Diabetes Control in China

Building a predictive machine learning model for diabetes detection in Mainland China based on living circumstances

Bachelor's thesis

Valkeakoski, International Business

Spring Semester 2019, 2019.

Nico Kling

HÄMEEN AMMATTIKORKEAKOULU
HÄME UNIVERSITY OF APPLIED SCIENCES

ABSTRACT

International Business
Valkeakoski

Author	Nico Kling	**Year** 2019
Subject	Diabetes Prevention in China	
Supervisor(s)	Sajal Kabiraj, Martina Heigl-Murauer	

ABSTRACT

In 2016, Healthy China 2030 has been announced and with it its five specific goals: controlling major risk factors, increasing the capacity of the health service, enlarging the scale of the health industry, perfecting the health service system and improving the health nationwide. Nevertheless, Diabetes continues to be a leading public health challenge in China. In this thesis, the author explores reasons for the rapid diabetes prevalence, as well as how a simple supervised machine learning model build in Python, based on the China Health and Retirement Longitudinal Study (CHARLS), can predict the risk of a Chinese citizen aged 45 and above having diabetes. Three different algorithms, Random Forest, Support Vector Machine and Logistic Regression are compared. The model is built with the help of SciKit-learn and imbalanced-learn. Findings obtained are correlations between diabetes and dyslipidemia, as well as correlations among diabetes and education level among other things. The most suited algorithm for prediction is Support Vector Machine, after introducing over-sampling. Generally, the findings demonstrate that the blueprint Healthy China 2030 is a thought-out and needed strategy change.

Keywords Healthy China 2030, Diabetes, Predictive Modelling, Supervised Machine Learning, China Health and Retirement Longitudinal Study, Health Care Management

Pages 66 pages including appendices 25 pages

CONTENTS

1 INTRODUCTION

1.1 Background

China´s role as a leader in the world's economy is in fact, nothing new with historical sources stating its technological leadership going back to the eighth century. However, during the mid-eighteenth century, China´s economy started stagnating because of the lack of industrialization (Maddison, 2007). The economic instability continued throughout the first half of the next century and only came to a halt 30 years after the proclamation of the People´s Republic of China, through the economic reforms under Deng Xiaoping, starting in 1979.

His reforms, a shift from central-planned economy to a more market-based one, enabled an economic miracle and allowed the economy to become the fastest growing one in the world (Witt, 2018). Nowadays, the People´s Republic of China, having a population of almost 1.4 billion, is the country with the second biggest gross domestic product (GDP) worldwide. Its GDP at purchasing power parity is the largest in the world, with an estimated 25.1 trillion USD in 2018 (International Monetary Fund, 2018).

Despite this, the Chinese health care sector, especially when comparing the rural and urban areas, did not evolve at the same rate. Paradoxically, health care is an integral part of the Chinese culture, with the beginning of Chinese medicine and health care going back as far as 2700 BC, to the legendary emperor Shennong who is said to have introduced acupuncture (Ho & Lisowski, 1997).

De facto, the Chinese health care sector was a subject of domestic and international criticism for its poor allocative efficiency performance in the early beginning of the millennium (Blumenthal & Hsiao, 2005). As the Chinese government under Xi Jinping fully recognized the need for a strategic shift, it launched the health care system reform in 2009. The overall goal of health care reform was the establishment and improvement of the basic health care system, covering urban as well as rural residents, and providing the people with affordable, secure, convenient and efficient health care services until 2020 (Ling & Hongqiao, 2017).

The health care system reform was further supported by the 12th 5-Year Plan for economic and social development of the people´s republic in China. Its healthcare development was split into three sequences.

The first sequence, taking place between 2009 and 2012, set the focus of the government on the health care reform. It was succeeded by the reform of the wasteful and inefficient public hospitals and financial investments into the health care sector between 2012 and 2015. The reform´s final sequence began in 2016 after the Chinese government announced the 13th

5-Year Plan for economic and social development of the people´s republic in China. It set the emphasis of the health care reform on prevention, as well as reconfirming the coverage of all basic health care services.

The plans crucial subjects include improving the medical insurance system for all citizens, improving major disease prevention and treatment, improving maternal and infant health care and childbirth services, optimizing the structure of the medical institution systems, improving the traditional Chinese medicine, implementing a fitness strategy and lastly implement a food and medicine safety strategy. Adding to that, the government introduced the Basic Health Care Law, which would define essential elements of the health care system (Central Committee of the Communist Party in China, 2016).

As a successor to the previous reform, the Chinese government announced a blueprint of Healthy China 2030 on October 25, 2016. It is China´s newest national medium- and long-term strategic plan for health care and made public health a priority for the economic and social development of the People´s Republic of China.

1.2 Purpose, State of the Art and Objectives

The research question the author seeks to answer is “How can Machine Learning, as a possible part of scientific development in Healthy China 2030, be utilized to reduce the prevalence of Diabetes among Chinese citizen aged 45 and above?”
The reason for that is that China has an ageing population, with an estimated average age of 50 by 2070, and a steadily increasing elderly dependency (Statista, 2019). Adding to that, the People´s Republic of China faces some healthcare problems, such as the rapid prevalence of diabetes (Xu et al., 2013).

This is relevant as diabetes and other chronic diseases increase the risk of disability or even premature death, if not treated properly and are thus leading to a burden on both the economy, as well as the society of the People´s Republic of China (Soumya & Srilatha, 2011; Papatheodorou et al., 2015; Nather et al., 2008). Also, health is the basis for human - and socio-economic development, and to support a prospering economy and society in the People´s Republic of China, the researcher decided to make use of the rapid advancements in technology that are occurring today, namely through Machine Learning, which is seeing many different use cases in today’s business world. One of its applications is the prediction of diabetes as a part of healthcare management.

In medicine, diabetes is diagnosed through fasting blood glucose, glucose tolerance and random blood glucose levels (Iancu et al., 2008; Cox and Edelmann, 2009; American Diabetes Association, 2012). Recent studies showed that Machine Learning can be used to make a preliminary judgment about diabetes through their daily physical examination data, as well as being a reference for doctors (Lee and Kim, 2016; Alghamdi et al., 2017; Kavakiotis et al., 2017).

Numerous different algorithms were used in recent research papers to predict diabetes. Zou et al. (2018) used Neural Networks, Random Forests and Decision Trees, and found out that fasting glucose is the most important index for prediction. Kavakiotis et al. (2017) conducted a study to create a systematic review of machine learning, data mining techniques and tools in diabetes research. The result showed that 85 percent of the approaches were supervised ones, while only 15 percent were unsupervised. Support vector machines proved to be the most successful algorithms in diabetes prediction. While Georga et al. (2013) used support vector machines and focused on glucose, Ravazian et al. (2015) used logistic regression for different onset type 2 diabetes predictions. Duygu and Esin (2011) focused on dimension reduction and feature extraction through Linear Discriminant Analysis.
Additionally, more studies emerge that use ensemble methods to improve accuracy (Kavakiotis et al., 2017).

Nearly all recent papers focus on fasting glucose level as their main predictor. In order to gain new insight, the author of this thesis uses predictors such as demographic information or activity level to predict diabetes. The reason for that is while diabetes type 1 is mostly caused by genes and environmental factors such as viruses, the most common form of diabetes, diabetes type 2, is caused mostly by high-risk behaviours, such as smoking, alcohol consumption or poor diets, or environmental factors such as air pollution (Batis et al. 2014; Chan et al., 2009). Therefore, a machine learning classifier based on the living circumstances of Chinese citizen can be built and used for generalization in order to predict diabetes risk patients.

The algorithms used for creating the prediction model, Random Forests, Linear Regression and Support Vector Machines, are chosen after discussing constraints such as processing power or the speed of prediction, as well as type of problem such as classification or regression.
Random forests are, as ensembled decision trees, through their classification power a favoured algorithm used in medical machine learning. Linear regression is an algorithm used to find the relationship between variables and forecasting. Therefore, it also used in this thesis. Additionally, as support vector machines are the most widely used algorithms in diabetes prediction, it is the final algorithm used.

Thus, the objects of this interdisciplinary thesis are to explain problems in the Chinese healthcare sector, especially the rapid prevalence of diabetes in China, to examine the policy changes of Healthy China 2030 as a possible control solution, and lastly also wishes to offer an example for possible helpful scientific development in the form of a supervised machine learning model for diabetes detection, that could help reduce the burden on society and economy of the People´s Republic of China.

1.3 **Thesis structure**

The thesis consists of seven sections: Introduction, Chinas Healthcare Development, Machine Learning, Analysis and Presentation of Diabetes Data, Supervised Machine Learning Solution, Recommendations and Limitations, Conclusion and Acknowledgment.
The first chapter served as an introduction to the thesis and briefly explained the history of China´s health care sector, as well as the objectives, purpose, state of the art and structure of this thesis. The second chapter offers basic insights and information on the health care reform Healthy China 2030, as well as the causes for the rapid prevalence of diabetes, and general healthcare problems that are contributing to the problems in the Chinese healthcare system. The third chapter gives the reader an overview of the theory of supervised machine learning, specifically based on three different classification algorithms, Logistic Regression, Support Vector Machines and Random Forests. The fourth chapter digs deeper into the reasons for diabetes for participants of CHARLS.
The fifth chapter explains the results of the supervised machine learning classifier and different statistic results. The seventh chapter discusses recommendations as well as limitations of the thesis. The final seventh chapter will conclude the thesis.

2 CHINAS HEALTHCARE DEVELOPMENT

The introduction of barefoot-doctors and ambitious health care reforms in the past culminated in a significant decrease in mortality, while steadily increasing life expectancy in China (Yang et al., 2008; Babiarz et al., 2015). Chinese life expectancy has increased to 76.2 years in 2015, which was 4.4 years higher than the international average of 71.8 (Tan et al., 2018). This development, as well as the in 2015 abandoned one-child policy, led to the ageing of the population.
In 2013, 15 percent of the people in China were 60 or older, while the amount is supposed to double until 2030 (China National Bureau of Statistics, 2014). This leads to issues, for example to increasing costs related to health and long-term care, as older people are more likely to suffer from health problems such as weaker functionality (Mihovska et al., 2014; Goodpaster et al., 2006). In the following section, China´s newest healthcare reform, Healthy China 2030, will be discussed, as well as general healthcare system problems, the rapid prevalence of diabetes.

2.1 General Healthcare System Problems

Although the Chinese government initiated a strategic drift in healthcare through Healthy China 2030, which will be discussed later in this chapter, it is still important to discuss the challenges that are contributing to the difficulties within the healthcare system. In total, there are three challenges: demographic and epidemiological trends, the quality of care and lastly internal system factors.

Demographic and epidemiological trends

Communicable diseases, injuries and nutritional as well as maternal conditions only accounted for 15 percent of deaths in 2014. Non-communicable diseases are responsible for 85 percent, with cancer alone being responsible for over 66 percent (World Health Organization, 2014). Non-communicable diseases are estimated to double or triple over the next 20 years for people above 40 (Wang et al., 2011).

Disparities in the quality of primary care

The disparities in the quality of care perceived by patients among different level of providers are one of the main reasons for the inability to re-direct patients to primary care facilities (Yang et al., 2008).

Another reason is the scarcity of competent health care professionals, especially in the urban regions, as doctors training varies strongly between the different levels of care. This scarcity further amplifies the disparity and leads to unnecessary and avoidable hospitalizations in the primary care level. A problem in all facilities is the drug over-prescription (World Bank, 2016; Qu et al., 2018).

Internal System Factors

The disparities in the quality of care in primary care, as well as the demographic and epidemiological trends, led to a decline of 6 percent in the number of primary care providers between 2002 and 2013, and an increase by 82 and 29 percent respectively in the number of tertiary and secondary hospitals (World Bank, 2016). Healthcare practitioners with a high-quality education are also moving away from primary care and can be found concentrated in hospitals (Sun, et al., 2015). This trend further increases the client/practitioner ratio in the tertiary and secondary healthcare facilities.

