a 5-fold cross validation was applied. Here, dataset D was partitioned into five folds, each representing 20% of the dataset. For each iteration, one of the fold subsets was held out and the system was trained on the other folds representing the remaining 80% of the dataset in order to extract a set of weighted fuzzy rules. Here, it is worth mentioning that it is standard practice in evaluating machine learning approaches

to split the dataset where 80% comprises of the training data and 20% comprise of the hold-out set.

The resulting fuzzy rules of the training process were used to build a fuzzy system to classify each instance in the hold-out sets D_h. The resulting classifications were compared with the actual associated linguistic labels for the data instances in the hold-out sets. This process was repeated five times for the five different folds where for each fold, the model's classification accuracy was calculated. Fig. 17 presents the prediction accuracies for each fold as well as the average prediction accuracy. It can be seen from the table that the highest accuracy of 96% was for the fold three at $k = 3$ and the lowest accuracy of 73% was for the fold two at $k = 2$. The average accuracy of the system was 89%, which shows a relatively good initial prediction accuracy using the proposed fuzzy modeling technique. Given the flexibility of the fuzzy rule based model, it is expected that this performance can be further improved with more data which can be summarised into more accurate models for predicting health-shocks.

6. Conclusions and Future Work

Currently, there is no publicly available dataset that can help to understand and monitor the health-shocks. The main goal of this research was to develop cloud based infrastructure to capture the first contextaware healthcare dataset based on the socio-economic, cultural, and geographic norms of Pakistan. The aim was to then analyse and model such a dataset to understand the relationships between socio-economic, demographic, and geographical conditions and their impact on health. The availability and interpret-ability of this data can be helpful to governments to determine policies for general practitioners and NGOs, in order to start community based health programs.

In this paper, we have developed a cloud based analytics framework for profiling and predicting health-shocks. The framework facilitates the collection of population based socio-economic, environmental and health related data using both manual and electronic survey tools which can be easily deployed in remote

and rural areas. Large amounts of data can be continuously collected for storage, processing and retrieval on the cloud. The framework was used to carry out a user study comprising of collecting a unique dataset from 1000 households belonging to 29 villages in rural areas of Pakistan. The data consisted of 47 features, which were pre-processed using health experts to derive four measures related to living standards, health risk, accessibility to health facilities and income allocation labeled with a level of health-shocks incurred.

The pre-processed data was used to generate a fuzzy rule based classification model for the prediction of health-shocks using the fuzzy LS technique which generated an interpretable rule based model to visualize and predict the magnitude of health-shocks experienced by individuals. The extracted fuzzy rules used quality measures that determined the strength of each rule in its ability to model the data, which provided stakeholders with a means of ranking and interpreting the quality of the rules. The generated fuzzy model was evaluated based on the interpretability of the rules in their ability effectively profile the factors affecting different levels of health-shocks and their modeling and classification accuracy for predicting health-shocks levels from unlabeled data. The results have shown that the generated rules provide sensible and meaningful profiles explaining the factors, corresponding to various levels of health-shocks that was also accepted by health experts with knowledge of the health issues affecting the sampled populations. The prediction accuracies of the fuzzy model based on a k-fold cross-validation of the data samples shows that the applied LS approach is also able to achieve good prediction accuracies which can be improved with larger data samples.

The paper has demonstrated that large-scale health data analytics facilitated through the cloud computing will not only help healthcare professionals to create and conduct surveys with minimal human and financial resources but will also help them to understand the socio-economic, environmental and cultural norms that directly or indirectly cause health-shocks. This study is one of the first initiatives to analyze and understand the healthcare system and the occurrence of health-shocks in rural and tribal areas of Pakistan. In the future, we

would like to extend our study to form Pakistan's first publicly available health informatics tool that can be helpful to government and healthcare professionals to form policies and healthcare reforms.

References

[1] E. Massad, N. R. S. Ortega, L. C. Barros, C. J. Struchiner, Fuzzy Logic in Action: Applications in Epidemiology and Beyond, Springer, 2008.

[2] J. Andreu-Perez, C. C. Y. Poon, R. D. Merrifield, S. T. C. Wong, G. Z. Yang, Big Data for Health, IEEE Journal of Biomedical and Health Informatics 19 (4) (2015) 1193–1208.

[3] S. Mahmud, R. Iqbal, F. Doctor, An integrated framework for the prediction of health shocks, in: Proceedings of The 2nd International Conference on Applied Information and Communications Technology, ICAICT, 2014.