Another internal system factor that negatively affects the healthcare system is institutional fragmentation. As too many governmental agencies are involved in the health care sector, with each one trying to reach its own bureaucratic goals, the development gets consequently hindered (Qian, 2015).

2.2 Prevalence of Diabetes in China

After the rapid economic growth, the changes in lifestyle and an increasing life expectancy, cardiovascular disease has become the leading cause of death in China (He et al., 2005; Zhao et al., 2019). Additionally, the prevalence of diabetes has reached more than 10 percent in Chinese citizen. It is increasing to more than 20 percent if aged 60 or above, but especially worrisome is the prevalence of diabetes in young people, which amounts to nearly 5 percent (Wang et al., 2013; Hu & Jia, 2018). With that, China has the largest number of people with diabetes worldwide. As diabetes is a major risk factor for cardiovascular diseases, this is especially concerning (Gu et al., 2003; Luk et al., 2014).

While diabetes type 1 is mostly caused by genes and environmental factors such as viruses, but the most common form of diabetes, diabetes type 2, is caused mostly by high-risk behaviours. Recent research showed that there are three major risk factors for type 2 diabetes: urbanisation, obesity and diet (Ma et al., 2014; Whiting et al., 2010).

Urbanization

Half of China´s population is nowadays living in cities, compared to 20 percent in the 1970s (Peng, 2011; Shin, 2015). Associated lifestyle changes, such as reduced physical activity or changes in dietary behaviour, are main reasons for an increase in obesity in China (Gong et al., 2012; Wu, 2006). Adding to that, the concentration of health care professionals in cities led to an especially increasing life expectancy in urban China.

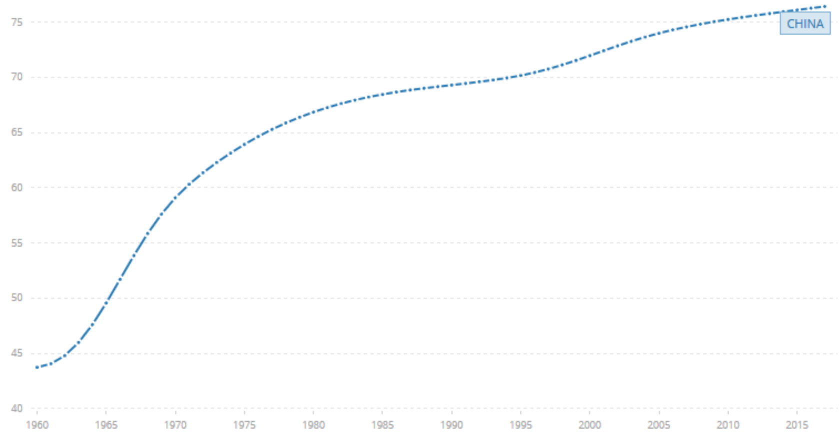

Figure 1 Life expectancy at birth in China

Decreasing birth rates, which can be explained by both the nowadays abolished one-child policy and urbanization, fundamentally changed the population structure, and simultaneously increases the people at risk of diabetes – as already mentioned, studies show that older people are at a higher risk of suffering from diabetes.

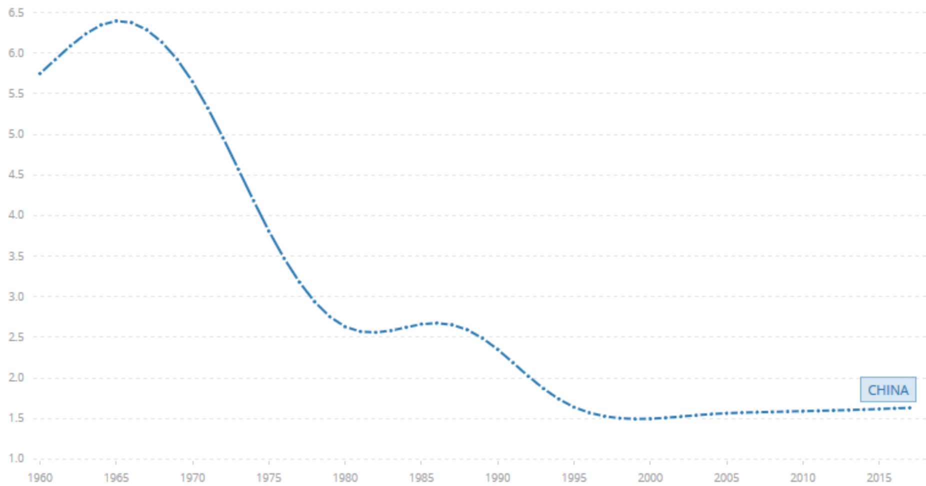

Figure 2 Birth rates in China

Obesity

Many factors, such as urbanization, changes in the diet or reduced physical activities, contribute to the increasing obesity in China. When compared to the WHO definitions obesity (BMI ≥30 kg/m²), the prevalence increased roughly 6-times between 1990 and 2016 for men, from 0.9 percent to 5.9

percent as seen in figure 3, and roughly 4-times for women, from 1.8 percent to 6.5 percent, as seen in figure 4. The tendency in the Asian population to have low muscle mass, coupled with visceral adiposity and the resulting metabolically obese phenotype could be an explanation for the increasing prevalence of diabetes in China (Ma et al., 2014).

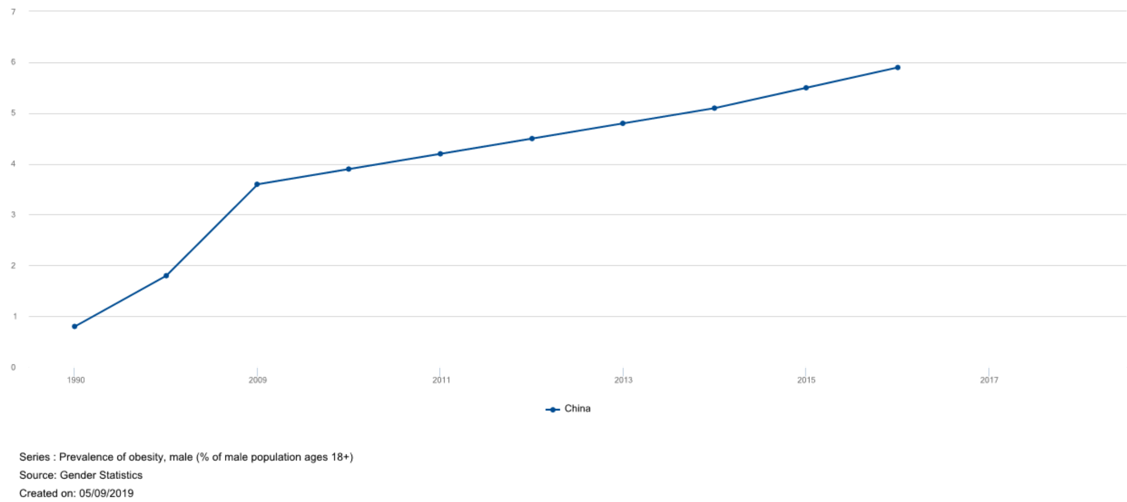

Figure 3 Male obesity in China

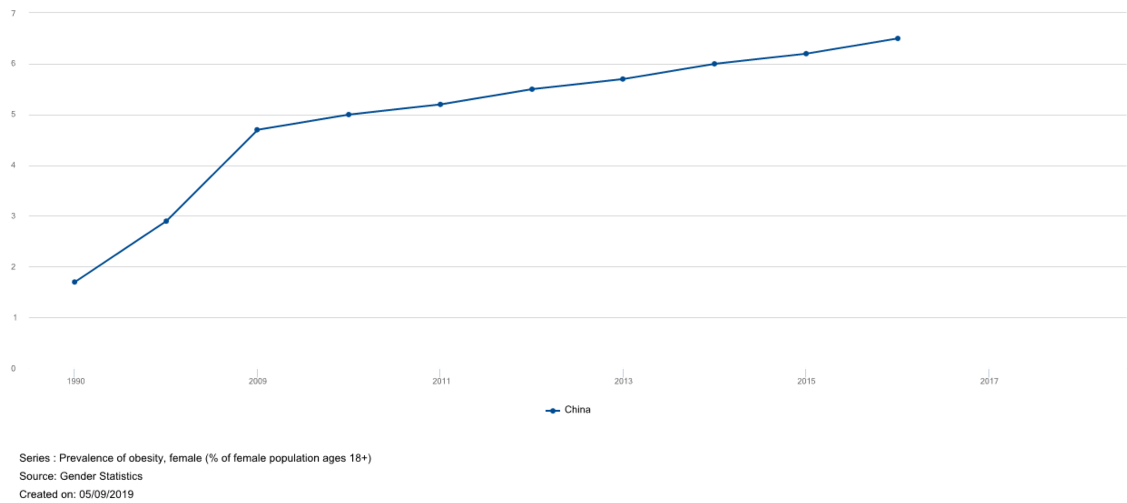

Figure 4 Female obesity in China

Diet

Healthy diets, which consist for example of fibre-rich foods, unrefined grain or unsaturated fats, is shown to decrease the risk of diabetes type 2 (Alhazmi et al., 2014). Simultaneously, following a so-called Western

pattern diet, characterized by high amounts of dietary fat, refined grain, high-sugar drinks or pre-packaged food, increases the risk (Kant, 2004).
In the last decades, the eating habits of the Chinese population were more and more characterized by the latter, with the growing influence of the West on the East (Hu et al., 2011). Adding to that, the high consumption of white rice, and its associated high glycemic index, might also increase the risk of diabetes in Chinese citizen (Villegas et al., 2007).

2.3 **Healthy China 2030**

As the Chinese government understood the need for a strategic drift in healthcare, the blueprint of Healthy China 2030, was passed on October 25, 2016. It is China´s national medium- and long-term strategic plan for health care and aims to make public health a priority for the economic and social development of the People´s Republic of China. This section is a summary of the blueprint, based on the "Outline of the Healthy China 2030 Plan" written by Ning Zhuang, the Deputy Director-General of the Department for Healthcare Reform in P.R. China. Its framework can be seen in figure 5.

<table>
<tr><td colspan="5">Goal
Put health on the priority list of development to a strategic position; promote the concept of health in the whole process of public policy implementation;enable everyone to be involved health and everyone to share health care services;focus on the health of all the people all their life in China.</td></tr>
<tr><td colspan="5">Principles</td></tr>
<tr><td>Health Priority</td><td>Reform and Innovation</td><td colspan="2">Scientific Development</td><td>Justice and Equity</td></tr>
<tr><td colspan="5">HC 2030: China's vision for health care</td></tr>
<tr><td>1. Health Level</td><td>2. Healthy life</td><td>3. Health Services and Health Security</td><td>4. Environmental Health</td><td>5. Health Industry</td></tr>
<tr><td colspan="5">The 13 Core Indicators</td></tr>
<tr><td>A. The average life expectancy
B. The mortality rate of infants
C. The mortality rate of children below 5 years of age
D. The mortality rate of pregnant women and mortality
E. The proportion of those meeting the national physique determination standard among urban and rural residents</td><td>A. The level of health literacy among residents
B. The number of people taking part in physical exercise</td><td>A. Premature mortality as a result of major non–communicable diseases
B. The number of registered doctors per 1000 residents and registered nurses per 1000 residents
C. The proportion of personal health spending in the total health expenses</td><td>A. Good air quality rate of all cities at prefecture level or above
B. The rate of surface water quality better than III</td><td>A. The total investment scale of health services</td></tr>
</table>

Figure 5 Healthy China 2030: A Vision for Health Care - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/The-framework-of-the-Healthy-China-2030-vision_fig1_317128652 [accessed 9 May, 2019]

Principles and Goals

Four principles are stated: health will be prioritized and set up in a strategic position in the public policy implementation. Government-led reforms and innovations, allowed through market mechanism, will help to speed up the improvement of people´s help. Through scientific development, the

Chinese healthcare sector is supposed to prevent first while supporting traditional Chinese medicine as well as Western medicine. It is also stated that the overall healthcare service mode is bound to change. The last principle, equity and justice, promotes equal access to basic public health services for rural areas, as well as maintaining public welfare.

The reform aims to reach five specific goals: improving the general health status, supporting a healthy life, optimizing healthcare services and security, building a healthy environment and further developing the healthcare industry.

Supporting a healthy life

Improving the health literacy among urban and rural citizens through health mentoring and interventions to high-risk groups and households, as well as promoting health education in school through integrating it into the national curriculum is supposed to improve the national health education. This should further support the encouragement of healthy habits, as knowledge is the key to the development of well-balanced diets. Tightening tobacco and alcohol control is supposed to reduce unhealthy lifestyles. Intervening in mental health disorders through an increasement of competence identifying such and reducing drug abuse as well as unsafe sexual behaviours are also tasks set in the blueprint.

Optimizing Healthcare Services and Security

Major illnesses are to be prevented and to be controlled, management of family planning services are to be reformed and improved, while also ensuring equity of primary healthcare services among urban and rural residents. Medical care delivery is to be improved, through increasing the nurse/permanent resident rate, and building more primary care facilities in close proximity to all communities. Medical care supply is to be improved by integrating curative, rehabilitate and long-term care, while also improving the quality and effectiveness through control systems. Traditional Chinese medicine is also to be enhanced, mainly through health increased capacity, further promotion and strengthening of Traditional Chinese medicine technologies. Priority groups, such as mothers and their respective children, or elderly and disabled citizen, are to get increased attention, for example through strengthening birth defect controlment, expanding elderly care or the development of barrier-free facilities. The basis for this enhanced healthcare system will be optimized health insurance, through universal health insurance coverage, as well as better management and promotion of commercial health insurances. Another structure to improve is the drug supply and security system. More reforms, specifically for pricing and the supply-chain of drugs, are to be implemented, while also changing drug policy to access to essential drugs, especially for children.

Building a healthy environment

Health campaigns, with the goal of improving the rural and urban health environment as well as sanitary conditions, are to be deepened. Healthy cities, towns and villages are planned. Environmental problems, such as air, water and land pollution, are to be diminished through legal actions, while simultaneously implementing plans on industrial discharge controlling. Health and environment are to be monitored, and risk assessment systems are to be implemented. Food safety will be increased through improving food safety standards, while also regulating drug safety.

Developing the healthcare industry

The pluralistic structure of medical care services is to be optimized through investments, for example in medical technology innovation, and political means, such as a change in intellectual property rights. New additions planned are for example internet-based health services or health tourism. International standards of drug and medical equipment are to be met by 2030. Public safety systems, such as road traffic safety systems or emergency management systems, are also to be improved.

Support Mechanisms

There are also additional mechanisms planned to ensure the success of Healthy China 2030. First, Health is set to be the priority in all policies. Additional financing mechanisms for healthcare are planned, as well as additional healthcare reforms and decentralization in the healthcare sector. Health personnel is to be strengthened, through means like better education, and additional task forces such as social sport mentors. Incentives, namely contract-based employment or remuneration of primary healthcare staff are further support mechanisms. Lastly, science and technology are to be promoted, and a national medical innovation system is planned.

2.4 Healthy Chinas effect on diabetes control

While Healthy China 2030 is supposed to be a general direction for Chinese healthcare development, the researcher tested whether it is also suitable to control the diabetes prevalence in mainland China. For that, he reviewed literature in the field of the five main tasks and added diabetes as a keyword for specified actions, e.g. searching for health education diabetes, in order to review the effectiveness of the policy change on diabetes control.

Health education, as well as improving physical activity, are shown to help in controlling diabetes. MakkiAwouda et al. (2014) conducted a study on diabetic patients to determine the effects of health education regarding the improvement of health status and its control. The results of the study show that health education is both suitable and helpful for different levels of age, sex and education. In the research article written by Qi et al. (2008), different prospective studies were compared, and their results consistently indicated that improving physical activity reduces the risk of type 2 diabetes.

Universal access to public health services also helps with the control of diabetes. Zhang et al. (2012) examined the relationship between access to healthcare access and diabetes control through a survey and concluded that people without access to healthcare are more likely to have worse diabetes control profiles than their counterparts.

A review of 84 different controlled clinical studies done by Zhao et al. (2006) focused on the effectiveness and safety of Chinese medicine in the treatment of diabetes type 2, and reported improvement in glycemia, insulin resistance, secondary failure and adverse effects, while simultaneously relieving diabetes symptoms.

Valk et al. (2004) conducted an observational study in two different diabetes cohorts with implemented quality improvement programs. After the implementation, the glycemic control improved at both cohorts.

Lee et al. (2006) reported a strong correlation between insulin resistance and serum concentrations of persistent organic pollutions, after using cross-sectional data from the 1999 – 2002 US National Health and Examination Survey. Therefore, reducing pollution, especially persistent organic pollutants, can help to reduce the prevalence of diabetes.

Medical technological innovations, for example, internet-based health services are expected to be usable in diabetes prevention and help in diabetes research through the generation of data. Fagherazzi & Ravaud (2018) discuss that digitosome data, a term they suggest to be used for data generated online by individuals as well as digital technologies, will profoundly change the way diabetes is controlled and prevented, as the patients will not only be characterized by glucose levels and glycated haemoglobin, but rather by for example real-time psychological well-being. Thus, the blueprint Healthy China 2030 is expected to help with controlling diabetes.

Fagherazzi & Ravaud (2018) also discuss the application of machine learning in diabetes research. Machine learning will be discussed in the next chapter.

3 MACHINE LEARNING

Machine learning in research seeks to provide knowledge to computers through data, observations and interacting with the world, in order to acquire knowledge to generalize new settings (Bengo, 2019).
Thus, it is the computational task of producing general hypotheses or learning correlations from data. In order to learn those, a training set is deployed and validated through various ways with a test set. The found information is then used by the data scientist to create a model, which is able to make predictions of future data. If the data is labelled, the process is called supervised machine learning, if unlabelled, it is called unsupervised machine learning.

In this section, the process of supervised machine learning, as the basis of the later build predictive classifier, will be discussed in detail, while also explaining the three classification algorithms used in the prediction model in section 4, namely Logistic Regression, Support Vector Machines and Random Forests, briefly and understandably for laymen, .

3.1 The Process of Supervised Machine Learning

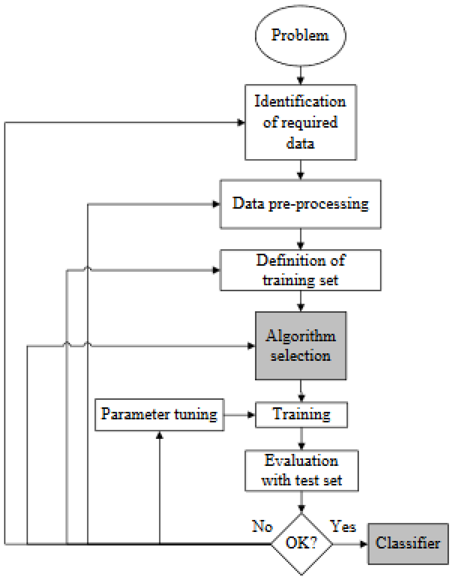

Figure 6 Supervised Machine Learning Model, adapted from "Supervised Machine Learning: A Review of Classification Techniques" by Kotsiantis et al., 2007. Copyright 2007 by the Authors.

If the data in machine learning process is labelled, meaning it has attached metadata which provides information about the initial data, one speaks of supervised machine learning. A general model of supervised machine learning can be seen in figure 6.

Understanding the Problem

To understand the problem, the data scientist must understand what kind of data he/she has at hand, either labelled or unlabelled, to categorize the problem into supervised or unsupervised.
In the process of supervised machine learning, is there are two problem categories: regression problems, which are solved by creating predictions on a continuous scale, and categorization problems, which are solved by predicting categories. Regression results fit the data, while classification results divide it. Categorization problems can furthermore be divided into binary classification, a classification with only two classes, and multi-class classification, classification with more than two classes.

Identification of required data

Once the problem is understood, the data that is required for solving the problem must be identified. Preferably, an expert suggests the attributes and features that are important. If that is not the case, brute-forcing, meaning simply measuring all data available, can be used. However, datasets that are collected by brute-forcing require significant pre-processing, as it often contains missing feature values or distortion in data, called noise (Zhang et al., 2002).

Data pre-processing

When trying to analyse data, data pre-processing is an important step, as for example some classification techniques, such as Logistic Regression, are highly vulnerable to missing data, noise in the data set or outliers. The process generally consists of data cleaning and/or data imputation and feature reduction.
The basic principle of data cleaning is to analyse the reason for so-called dirty data, and to propose cleaning rules in order to improve the quality of data (Yang et al., 2015). An example for dirty data are, depending on the circumstance, NULL values. They signify missing or unknown values.
Three main types of missing data exist: Missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR). MCAR data is not related to other values or the missingness of the hypothetical value, in other words, the missingness is not systematic. MAR data has a systematic relationship between observed data and the propensity of missing values, but not the missing data, e.g. if men are more likely to tell their weight then woman, weight data would be MAR. MNAR data is data with a relationship between the propensity of a missing value and its hypothetical value, e.g. when sick people drop out of a longitudinal study (Graham, 2009).
When faced with them, the data scientist must understand why the data is missing, in order to proceed with either deletion or imputation through various means (Batista & Monard, 2003). In order to increase the operation time and efficiency of supervised machine learning models,

feature selection is used. It is the process of first identifying and then removing irrelevant and redundant variables. (Yu & Lui, 2004).

Definition of the training set

When the data set was cleaned, a training set needs to be deployed. In a supervised classification problem, training data is the split of the initial set of information used to learn the rules of assigning the different instances to the different groups. The idea of using training data in machine learning is while being simple, the very foundation of the learning process. There are two competing concerns when it comes to choosing the correct split: with less training data and more testing data, parameter estimates have greater variance. On the other hand, with more training data and less testing data, there will be greater variance. In the end, the split depends on the total amount of instances.

Another concern specifically in classification problems are imbalanced training sets, which have a different number of data points available for the different classes and are therefore not equally presented, which leads to faulty learning and data understanding (Japkowicz & Stephen, 2002).

Training, evaluation of results with the test set and parameter tuning

Once the training set is defined, and the algorithms are chosen, the actual training of the model is done. How the different algorithms learn is briefly discussed in chapter 3.2, 3.3 and 3.4 respectively.

When evaluating the results of the training, a test set is deployed. It is the other split of the initial data set to access the use of the classification model. When evaluating the accuracy of a machine learning model for classification, there are three different metrics to evaluate. First, there is the accuracy, which is the ratio of correctly predicted observations to the total observations. The second is called precision and is the ratio of correctly predicted positive observations to the total predicted positive observation. The recall is the ratio of correctly predicted positive observations to all observations in the actual class.
If the received metrics are low, parameter tuning can be done, which again depends on the different used algorithms. Once better results are achieved, the model can be deployed.

3.2 Logistic Regression

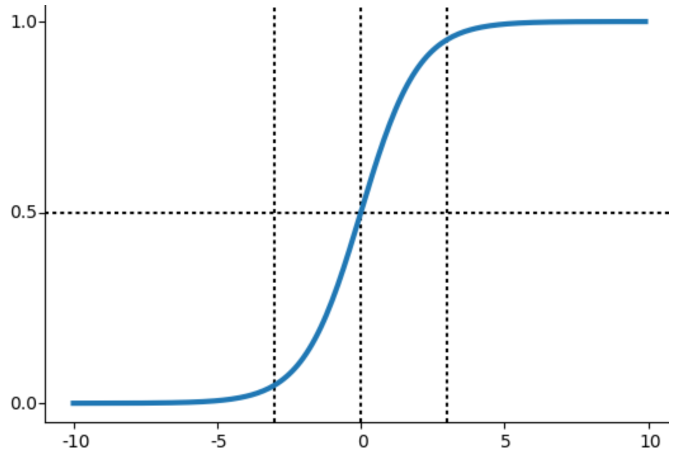

Figure 7 Sigmoid Function

Logistic regression is based on the idea of finding the relationship of a feature, and the probability of a particular outcome. Its origin is the sigmoid function as seen in Figure 7, and its output is a probability that a given input belongs to a certain class. Therefore, the output always lies between 0 and 1. The algorithm learns through maximum likelihood estimation (Pant, 2019).