[4] E. Binnendik, R. Koren, D. M. Dror, Hardship financing of healthcare among rural poor in Orissa, India, BMC Health Services Research 12 (2012) 23.

[5] A. Balasubramanian, Changes in health-care financing and organization, Robert Wood Johnson foundation, 2009.

[6] R. Victor, F. Jeffrey, E. Harris, Review: Who Shall Live? Health, Economics, and Social choice, The Bell Journal of Economics 7 (1976) 340–343.

[7] R. Chandrashekar, M. Kala, D. Mane, Integration of Big Data in Cloud computing environments for enhanced data processing capabilities, International Journal of Engineering Research and General Science 3 (2015) 240–245.

[8] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, M. Zaharia, A view of cloud computing, Communications of the ACM 53 (4) (2010) 50–58.

[9] P. Priyanga, V. P. MuthuKumar, Cloud computing for healthcare organization, International Journal of Multidisciplinary Research and Development 2 (2015) 487–493.

[10] J. Bryce, F. Arnold, A. Blanc, A. Hancioglu, H. Newby, J. Requejo, T. Wardlaw, Measuring Coverage in MNCH: New Findings, New Strategies and Recommendations for Action, PLOS Medicine.

[11] I. Rudan, J. Lawn, S. Cousens, A. K. Rowe, C. Boschi-Pinto, L. Tomaskovi, W. Mendoza, C. F. Lanata, A. Roca-Feltrer, L. Carneiro, J. A. Schellenberg, O. Polasek, M. Weber, J. Bryce, S. Morris, R. E. Black, H. Campbell, Gaps in policy relevant information on burden of disease in children: A systematic review, The Lancet 365 (2005) 2031–2040.

[12] H. K. Herb, A. C. Chandran, I. Rudan, A. H. Baqui, Care seeking for neonatal illness in low and middle income countries: A systematic review, PLOS Medicine.
URL http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001183

[13] M. H. V. Velthoven, J. Car, Y. Zhang, A. Maruic, mHealth series: New ideas for mHealth data collection implementation in low and middle income countries, PLOS Medicine 3.

[14] A. Abraham, B. Nath, Hybrid Intelligent Systems: A Review of Decade of Research, Technical report, School of Computing and Information Technology, Faculty of Information Technology, Monash University, Australia (2007).

[15] P. Szolovits, R. S. Patil, W. B. Schwartz, Artificial Intelligence in Medical Diagnosis, Journal of Internal Medicine 108 (1988) 80–87.

[16] P. Szolovits, Uncertainty and Decision in Medical Informatics, Methods of Information in Medicine 34 (1995) 111–121.

[17] M. R. Howlader, Analysing the socio-demographic variables impact on health status of Bangladesh, Social Science Research Network.
URL http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2294871

[18] K. William, Medicines Dilemmas: Infinite need versus finite resources, 1994.

[19] R. Iqbal, N. Shah, A. James, J. Duursma, : From work practices to redesign for usability, Journal of Expert Systems with Applications 38 (2011) 1182–1192.

[20] J. Bernstein, Impact of the economy on health-care, Changes in Health Care Financing and Organization (HCFO) (2009) 1–8.

[21] J. P. Smith, Consequences and predictors of New Health Events: Analysis in the Economics of Aging, Working paper no. 10063, Chicago: University of Chicago Press (2007).

[22] M. Suhrcke, M. McKee, R. S. Arce, S. Tsolova, J. Mortensen, The Contribution of health to the economy in the European Union, Public Health 120 (2006) 94–1001.

[23] B. Lindberg, C. Nilsson, D. Zotterman, S. Sderberg, L. Skr, Using Information and Communication Technology in Home Care for Communication between Patients, Family Members, and Healthcare Professionals: A Systematic Review, International Journal of Telemedicine and Applications (2013) 1–31.

[24] E. Ammenwerth, J. Brender, P. Nykanen, H. U. Prokosch, M. Rigby, J. Talmon, Visions and strategies to improve evaluation of health information systems Reflections and lessons based on the HIS-EVAL workshop in Innsbruck, International Journal of Medical Informatics 73 (2004) 479–491.

[25] B. Callaway, V. Ghosal, Adoption and Diffusion of Health Information Technology: The Case of Primary Care Clinics, CESifo Working paper:

Industrial Organization No. 3925, Leibniz Institute for Economic Research at the University of Munich (2012).