3.3 Support Vector Machines

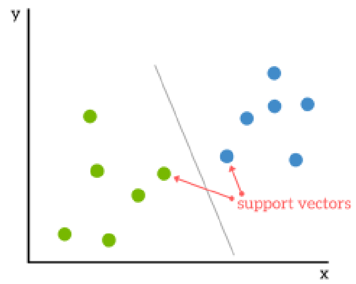

Figure 8 5 A 2-dimensional Support Vector Machine, "Support Vector Machines: A Simple Explanation" by Bambrick N., 2016, KD Nuggets, retrieved April 29, 2019, from https://www.kdnuggets.com/2016/07/support-vector-machines-simple-explanation.html

Support Vector Machines (SVM) are based on the idea of a margin – either side of a hyperplane which separates two classes. When the margin is maximized, the largest possible distance between hyperplane and instance is created, which reduces the expected generalization error. If data is linearly separable and an optimum separating hyperplane has been found, the data points that lie on its margin are called support vectors, and the results are a linear combination of those points (Suykens & Vandewalle) 1999).

In classification, the hyperplane finds an optimum hyperplane, or margin maximizing hyperplane, which best separates the features into different domains, as for example seen in figure 8.

3.4 Random Forest

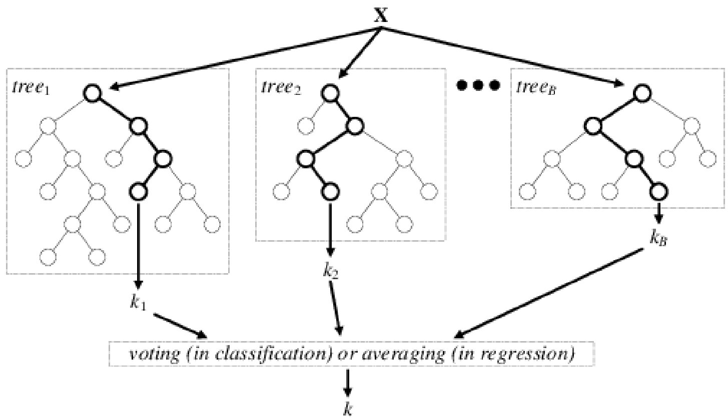

Figure 9 Random Forest Model, adapted from "Electromyographic Patterns during Golf Swing: Activation Sequence Profiling and Prediction of Shot Effectiveness" by Verikas A. et al., 2016. Copyright 2016 by the Authors.

Random forests are ensembles of many classification trees. The algorithm fits many classification trees into a data set and uses them to create predictions from all the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features, hence it being random. In classification, a prediction is done by taking a majority vote for the predicted class (Breiman, 2001).

4 ANALYSIS AND PRESENTATION OF DIABETES DATA

This chapter discusses the data analysis and findings through statistical means, in order analyse the impact of the blueprint Healthy China 2030 specifically on diabetes prevention on the elderly in China. The data will be analysed with SPSS 24, and results are given in valid percent, as some cases are missing. It is mentioned if missing cases reach a higher threshold than 10 percent.

4.1 Frequency Analysis on demographic information

The dataset consists of 20.411 participants in total, after deleting cases born after 1979, as they are not 40 yet. Out of those 48 percent are male, and 52 percent are female. Most of the interviewed people, 70,8 percent, live in a village, and 14,6 percent in a main city zone, the other five options, such as combination zones or special area, amount to a total of 14,6 percent. This result seems quite contradictory to chapter 2.2, specifically urbanization, but can be explained by the nature of the study, as the majority of the survey was conducted in villages.
The birth time ranges from 1910 to 1978, the mean being 1956. The standard deviation with a confidence interval of 95 percent is rounded 11, which means that 95 percent of the participants are born between 1945 and 1967. Only a low of amount of people, 37,1 valid percent, is literate, which can partly be explained by the difference of rural and urban areas, as well as with missing cases of 88 percent.

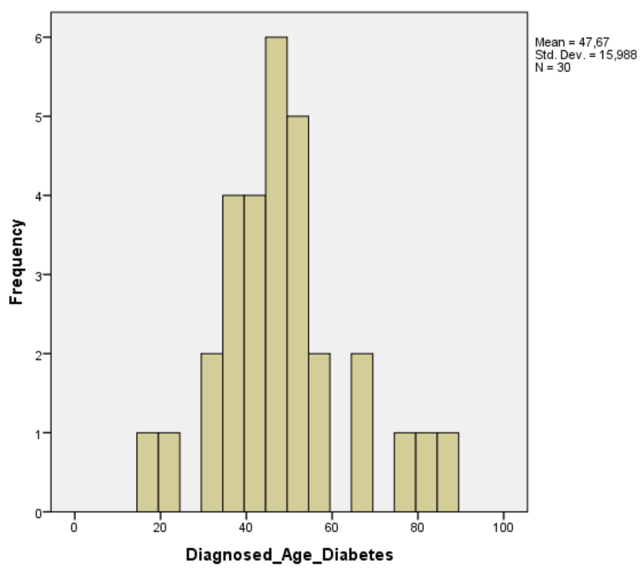

Figure 10 Histogram Diagnosed Age Diabetes

As seen in figure 10, the earliest age someone was diagnosed with diabetes was 17, the oldest 87. 95 percent of the respondents are within 33 and 63 when diagnosed with diabetes, with the mean being rounded 48, and the standard deviation 15. There is only a small amount of people that responded to the question, n = 30, which is why the result is not representative. Nevertheless, it still shows that an increase in age increases the chance of suffering from diabetes.

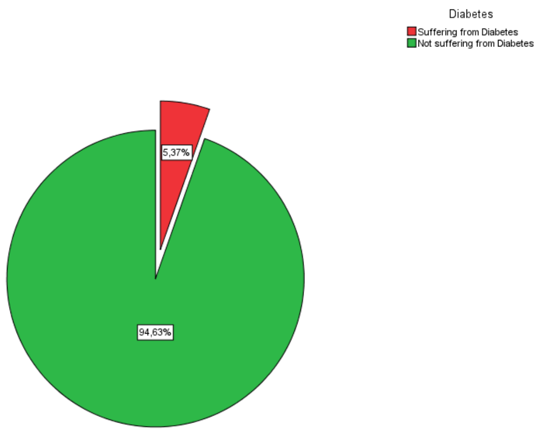

Figure 11 Pie Chart Diabetes Status

1090 people in total suffer from diabetes, which amounts to 5.4 percent of the population. In the next chapter, relationships between diabetes and other variables will be explored. The results discussed can be seen in Appendix A.

4.2 Cross-Tabulation Analysis

To get a better understanding of the data at hand, the researcher conducted cross-tabulations between diabetes and various other variables to establish possible relationships.

Differences between gender

Out of 1090 people suffering from diabetes in total, 56.7 percent are female, and 43.3 percent are male.

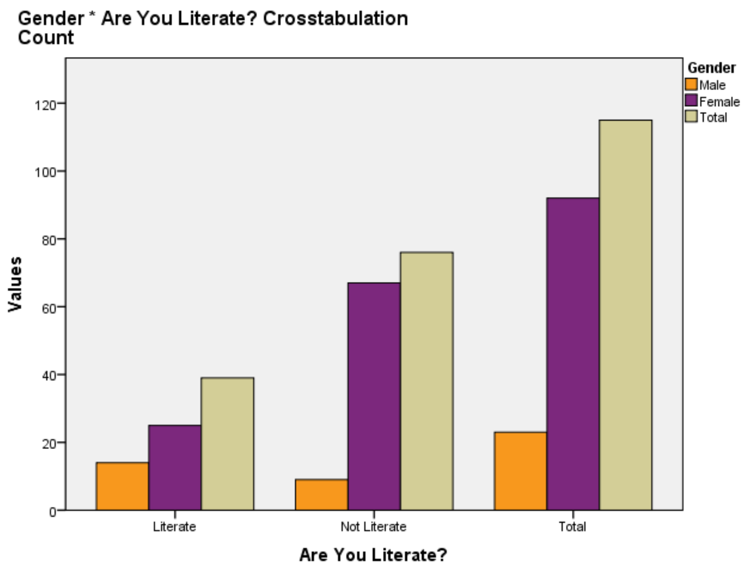

Figure 12 Cross-Tabulation Literacy/Gender

One of the reasons for the increase could be the higher likelihood of women to be obese, as shown and explained in chapter 2.2. Also, as explained in chapter 4.1, the database consists of more male participants than female. Adding to that, more females than men are not literate in this database, as seen in figure 12, with 88 percent and 12 percent respectively, which consequently is a sign for a lower education level among the female population. This is explainable by the patriarchal structure, the male dominance in both society and culture, that China used to have (Liu & Carpenter, 2005).

Additional Disease

The results show that if suffering from diabetes, participants are likely to have another illness as well. Predominantly hypertension, which 56.3 percent of cases suffer from, followed by dyslipidemia with 45,8 percent, heart attacks with 26,7 percent, stomache diseases with 25,6 percent, chronic lung disease with 14, 2 percent, kidney disease with 13,4 percent, liver disease with 8,4 percent, other medical diseases with 6,7 percent, stroke with 6,4 percent, and cancer with 2,4 percent. A visualization of the hypertension/diabetes crosstabulation can be seen in figure 13.

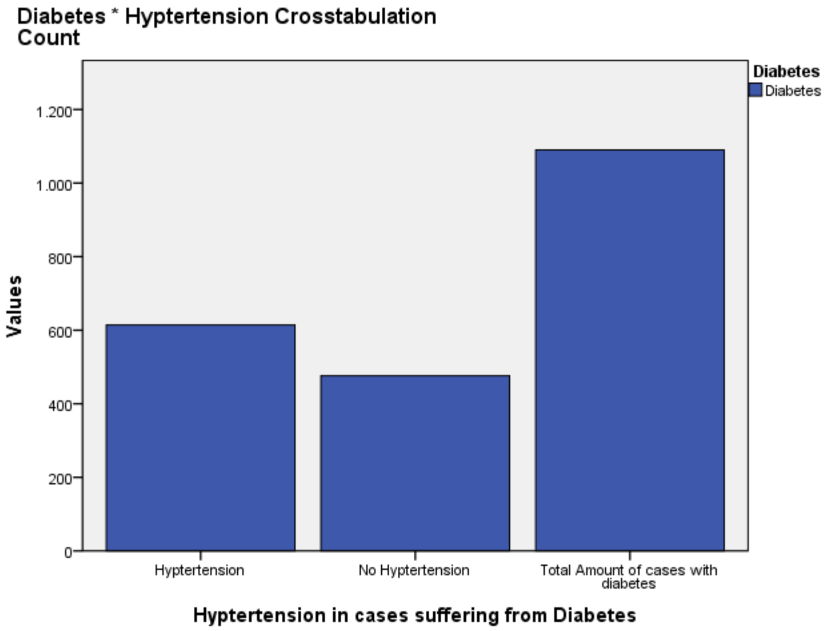

Figure 13 Crosstabulation Suffering from Diabetes and Hypertension

This is not surprising, as diabetes is known to be a catalyst for cardiovascular diseases, such as hypertension, heart attacks or strokes, as mentioned in chapter 2.2. Dyslipidemia is caused by either genetic factors or lifestyle factors, such as diet or obesity level (Huizen, 2018). As the other diseases, such as kidney diseases or stomache diseases, are not specified, the researcher was not able to establish a link between diabetes and the illnesses. Additional data can be found in Appendix C.

Activity Level

The activity level of participants was assessed on a weekly basis, consisting of either high-intensity, medium-intensity or low intensity sports, also separated by time – a minimum of 10 minutes, at least 30 minutes, at least two hours, at least four hours. The highest amount of cases suffering from diabetes was found in the low-intensity category specifically at least 10 minutes, with the amount significantly decreasing when increasing the time as well as the activity level.

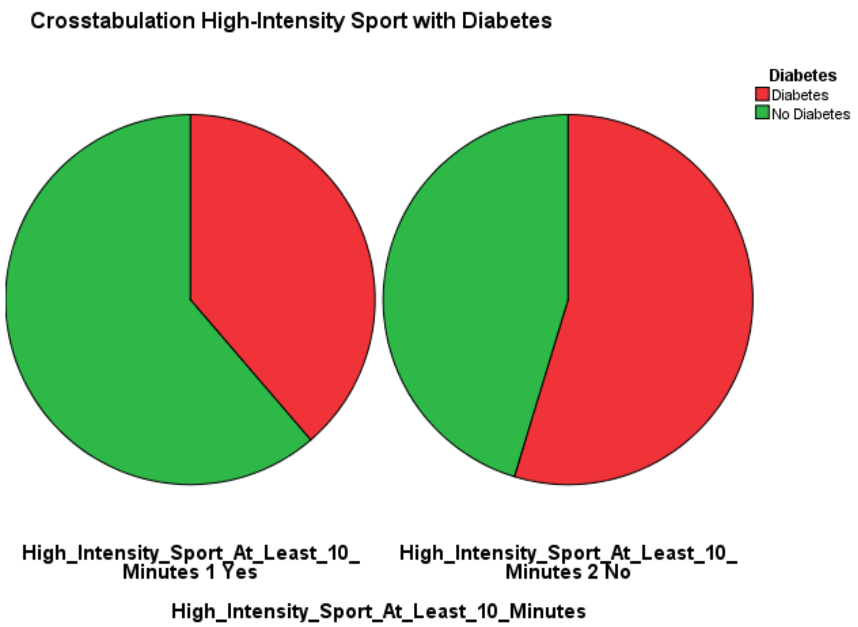

Figure 14 Crosstabulation High-Intensity Sport at least 10 minutes Diabetes/No Diabetes

Also, a clear trend can be observed when comparing people suffering from diabetes and participants not suffering. While having nearly the same ratios of doing low-intensity sports for at least 10 minutes (78,8 percent and 79.0 percent are doing at least 10 minutes of low-intensity activity, for diabetes and non-diabetes respectively), the ratios start to differ in the medium-activity category, 50,4 percent and 56,4 percent. When comparing high-intensity level, an even bigger discrepancy can be observed. Only 22,7 percent of participants suffering from diabetes are doing high-intensity sports for at least 10 minutes, compared to 36,0 percent of the healthy people. The same trend can be observed when comparing data for 30-minutes, two hours or 4 hours. The data can be found in Appendix B.

Smoking and Alcohol Consumption

Interestingly, when comparing drinking habits of Chinese citizen suffering from diabetes and healthy citizen, ill people are less likely to drink more than once a month, with only 19,8 percent compared to 27,3 percent, as

well as less than once a month, 6,5 percent compared to 9,0 percent. A third answer, none of the other two options, was chosen by 73,7 percent for ill persons, and 63,7 persons, which could both indicate drinking more heavily than only once or less a month or being completely abstinent.
Ill people are also less likely to smoke, 8,4 percent compared to 13,8 percent for healthy people. A possible explanation for that could be advice given by doctors regarding smoking and drinking, as it can make the sickness worse, and lead to complications (Vieira, n.d). The data can be found in Appendix D.

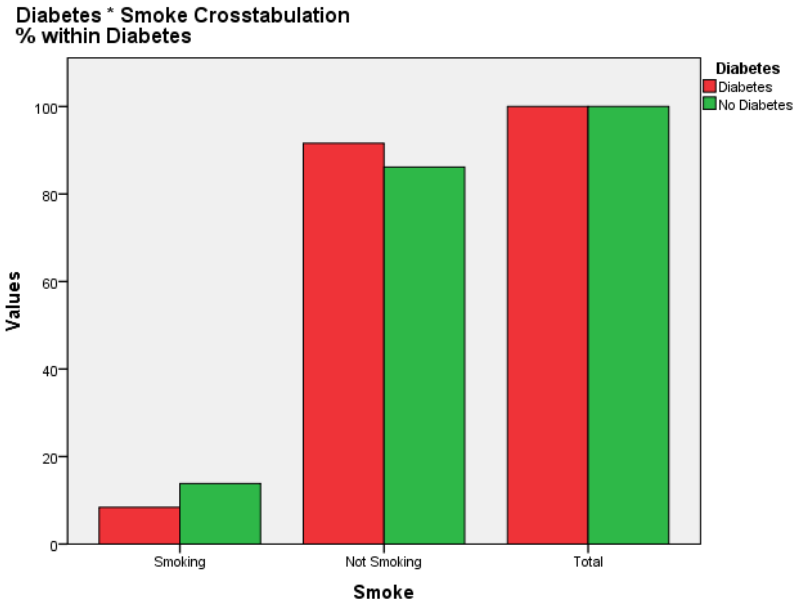

Figure 15 Smoking Cross-Tabulation Diabetes/No Diabetes

4.3 Chi-Square Test of Independence

To calculate whether relationships between the different categorical values do have a statistical significance, the Chi-Square test is used. Dependent variable is always Diabetes, independent ones are: Hypertension, Dyslipidemia, Heart Attack, Stroke, Literacy, Village or Town, Gender, Drinking Habits, Smoking Habits, High-Intensity Sports for at least 10 minutes, Medium-Intensity Sports for at least 10 minutes, Low-Intensity Sports for at least 10 minutes. Two hypotheses are tested, with a significance level of α = 0.05:

H_0 = Diabetes is not associated with variable X

H_1 = Diabetes is associated with variable X

H_0 is rejected if the result in form of the p-value is equal or smaller than the significance level α = 0.05. If the p-value is larger then the significance level α = 0.05, there is not enough evidence to reject H_0. The author will demonstrate the process of Chi-Square test on diabetes paired with hypertension in detail, while only explaining results for the other pairings.

Process of Chi-Square testing in detail

Crosstab

			Hyptertension 1	Hyptertension 2	Total
Diabetes	1	Count	614	476	1090
		Expected Count	225,1	864,9	1090,0
	2	Count	3580	15640	19220
		Expected Count	3968,9	15251,1	19220,0
Total		Count	4194	16116	20310
		Expected Count	4194,0	16116,0	20310,0

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	894,902[a]	1	,000		
Continuity Correction[b]	892,602	1	,000		
Likelihood Ratio	713,029	1	,000		
Fisher's Exact Test				,000	,000
Linear-by-Linear Association	894,858	1	,000		
N of Valid Cases	20310				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 225,08.

b. Computed only for a 2x2 table

Figure 16 Example of Crosstab: Hypertension with expected count and Chi-Square Tests

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	,210	,000
	Cramer's V	,210	,000
N of Valid Cases		20310	

Figure 17 Symmetric Measures of Chi-Square Test as seen in Figure 16

If the different variables, in the case of figure 16 Hypertension and Diabetes, were independent or not associated, the scientist would expect a count of 225 people with Hypertension and Diabetes, 865 with Diabetes and no Hypertension, 3969 without Diabetes but with Hypertension, and 15251 without both Hypertension and Diabetes, as seen in the expected count in figure 16. The Chi-Square count shows whether these differences between count and expected count are enough for the test to be significant. The two assumptions, both found in figure 16 under the table, called a and b, which are needed for the test, are met.

Therefore, the result important for the research, called p-value, found under Asymptotic Significance (2-sided), found in figure 16, can be explained. As the observed p-value 0,000, is smaller than the significance level α = 0.05, the alternative Hypothesis H_1 can be accepted. Thus, the result is statistically significant. In other words, Hypertension is dependent on Diabetes.

Next up is the Phi coefficient, found in figure 17. It tells the effect size, or correlation coefficient, and ranges from -1 to 1 (Davenport & El-Sanhury, 1991).
The researcher is able to interpret the value as a weak positive relationship meaning that if the class of Diabetes increases, it is likely that the group of Hypertension also increases.

Other results

When comparing observed counts and expected counts of cases suffering from diabetes, the observed ones were always greater or smaller than expected ones, especially for the different illnesses. Thus, the scientist was able to reject H_0 for all twelve variables. The effect ranges differed but were especially high for illnesses. Also, increasing the activity level resulted in a bigger negative relationship. Overall, the results already mentioned in chapter 2.2 as well as 4.2, were proven correct. Additional data is found in Appendix B, C and D.

5 SUPERVISED MACHINE LEARNING SOLUTION

As diabetes was now discussed in detail, a predictive machine learning classifier will now be introduced. The purpose is the demonstration of the capabilities of machine learning for diabetes detection. Three different models are written in Python and will be trained with the help of the CHARLS 2015 dataset, in order to make decisions of the diabetes class of participants in the survey.

5.1 Python and SciKit-Learn

Python is an object-oriented, high-level programming language, with integrated dynamic semantics. It supports the use of modules and packages called libraries, such as SciKit-learn. The later is a library which provides algorithms specifically for machine-learning, which was the reason why the researcher chose this specific language. Prerequisites for the use of SciKit-learn are NumPy, SciPy, and Pandas. Those libraries are used for the manipulation of data, specifically its structure and dimensionality.

5.2 Methodology of the Machine Learning model

The foundation of a well-functioning machine learning model is the data that it uses. Therefore, the researcher used parts of the China Health and Retirement Longitudinal Study, in short CHARLS. It is a longitudinal survey of persons in China that are 45 years of age or older. Using CHARLS ensures that the research fits the needs of the aging Chinese society.
The survey is of national representation and includes information about the health, social and economic conditions of the participants.
The parts used for building the model in Python were the health status and functioning section, as well as parts of the demographic section, and are from CHARLS 2015. The reason for choosing those specific sections is that they contain information about diabetes of the participants and living conditions of the participants, while also being of national representation. The methodology applied in the thesis is thus of quantitative nature, specifically of secondary quantitative nature, as the researcher uses CHARLS.
The cleaning of the dataset is described in chapter 4.3. As the datasets dependent variable is heavily unbalanced, with less than 5 percent of the participants suffering from diabetes, the synthetic minority over-sampling technique is used on the training set. Diabetes is used as the binary predictor, and the evaluation of the different models is done through a

comparison of accuracy, precision and recall scores, as well as the F1-score.