[26] B. Hayes, Cloud computing, Communications of the ACM 51 (7) (2008) 9–11.

[27] V. Chang, An overview, examples and impacts offered by Emerging Services and Analytics in Cloud Computing, International Journal of Information Management.
URL http://www.sciencedirect.com/science/article/pii/S0268401215000924

[28] V. Chang, G. Wills, A Model to Compare Cloud and non-Cloud Storage of Big Data, Future Generation Computer Systems.
URL http://eprints.soton.ac.uk/382709/

[29] V. Chang, Y. H. Kuo, M. Ramachandran, Cloud Computing Adoption Framework? A security framework for business clouds, Future Generation Computer Systems.
URL http://eprints.soton.ac.uk/382704/

[30] A. Manekar, G. Pradeepini, A Review on Cloud Based Big Data Analytics, ICSES Journal on Computer Networks and Communication 1 (2015) 6–9.

[31] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. H. Katz, A. Konwinski, G. Lee, D. A. Patterson, A. Rabkin, M. Zaharia, Above the Clouds: A Berkeley View of Cloud Computing, Tech. rep., University of Berkeley (2009).

[32] D. P. Shukla, S. B. Patel, A. K. Sen, A Literature Review in Health Informatics Using Data Mining Techniques, International Journal of Software and Hardware Research in Engineering 2 (2) (2014) 123–129.

[33] K. Choi, S. Chung, H. Rhee, Y. Suh, Classification and Sequential Pattern Analysis for Improving Managerial Efficiency and Providing Better Medical Service in Public Healthcare Centers, Healthcare Informatics Research 16 (2) (2010) 67–76.

[34] J. Liu, Z. H. Tang, F. Zeng, Z. Li, L. Zhou, Artificial neural network models for prediction of cardiovascular autonomic dysfunction in general Chinese population, BMC Medical Informatics and Decision Making 13 (80) (2013) 1–7.

[35] W. Yu, T. Liu, R. Valdez, M. Gwinn, M. J. Khoury, Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes, BMC Medical Informatics and Decision Making 10 (16) (2013) 1–7.

[36] M. Jain, P. Dua, S. Dua, W. J. Lukiw, Data Adaptive Rule-based Classification System for Alzheimer Classification, Journal of Computer Science and Systems Biology 6 (5) (2013) 291–297.

[37] G. Fleming, M. V. D. Merwe, G. McFerren, Fuzzy expert systems and GIS for cholera health risk prediction in southern Africa, Environmental Modelling and Software 22 (2007) 442–448.

[38] X. Djam, G. M. Wajiga, Y. H. Kimbi, j. . I. y. . . v. . . n. . . p. . . N. V. Blamah, title = A Fuzzy Expert System for the Management of Malaria.

[39] F. Doctor, C. H. Syue, Y. X. Liu, J. S. Shieh, R. Iqbal, Type-2 fuzzy sets applied to multivariable self-organizing fuzzy logic controllers for regulating anesthesia, Applied Soft Computing.
URL http://www.sciencedirect.com/science/article/pii/S156849461500647X

[40] P. R. Innocent, R. I. John, Computer Aided Fuzzy Medical Diagnosis, Information Science 162 (2004) 81–103.

[41] M. Poongodi, L. Manjula, S. Pradeepkumar, M. Umadevi, Cancer prediction technique using fuzzy logic, International Journal of Current Research 4 (2012) 106–110.

[42] A. Altunkaynak, M. zger, M. akmakci, Water Consumption Prediction of Istanbul City by Using Fuzzy Logic Approach, Water Resources Management 19 (5) (2005) 641–654.

[43] F. Doctor, R. Iqbal, R. N. Gorgui-Naguib, A fuzzy ambient intelligent agents approach for monitoring disease progression of dementia patients, Journal of Ambient Intelligence and Humanized Computing 5 (1) (2014) 147–158.

[44] J. Mendel, Uncertain Rule-Based Fuzzy Logic Systems: Introduction and New Directions, 1st Edition, Prentice Hall PTR, Prentice Hall Inc, 2001.

[45] GoP, Pakistan poverty alleviation fund (PPAF) livelihoods programme: Implementation manual.
URL http://www.ppaf.org.pk/What_We_Do_detail.aspx?component_id=1

[46] GoP, Pakistan poverty scorecard: Assessment of measuring impact of ppaf interventions using pakistan poverty scorecard (2012) 1–33.