5.3 Pre-emptive feature selection of the dataset

The dataset consists of 1135 variables without pre-processing. Therefore, pre-emptive feature selection must be done, as the computing time in Python is otherwise very high. In order to do so, NMAR variables were deleted, and the others selected on the basis of Healthy China 2030 goals. The total amount of variables after feature selection was 21. Categories used were illnesses, information on the physical activity of respondents, and demographic information such as gender or education level. The different variables were discussed in chapter 4. After that, Null values were deleted, as reasonable imputation wouldn't have been possible with the system at hand. This left the researcher with a total of 5944 cases, split into 5654 cases without diabetes (group 2), and 290 with diabetes (group 1), as seen in figure 18

```
2.0    5654
1.0     290
Name: Diabetes, dtype: int64
```

Figure 18 Dataset after cleaning, split into Diabetes types

5.4 Diabetes prediction model

The first model introduced by the author is built with the Random Forest algorithm, which is discussed in chapter 3.3. The scientist will discuss the process of building the model in detail. Following algorithms will only discuss results.

```
In [1]: import pandas as pd
        import numpy as np
        from sklearn.ensemble import RandomForestClassifier
        from imblearn.over_sampling import SMOTE
        from sklearn.metrics import  accuracy_score, recall_score,precision_score

        from sklearn.model_selection import train_test_split

        df_for_modeling = pd.read_excel("C:/Users/Nico/Desktop/Health2.xlsx")
```

Figure 19 Importing libraries into Python

The first step was to import the libraries needed for building the classifier, as discussed in chapter 5.1. The process can be seen in figure 19. The name of the dataset is df_for_modeling, in the following called dataset.

```
In [2]: df_for_modeling.dropna(subset=["Birthyear", "Other_Medical_Diseases", "High_Intensity_Sport_At_Least_10_Minutes",
                                       "Medium_Intensity_Sport_At_Least_10_Minutes", "Low_Intensity_Sport_At_Least_10_Minutes",
                                       "Smoke", "Drink", "Gender", "Adresstype", "Village_or_Town", "Martial_Status", "Hyptertension"
                                       , "Diabetes", "Dyslipedimia", "Cancer", "Chronic_Lung_Disease", "Liver_disease",
                                       "Heart_Attack", "Stroke","Kidney_disease","Stomage_disease"], inplace=True)
```

Figure 20 Removing NULL values

Next up was the removing of the NULL values from the dataset, as all Random Forest cannot work with them, as seen in figure 20.

```
In [3]: X = df_for_modeling.loc[:, df_for_modeling.columns != 'Diabetes']
        y = df_for_modeling.loc[:, df_for_modeling.columns == 'Diabetes']
```

Figure 21 Defining dependent and independent values

Once they were removed, the dataset was split into two parts, dependent variable, set as Diabetes as this was the variable the scientist wanted to predict, and independent variables, which were all other variables. The process can be seen in figure 21.

```
In [4]: from sklearn.model_selection import train_test_split

        X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
```

Figure 22 Splitting the dataset into training data and testing data

As already explained in chapter 3.1, the dataset must be split into testing data and training data. This was done for both, dependent and independent variables, which amounted to a total of four different sets. The training-test split was 80/20, meaning 80 percent of the data in the training set, and 20 percent of the data in the testing set, as seen in figure 22 under test_size=0.2.

```
In [5]: random_forest = RandomForestClassifier(n_estimators=100)
        random_forest.fit(X_train, y_train)
        y_pred = random_forest.predict(X_test)
```

Figure 23 Training the Random Forest algorithm

The algorithm was then trained with the help of training data, as seen in figure 23. This was followed by checking of the three different metrics used for evaluating a machine learning model, as seen in figure 24.

```
In [6]: precision_score(y_test, y_pred)

Out[6]: 0.3333333333333333

In [7]: recall_score(y_test, y_pred)

Out[7]: 0.14285714285714285

In [8]: accuracy_score(y_test, y_pred)

Out[8]: 0.9596299411269975
```

Figure 24 Random Forest Metrics without SMOTE

The first is called precision and is the ratio of correctly predicted positive observations to the total predicted positive observation. The score is 33 percent, which in other words means that it is the amount of predicted diabetes instances that actually have diabetes is 33 percent. Therefore, 67 percent of healthy people were falsely classified.
Following is recall, the ratio of correctly predicted positive observations to all observations in the actual class. Thus, recall shows that out of all the instances that truly have diabetes, rounded 14,3 percent were predicted by the algorithm.
Last, there is the accuracy, which is the ratio of correctly predicted observations to the total observations. As the accuracy is quite high with nearly 96 percent, as it was able to nearly correctly classify all participants without diabetes.
The reason for the low precision and recall values, as well as the high accuracy, is overfitting. It is briefly explained in chapter 3.1. A possible solution for that is the synthetic minority over-sampling technique (SMOTE), which under-samples the majority class, and over-samples the minority class, in order to generate a balanced training set (Chawla et al., 2002).

```
In [10]: smt = SMOTE()
         X_train, y_train = smt.fit_sample(X_train, y_train)
```

Figure 25 Introduction of SMOTE into the model

Once SMOTE was introduced to the model, as seen in figure 25, the classifier had to be trained again, which was the same code seen in figure 23. Interestingly, SMOTE didn´t increase the precision, accuracy or recall, but rather decreased all three, as seen in figure 26.

```
In [12]: precision_score(y_test, y_pred)

Out[12]: 0.3181818181818182

In [13]: accuracy_score(y_test, y_pred)

Out[13]: 0.946164199192463

In [14]: recall_score(y_test, y_pred)

Out[14]: 0.09722222222222222
```

Figure 26 Random Forest Metrics after SMOTE

Logistic Regression Model

As already mentioned previously, only results will be discussed now, as the only things that changed were algorithm and results.

```
In [6]: precision_score(y_test, y_pred)
Out[6]: 0.4

In [7]: recall_score(y_test, y_pred)
Out[7]: 0.03571428571428571

In [8]: accuracy_score(y_test, y_pred)

Out[8]: 0.9520605550883096
```

Figure 27 Logistic Regression Metrics without SMOTE

In figure 27, the results of Logistic Regression without SMOTE can be seen. It achieves the highest precision yet with 40 percent but has a very low recall score with only 3,5 percent. The accuracy is also quite high with 95,2 percent. The results change quite drastically when introducing SMOTE to the model. The precision reduces to 12,9 percent, the accuracy by nearly 17 percent to 78,5 percent, but recall increases by 56 percent to a total of 59,7.

```
In [12]: precision_score(y_test, y_pred)
Out[12]: 0.12912912912912913

In [13]: accuracy_score(y_test, y_pred)
Out[13]: 0.7853297442799462

In [14]: recall_score(y_test, y_pred)
Out[14]: 0.5972222222222222
```

Figure 28 Logistic Regression Metrics with SMOTE

Support Vector Machine

Interestingly, without SMOTE, Support Vector Machines did not produce any results for diabetes but was still able to predict all non-diabetes cases, reflected in its high accuracy score seen in figure 29.

```
In [8]: accuracy_score(y_test, y_pred)

Out[8]: 0.9545836837678722
```

Figure 29 Support Vector Machine Accuracy without SMOTE

Once SMOTE was introduced, the Support Vector Machine was able to increase its precision score to 14,5 percent, its recall score to 58,3 percent and decreased its accuracy to 81,3 percent, as seen in figure 30.

```
In [12]: precision_score(y_test, y_pred)
Out[12]: 0.14482758620689656

In [13]: accuracy_score(y_test, y_pred)
Out[13]: 0.8129205921938089

In [14]: recall_score(y_test, y_pred)
Out[14]: 0.5833333333333334
```

Figure 30 Support Vector Machines Metrics with SMOTE

Concluding the machine learning algorithm

While the accuracy is consistently the highest score, it is the least important one in this specific thesis, as it only works well if there are an equal number of samples in both classes, which is not the case as seen in figure 18. Therefore, precision and recall are more important, as they give us information regarding the precision of diabetes prediction, and whether it misses many cases. Naturally, there is always a trade-off between recall and precision, which can be concluded from the different results in the different results.

Thus, the F1-score will be calculated, which is the harmonic mean between the two metrics, and used to select a model. Its formula can be seen in figure 31.

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

Figure 31 Calculation of the F1 Score

The Support Vector Machine with SMOTE had the highest F1 score with 22,5 percent, followed by Logistic Regression with SMOTE with 20 percent. The Random Forest model without SMOTE was able to score 18 percent and with SMOTE 14 percent. Logistic Regression without SMOTE had 5 percent, Support Vector Machine 0 percent.

Therefore, the Support Vector Machine with SMOTE is the best model for diabetes detection, with an accuracy of 81,3 percent.

6 RECOMMENDATIONS AND LIMITATIONS

When comparing the results of chapter 2.1, 2.2, section 4 and section 5 to the contents of Healthy China 2030, it is apparent that the Chinese government understood the need for the strategic drift. The five stated goals, improving the general health status, supporting a healthy life, optimizing healthcare services and security, building a healthy environment and further developing the healthcare industry, are all interconnected, and will help to reduce the prevalence of diabetes, as they will diminish the risk factors of diabetes. Specified actions, such as reducing the environmental pollution, are proven to help, as seen in chapter 2.4.

Strengthening the health education as well as the encouragement of healthy habits will decrease both obesity and dietary causes. Also, an improvement of the physical fitness of individuals will help reducing the prevalence of diabetes.

Once universal access to public health services is granted and high-quality and efficient care are in place, the negative effects of urbanization will diminish. Also, healthcare services for priority groups such as the elderly, who are as shown in more need of healthcare, will help to reduce the diabetes prevalence as well. Also, once the pluralistic structure of medical care services is optimized through investments, and science and technology were promoted, the researcher recommends both investment as well as additional research in the field of Machine Learning.

As shown in this thesis, predictive machine learning classifier can be built, and help both the detection of diabetes, and even prevention if detected early enough. It must be mentioned that this does not happen with the same accuracy as traditional classifiers that are based on glucose level. China does not only have a government that understands the need for change, but also the resources to implement these measures. With those policy changes coming, China is bound to reduce the diabetes rate, and to flourish again.

6.1 Limitations of the thesis

In this thesis, some limitations were taking place. First, the scientist does not speak Chinese and was thus not able to access all literature regarding Healthy China 2030. Additionally, due to a lack of monetary resources, a more detailed literature review was not possible.

More importantly, the database, while being of high-quality, does not differentiate between Diabetes Type 1 or Diabetes Type 2. Adding to that, many of its variables contain NULL values, which is generally a big problem in research. The author used the assumption of not mentioning being sick means being healthy, which is the reason why there were no NULL values for diabetes, hypertension, etc.

As the author was not able to build the models on a strong computer system, no imputation for missing values was possible, which decreased both precision and recall. Therefore, the model cannot be deployed in a real-case scenario, but only used for demonstration purposes of machine learning usage in healthcare. Additionally, as a result of the scope of this thesis, only problems in the Chinese healthcare sector are discussed, while in a larger research, positive aspects could have been provided. Hence, no evaluation of the Chinese healthcare sector was possible.
Due to the methodological choices for the thesis as well as through the systemic restriction, found data did not provide sufficient details to answer the research question comprehensively.

7 CONCLUSION

Problematic demographic and epidemiological trends, such as the aging society, or the discrepancy of quality of care in urban and rural regions are problems from which the Chinese healthcare sector is still suffering from. Another major challenge is the rapid prevalence of diabetes in China, especially among the elderly. Reasons for that are urbanization and associated lifestyles, such as the diet, which consequently lead to obesity and a low level of activity. Other reasons are the absence of education among older residents of predominantly rural areas or low activity level, what also leads to additional diseases, such as hypertension or dyslipidemia, particularly found in people suffering from diabetes.
The Chinese government understood the need for a strategic shift, and announced Healthy China 2030, a blueprint for the national healthcare policy as a medium- and long-term strategic plan for healthcare. It aims to make public health a priority for the economic and social development of the People´s Republic of China. That is made possible by government-led reforms and innovations, scientific development, a change in the healthcare service mode and lastly through the promotion of equity and justice. Three different machine algorithms, Random Forests, Logistic Regression and Support Vector Machines, were used to create a machine learning classifier, as an example of the possible scientific development, that could be used for both prevention and detection, and ultimately control of diabetes.
In conclusion, it can be said that the Chinese government is steering the right course and tries its best do reduce or completely remove underlying causes of the healthcare problems. The Central Council deserves praise for the blueprint of Healthy China 2030.

8 ACKNOWLEDGMENT

The researcher would like to thank the National School of Development at Peking University in Bejing, PR China for giving us access to the CHARLS database. http://charls.pku.edu.cn/en

REFERENCE

Alghamdi M., Al-Mallah M., Keteyian S., Brawner C., Ehrman J., Sakr S. (2017). *Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: the henry ford exercise testing (FIT) project. PLoS One.* 12:e0179805. doi:10.1371/journal.pone.0179805

Alhazmi A., Stojanovski E., McEvoy M., Garg ML. (2014). *The association between dietary patterns and type 2 diabetes: a systematic review and meta-analysis of cohort studies*. Journal of Human Nutrition and Dietics. 27: 251–60.

American Diabetes Association (2012). *Diagnosis and classification of diabetes mellitus. Diabetes Care,* 35(1): 64–71. doi:10.2337/dc12-s064

Babiarz, K. S., Eggleston, K., Miller, G., & Zhang, Q. (2015). *An exploration of China's mortality decline under Mao: A provincial analysis, 1950-80*. Population studies. 69(1): 39–56. doi:10.1080/00324728.2014.972432

Bao, Y., Deng, S., Lin, W. (2015). *Research of Data Cleaning Methods Based on Dependency Rules*. World Academy of Science, Engineering and Technology, International Science Index 106, International Journal of Computer, Electrical, Automation, Control and Information Engineering. 9(10): 2189 - 2193.

Batis, C., Sotres-Alvarez, D., Gordon-Larsen, P., Mendez, M. A., Adair, L., & Popkin, B. (2014). *Longitudinal analysis of dietary patterns in Chinese adults from 1991 to 2009*. British Journal of Nutrition. 111(8): 1441-1451.

Batista, G., & Monard, M.C. (2003). An Analysis of Four Missing Data Treatment Methods for Supervised Learning. Applied Artificial Intelligence. 17: 519-533.

Bengo, Y. (2019). *The Rise of Neural Networks and Deep Learning in Our Everyday Lives – A Conversation with Yoshua Bengio*. Retrieved April 29, 2019, from https://emerj.com/ai-podcast-interviews/the-rise-of-neural-networks-and-deep-learning-in-our-everyday-lives-a-conversation-with-yoshua-bengio/

Blumenthal, D., Hsiao, W. (2005). *Privatization and its discontents - the evolving Chinese health care system*. New England Journal of Medicine. 353(11): 1165-1170.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1): 5-32. doi:10.1023/A:1010933404324

Central Committee of the Communist Party of China. (2016). *The 13th Five-Year plan for economic and social development of the People´s Republic of China 2016 – 2020*. (Compilation and Translation Bureau of the Central Committee of the Communist Party of China, trans.). Bejing, China.

Chan, JC., Malik, V., Jia, W., et al. (2009) *Diabetes in Asia: epidemiology, risk factors, and pathophysiology*. JAMA. 301: 2129-2140.

Chawla, N. V., Bowyer K. W., Hall L. O., Kegelmeyer W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. JAIR. *16.* doi:10.1613/jair.953

China National Bureau of Statistics. (2014). *China Statistical Yearbook*. Bejing, China

Cox M. E., Edelman D. (2009). *Tests for screening and diagnosis of type 2 diabetes. Clin. Diabetes,* 27: 132–138. doi:10.2337/diaclin.27.4.132

Davenport, E., & El-Sanhury, N. (1991). *Phi/Phimax: Review and Synthesis*. Educational and Psychological Measurement. 51: 821–828. doi:*10.1146/annurev.psych.58.110405.085530*

Duygu ç., Esin D. (2011). *An automatic diabetes diagnosis system based on LDA-wavelet support vector machine classifier. Expert Syst. Appl.* 38: 8311–8315.

Fagherazzi G., Ravaud, P. (2018). *Digital diabetes: Perspectives for diabetes prevention, management and research*. Diabetes & Metabolism. doi:10.1016/j.diabet.2018.08.012.

Georga E. I., Protopappas V. C., Ardigo D., Marina M., Zavaroni I., Polyzos D., et al. (2013). *Multivariate prediction of subcutaneous glucose concentration in type 1 diabetes patients based on support vector regression. IEEE J. Biomed. Health Inform.* 17: 71–81. doi:10.1109/TITB.2012.2219876

Gong P, Liang S, Carlton EJ, et al. (2012). *Urbanisation and health in China*. Lancet. 379: 843–52.

Goodpaster BH, Park SW, Harris TB, et al. (2006). *The loss of skeletal muscle strength, mass, and quality in older adults: the health, aging and body composition study*. J Gerontol A Biol Sci Med Sci. 61: 1059-1064.

Graham, J. W. (2009). Missing Data Analysis: Making it Work in the Real World. Annual Review of Psychology. Vol. 60:549-576.

Gu D, Reynolds K, Duan X, et al. (2012). *Prevalence of diabetes and impaired fasting glucose in the Chinese adult population: International Collaborative Study of Cardiovascular Disease in Asia (InterASIA).* Diabetologia 2003. 46: 1190-1198.

He J, Gu D, Wu X, et al. (2005). *Major causes of death among men and women in China.* New England Journal of Medicine. 353: 1124-1134. doi:10.1056/NEJMsa050467

Ho P. Y., Lisowoski F. P. (1997*). A Brief History of Chinese medicine (2nd edition)*. Singapur, Singapur: World Scientific Pub Co Inc. doi:10.1017/S0007114513003917.

Hu C., Jia W. (2018). *Diabetes in China: Epidemiology and Genetic Risk Factors and Their Clinical Utility in Personalized Medication*. Diabetes. 67(1): 3-11. doi:10.2337/dbi17-0013

Hu F.B., Liu Y., Willett W.C. (2011). *Preventing chronic diseases by promoting healthy diet and lifestyle: public policy implications for China*. Obesity Reviews. 12: 552–59.

Huizen, J. (2018). *Dyslipidemia: Everything you need to know*. Retrieved May, 14, 2019 from https://www.medicalnewstoday.com/articles/321844.php

Iancu I., Mota M., Iancu E. (2008). *Method for the analysing of blood glucose dynamics in diabetes mellitus patients*. *Proceedings of the 2008 IEEE International Conference on Automation, Quality and Testing, Robotics*, Cluj-Napoca. doi: 10.1109/AQTR.2008.4588883

International Monetary Fund. (2017). *Report for Selected Countries and Subjects*. Retrieved March 13, 2019, from https://bit.ly/2B2YJD5

Japkowicz N., Stephen, S. (2002). The Class Imbalance Problem: A Systematic Study. Intelligent Data Analysis. 6: 429-449.

Kant, Ashima K. (2004). Dietary patterns and health outcomes. Journal of the American Dietetic Association. 104(4): 615–635. doi:10.1016/j.jada.2004.01.010.

Kavakiotis I., Tsave O., Salifoglou A., Maglaveras N., Vlahavas I., Chouvarda I. (2017). *Machine learning and data mining methods in diabetes research. Comput. Struct. Biotechnol. J.* 15: 104–116. doi:10.1016/j.csbj.2016.12.005

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering,* 160: 3-24.

Lee B. J., Kim J. Y. (2016). *Identification of type 2 diabetes risk factors using phenotypes consisting of anthropometry and triglycerides based on machine learning. IEEE J. Biomed. Health Inform.* 20: 39–46. doi:10.1109/JBHI.2015.2396520

Lee, DH., Lee, IK., Song, K., et al. (2006). *A strong dose-response relation between serum concentrations of persistent organic pollutants and diabetes: results from the national health and examination survey 1999–2002*. Diabetes Care. 29: 1638-1644

Ling, L., Hongqiao, F. (2017). *China´s Health care system reform: Progress and prospects*. The International Journal of Health Planning and Management. 32: 240 – 253.

Lui, J., Carpenter, M. (2005). *Trends and Issues of Women´s Education in China*. The Clearing House. 78(6): 277-281.

Luk, AO., Lau ES., So WY., et al. (2014*). Prospective study on the incidences of cardiovascular-renal complications in chinese patients with young-onset type 1 and type 2 diabetes*. Diabetes Care. 37: 149-157.

Ma, R.C., Lin, X., Jia, W. (2014). *Causes of type 2 diabetes in China.* The Lancet Diabetes & Endocrinology, 2(12): 980-991. doi:10.1016/S2213-8587(14)70145-7.

Maddison, A. (2007)*. Chinese Economic Performance in the Long Run (2nd edition)*. Paris: OECD.

MakkiAwouda, F. O., Elmukashfi, T. A., & Hag Al-Tom, S. A. (2014*). Effects of health education of diabetic patient's knowledge at Diabetic Health Centers, Khartoum State, Sudan: 2007-2010*. *Global journal of health science*. *6*(2): 221–226. doi:10.5539/gjhs.v6n2p221

Mihovska A., Kyriazakos S. A., Prasad R. (2014). *eWall for active long living: Assistive ICT services for chronically ill and elderly citizens*. *IEEE International Conference on Systems, Man, and Cybernetics (SMC),* San Diego, CA. 2014. 2204-2209. doi:10.1109/SMC.2014.6974251

Nather A., Bee C.S., Huak C.Y., Chew J.L.L., Lin C.B., Neo S., Sim E.Y. (2008). *Epidemiology of diabetic foot problems and predictive factors for limb loss*. J. Diab. Complic., 22(2): 77-82.

Pant, A. (2019). Introduction to Logistic Regression. Retrieved April 29, 2019, from https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148

Papatheodorou K., Banach M., Edmonds M., Papanas N., Papazoglou, D. (2015). *Complications of diabetes.* J. Diabetes Res. 1-5.

Peng X. (2011). *China's demographic history and future challenges*. Science. 333: 581–87.

Qi, L., Hu, FB., Hu, G. (2008). *Genes, Environment, and Interactions in Prevention of Type 2 Diabetes: A Focus on Physical Activity and Lifestyle Changes*. Current Molecular Medicine. 8(6): 519-532.

Qian, J. (2015). Reallocating authority in the Chinese health system: an institutional perspective. *Journal of Asian Public Policy*, 8(1): 19-35.

Qu, X., Yin, C., Sun, X., Huang, S., Li, C., Dong, P., et al. (2018). *Consumption of antibiotics in Chinese public general tertiary hospitals (2011-2014): Trends, pattern changes and regional differences*. PloS one. 13(5). doi:10.1371/journal.pone.0196668

Razavian N., Blecker S., Schmidt A. M., Smith-McLallen A., Nigam S., Sontag D. (2015). *Population-level prediction of type 2 diabetes from claims data and analysis of risk factors. Big Data.* 3: 277–287. doi:10.1089/big.2015.0020

Shin, Hyun. (2015). *Urbanization in China*. London, UK: School of Economics and Political Science Publishing. doi:o10.1016/B978-0-08-097086-8.72095-2.

Soumya D., Srilatha B. (2011). *Late stage complications of diabetes and insulin resistance.* J. Diabetes Metab. 2(167): 2-7.

Statista – The Statistics Portal. (2019). *Children and old-age dependency ratio in China from 1990 to 2100*. Retrieved March 13, 2019, from https://www.statista.com/statistics/251535/child-and-old-age-dependency-ratio-in-china/

Sun, Z., Wang, S., & Barnes, S. R. (2015). Understanding congestion in China's medical market: an incentive structure perspective. Health Policy and Planning. 31(3): 390-403. doi:10.1093/heapol/czv062

Suykens, J. A., & Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural processing letters*, *9*(3), 293-300. doi:10.1023/A:1018628609742

Tan, X., Wu, Q., & Shao, H. (2018). *Global commitments and China's endeavors to promote health and achieve sustainable development goals*. Journal of health, population, and nutrition. *37*(1): 8. doi: 10.1186/s41043-018-0139-z

Valk, G. D et al. (2004). *Quality of care for patients with type 2 diabetes mellitus in the Netherlands and the United States: a comparison of two quality improvement programs*. *Health services research*, *39*(4): 709–725. doi:10.1111/j.1475-6773.2004.00254.x

Vieira, K. (n.d.). *Drug and Alcohol use with Diabetes*. Retrieved May, 14, 2019, from https://drugabuse.com/guides/substance-abuse-and-diabetes/

Villegas R, Liu S, Gao YT, et al. (2007). Prospective study of dietary carbohydrates, glycemic index, glycemic load, and incidence of type 2 diabetes mellitus in middle-aged Chinese women. Arch Intern Med. 167: 2310–2316.

Wang et al. (2017). Prevalence and Ethnic Pattern of Diabetes and Prediabetes in China in 2013. JAMA. doi:317. 2515. 10.1001/jama.2017.7596.

Wang, S., Marquez, P., Langenbrunner, J., Niessen, L., Suhrcke, M., & Song, F. (2011). *Toward a healthy and harmonious life in China: stemming the rising tide of non-communicable diseases*. Washington, DC: World Bank, 1-48.

Whiting, D., Unwin, N., Roglic, G. (2010). *Diabetes: equity and social determinants. Blas E Kurup A Equity, social determinants and public health programmes*. World Health Organization, Geneva, Switzerland. 77-94.