[47] M. Schreiner, A simple poverty scorecard for pakistan, Journal of Asian and African Studies 45 (3) (2010) 326–349.

[48] GoP, Government of Khyber Pakhtunkhwa: Khyber Pakhtunkhwa Health Sector Situation Analysis, Tech. rep. (2010).
URL http://www.healthkp.gov.pk/downloads/HSSA-KP.pdf

[49] GoP, Rapid Assessment Survey of Children's Involvement in Worst Forms of Child Labour in District Haripur, KPK, Tech. rep. (2013).
URL http://www.khyberpakhtunkhwa.gov.pk/cms/pages/139583029953328b426e2f0.pdf

[50] D. Wu, J. M. Mendel, J. Joo, Linguistic Summarization Using If-Then Rules, in: Proceedings of the IEEE International Conference on Fuzzy Systems, FUZZ-IEEE, 2010, pp. 1–8.

[51] F. Doctor, R. Iqbal, An Intelligent Framework for Monitoring Student Performance Using Fuzzy Rule-Based Linuistic Summarization, in: FUZZ-IEEE, 2012, pp. 1–8.

[52] L. A. Zadeh, Fuzzy sets, Information and Control 8 (3) (1965) 338–353.

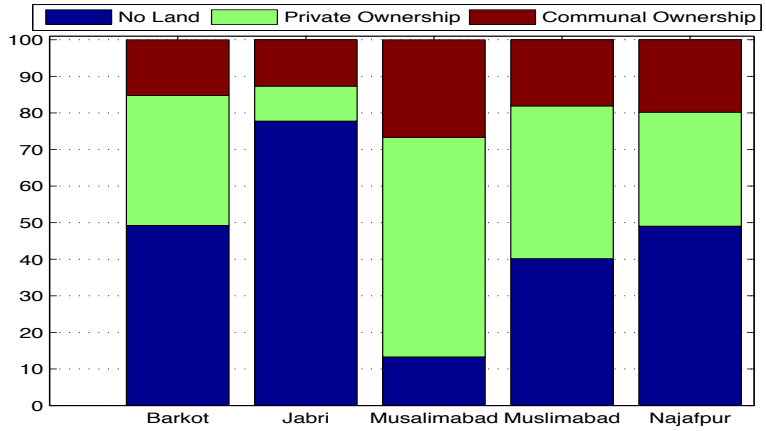
[53] L. X. Wang, The MW method completed: A flexible system approach to data mining, IEEE Transactions on Fuzzy System 11 (6) (2003) 768–782.

[54] F. Doctor, H. Hagras, V. Callaghan, An Intelligent Fuzzy Agent Approach for Realising Ambient Intelligence in Intelligent inhabited environments, IEEE Transactions on System, Man and Cybernetics 35 (1) (2005) 55–65.

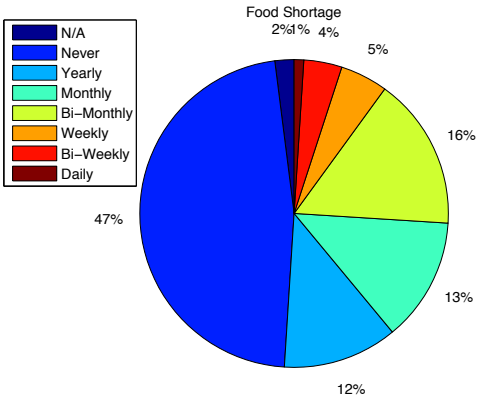
[55] H. Ishibuchi, T. Yamamoto, Rule Weight Specification in Fuzzy Rule-Based Classification Systems, IEEE Transactions on Fuzzy Systems 13 (4) (2005) 428–435.

[56] Gacto, M. Jos, R. Alcal, F. Herrera, Interpretability of linguistic fuzzy rule-based systems: An overview of interpretability measures, Information Sciences 181 (20) (2011) 4340–4360.

[57] S. Arlot, A. Celisse, A survey of cross-validation procedures for model selection (2009).

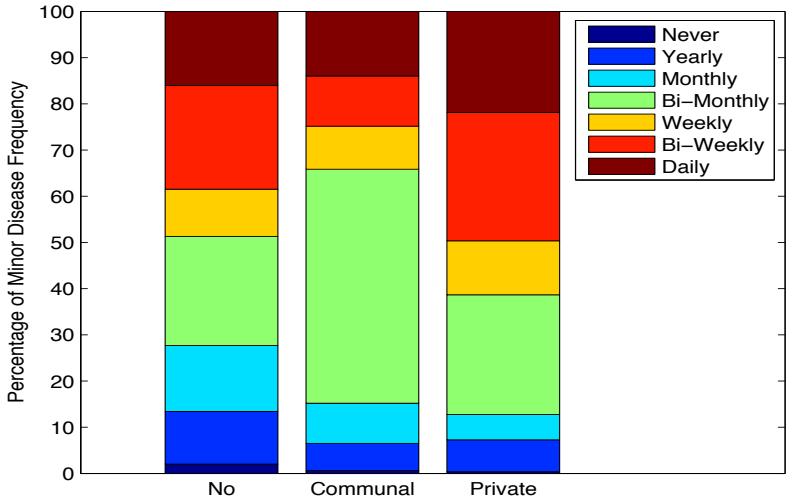


(a)

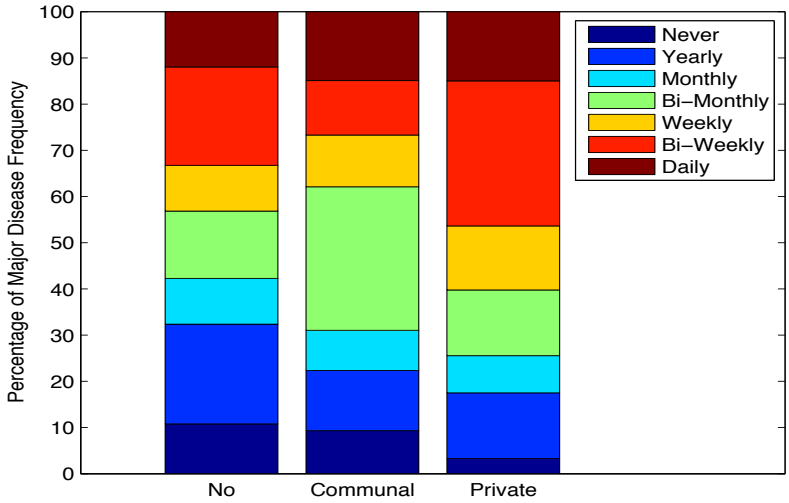


(b)

Figure 13: a) Land ownership. b) Food shortage.



(a)



(b)

Figure 14: (a) Frequency of minor disease versus toilet facility. Here, communal represents a toilet shared by more than 3 people. b) Frequency of major disease versus toilet facility.

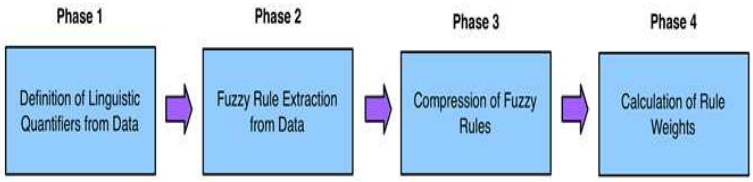


Figure 15: Flow diagram showing the phases of the fuzzy LS approach.

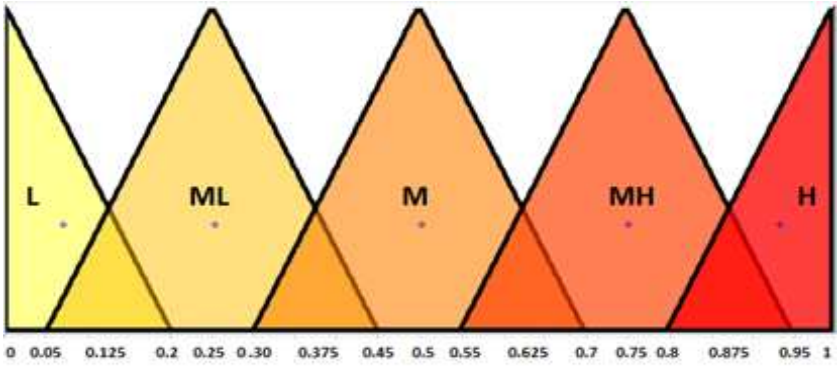


Figure 16: Fuzzy Sets for Input and Output Variables.

Figure 17: Overall model accuracy for prediction health-shocks on seen and unseen data.