Witt, M. (2018). *China: a concise profile 2018*. INS809. Boston, MA: Harvard Business School Publishing.

World Bank. (2016). *Deepening health reform in China : building high-quality and value-based service delivery - policy summary (English)*. Washington, D.C.: World Bank Group.

World Health Organization. (2014). *Noncommunicable Diseases (NCD) Country Profiles China*. Retrieved March 13, 2019, from https://www.who.int/nmh/countries/chn_en.pdf

Wu Y. (2006). *Overweight and obesity in China*. BMJ (Clinical research ed.). 333(7564): 362–363. doi:10.1136/bmj.333.7564.362

Xu, Y., Wang, L., He, J., Bi, Y., Li, M., Wang, T., et al. (2013) *Prevalence and control of diabetes in Chinese adults*. JAMA. 310:948–959.

Yang, G., Kong, L., Zhao, W., Wan, X., Zhai, Y., Chen, L. C., & Koplan, J. P. (2008). Emergence of chronic non-communicable diseases in China. *The Lancet*, 372(9650): 1697-1705. doi:10.1016/S0140-6736(08)61366-5

Yu, L., Liu, H. (2004), Efficient Feature Selection via Analysis of Relevance and Redundancy. JMLR, 5: 1205-1224.

Zhang, S., Zhang, C., Yang, Q. (2002). Data Preparation for Data Mining. Applied Artificial Intelligence. 17: 375 - 381.

Zhang, X., Bullard, K. M., Gregg, E. W., Beckles, G. L, et al. (2012). *Access to health care and control of ABCs of diabetes*. *Diabetes care*, *35*(7): 1566–1571. doi:10.2337/dc12-0081

Zhao D., Liu, J., Wang, M., Zhang , X., Zhou, M. (2019). *Epidemiology of cardivascular disease in China: current features and implications*. Nature Reviews Cardiology. 16: 203–212.

Zhao H.L., Tong P., Chan J.(2006). *Traditional Chinese Medicine in the Treatment of Diabetes*. Nestlé Nutr Workshop Ser Clin Perform Program. Nestec Ltd., Vevey/S. Karger AG, Basel.
11: 15-29. doi:10.1159/000094399

Zhuang, N. (n.d.). *Outline of the Healthy China 2030 plan*. Retrieved April 29, 2019, from https://www.sahealth.sa.gov.au/wps/wcm/connect/d39abd8041032c76a711ff1afc50ebfc/1645+Ning+Zhuang.pdf?MOD=AJPERES&CACHEID=ROOTWORKSPACE-d39abd8041032c76a711ff1afc50ebfc-mwMWQZ3

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). *Predicting Diabetes Mellitus With Machine Learning Techniques.* Frontiers in genetics*.* *9*: 515. doi:10.3389/fgene.2018.0051

APPENDIX A

Frequency Tables and Descriptives for Chapter 4.1

Are You Literate?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Yes	919	4,5	37,1	37,1
	2 No	1555	7,6	62,9	100,0
	Total	2474	12,1	100,0	
Missing	System	17937	87,9		
Total		20411	100,0		

Gender

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	9746	47,7	48,0	48,0
	2	10564	51,8	52,0	100,0
	Total	20310	99,5	100,0	
Missing	System	101	,5		
Total		20411	100,0		

Village_or_Town

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	2873	14,1	14,6	14,6
	2	956	4,7	4,8	19,4
	3	1049	5,1	5,3	24,7
	4	538	2,6	2,7	27,5
	5	81	,4	,4	27,9
	6	270	1,3	1,4	29,2
	7	13952	68,4	70,8	100,0
	Total	19719	96,6	100,0	
Missing	System	692	3,4		
Total		20411	100,0		

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Birthyear	20411	1910	1978	1955,54	10,590
Valid N (listwise)	20411				

APPENDIX B

Crosstabulation and Chi-Square for Activity Levels, Chapter 4.2 + 4.3

Diabetes * High_Intensity_Sport_At_Least_10_Minutes

Crosstab

			High_Intensity_Sport_At_Least_10_Minutes		
			1	2	Total
Diabetes	1	Count	119	399	518
		Expected Count	182,5	335,5	518,0
	2	Count	3349	5978	9327
		Expected Count	3285,5	6041,5	9327,0
Total		Count	3468	6377	9845
		Expected Count	3468,0	6377,0	9845,0

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	35,977[a]	1	,000		
Continuity Correction[b]	35,413	1	,000		
Likelihood Ratio	38,406	1	,000		
Fisher's Exact Test				,000	,000
Linear-by-Linear Association	35,973	1	,000		
N of Valid Cases	9845				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 182,47.

b. Computed only for a 2x2 table

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	-,060	,000
	Cramer's V	,060	,000
N of Valid Cases		9845	

Diabetes * Medium_Intensity_Sport_At_Least_10_Minutes

Crosstab

			Medium_Intensity_Sport_At_Least_10_Minutes 1	2	Total
Diabetes	1	Count	260	257	517
		Expected Count	289,2	227,8	517,0
	2	Count	5241	4077	9318
		Expected Count	5211,8	4106,2	9318,0
Total		Count	5501	4334	9835
		Expected Count	5501,0	4334,0	9835,0

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	7,049[a]	1	,008		
Continuity Correction[b]	6,810	1	,009		
Likelihood Ratio	7,004	1	,008		
Fisher's Exact Test				,008	,005
Linear-by-Linear Association	7,049	1	,008		
N of Valid Cases	9835				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 227,83.

b. Computed only for a 2x2 table

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	-,027	,008
	Cramer's V	,027	,008
N of Valid Cases		9835	

Diabetes * Low_Intensity_Sport_At_Least_10_Minutes

Crosstab

			Low_Intensity_Sport_At_Least_10_Minutes		
			1	2	Total
Diabetes	1	Count	406	112	518
		Expected Count	410,2	107,8	518,0
	2	Count	7386	1936	9322
		Expected Count	7381,8	1940,2	9322,0
Total		Count	7792	2048	9840
		Expected Count	7792,0	2048,0	9840,0

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	,217[a]	1	,641		
Continuity Correction[b]	,168	1	,682		
Likelihood Ratio	,215	1	,643		
Fisher's Exact Test				,656	,338
Linear-by-Linear Association	,217	1	,641		
N of Valid Cases	9840				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 107,81.

b. Computed only for a 2x2 table

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	-,005	,641
	Cramer's V	,005	,641
N of Valid Cases		9840	

Diabetes * High_Intensity_Sport_2_hours

Crosstab

			High_Intensity_Sport_2_hours		Total
			1	2	
Diabetes	1	Count	36	82	118
		Expected Count	23,7	94,3	118,0
	2	Count	663	2695	3358
		Expected Count	675,3	2682,7	3358,0
Total		Count	699	2777	3476
		Expected Count	699,0	2777,0	3476,0

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	8,222[a]	1	,004		
Continuity Correction[b]	7,566	1	,006		
Likelihood Ratio	7,421	1	,006		
Fisher's Exact Test				,007	,004
Linear-by-Linear Association	8,220	1	,004		
N of Valid Cases	3476				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 23,73.

b. Computed only for a 2x2 table

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	,049	,004
	Cramer's V	,049	,004
N of Valid Cases		3476	

Diabetes * Medium_Intensity_Sport_2_hours

Crosstab

			Medium_Intensity_Sport_2_hours		Total
			1	2	
Diabetes	1	Count	163	97	260
		Expected Count	126,7	133,3	260,0
	2	Count	2508	2711	5219
		Expected Count	2544,3	2674,7	5219,0
Total		Count	2671	2808	5479
		Expected Count	2671,0	2808,0	5479,0

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	21,237[a]	1	,000		
Continuity Correction[b]	20,656	1	,000		
Likelihood Ratio	21,411	1	,000		
Fisher's Exact Test				,000	,000
Linear-by-Linear Association	21,234	1	,000		
N of Valid Cases	5479				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 126,75.

b. Computed only for a 2x2 table

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	,062	,000
	Cramer's V	,062	,000
N of Valid Cases		5479	

Diabetes * Low_Intensity_Sport_2_hours

Crosstab

			Low_Intensity_Sport_2_hours 1	2	Total
Diabetes	1	Count	285	120	405
		Expected Count	250,5	154,5	405,0
	2	Count	4504	2833	7337
		Expected Count	4538,5	2798,5	7337,0
Total		Count	4789	2953	7742
		Expected Count	4789,0	2953,0	7742,0

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	13,127[a]	1	,000		
Continuity Correction[b]	12,749	1	,000		
Likelihood Ratio	13,568	1	,000		
Fisher's Exact Test				,000	,000
Linear-by-Linear Association	13,125	1	,000		
N of Valid Cases	7742				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 154,48.

b. Computed only for a 2x2 table

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	,041	,000

	Cramer's V	,041	,000
N of Valid Cases		7742	

Diabetes * High_intensity_Sport_30_minutes

Crosstab

			High_intensity_Sport_30_minutes 1	2	Total
Diabetes	1	Count	5	30	35
		Expected Count	8,3	26,7	35,0
	2	Count	162	508	670
		Expected Count	158,7	511,3	670,0
Total		Count	167	538	705
		Expected Count	167,0	538,0	705,0

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1,801[a]	1	,180		
Continuity Correction[b]	1,295	1	,255		
Likelihood Ratio	2,003	1	,157		
Fisher's Exact Test				,223	,125
Linear-by-Linear Association	1,798	1	,180		
N of Valid Cases	705				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 8,29.

b. Computed only for a 2x2 table

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	-,051	,180
	Cramer's V	,051	,180
N of Valid Cases		705	

Diabetes * Medium_Intensity_Sport_30_minutes

Crosstab

			Medium_Intensity_Sport_32_minutes 1	2	Total
Diabetes	1	Count	45	119	164
		Expected Count	40,9	123,1	164,0
	2	Count	624	1895	2519
		Expected Count	628,1	1890,9	2519,0
Total		Count	669	2014	2683
		Expected Count	669,0	2014,0	2683,0

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	,585[a]	1	,444		
Continuity Correction[b]	,451	1	,502		
Likelihood Ratio	,574	1	,449		
Fisher's Exact Test				,456	,248
Linear-by-Linear Association	,585	1	,444		
N of Valid Cases	2683				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 40,89.

b. Computed only for a 2x2 table

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	,015	,444
	Cramer's V	,015	,444
N of Valid Cases		2683	

Diabetes * Low_Intensity_Sport_30_minutes

Crosstab

			Low_Intensity_Sport_32_minutes		
			1	2	Total
Diabetes	1	Count	50	236	286
		Expected Count	69,5	216,5	286,0
	2	Count	1118	3401	4519
		Expected Count	1098,5	3420,5	4519,0
Total		Count	1168	3637	4805
		Expected Count	1168,0	3637,0	4805,0

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	7,700[a]	1	,006		
Continuity Correction[b]	7,311	1	,007		
Likelihood Ratio	8,247	1	,004		
Fisher's Exact Test				,005	,003
Linear-by-Linear Association	7,698	1	,006		
N of Valid Cases	4805				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 69,52.

b. Computed only for a 2x2 table

Symmetric Measures

	Value	Approximate Significance

Nominal by Nominal	Phi	-,040	,006
	Cramer's V	,040	,006
N of Valid Cases		4805	

Diabetes * High_Intensity_Sport_4_hours

Crosstab

			High_Intensity_Sport_4_hours		
			1	2	Total
Diabetes	1	Count	20	62	82
		Expected Count	23,1	58,9	82,0
	2	Count	762	1933	2695
		Expected Count	758,9	1936,1	2695,0
Total		Count	782	1995	2777
		Expected Count	782,0	1995,0	2777,0

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	,594[a]	1	,441		
Continuity Correction[b]	,417	1	,518		
Likelihood Ratio	,611	1	,435		
Fisher's Exact Test				,533	,263
Linear-by-Linear Association	,593	1	,441		
N of Valid Cases	2777				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 23,09.

b. Computed only for a 2x2 table

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	-,015	,441
	Cramer's V	,015	,441
N of Valid Cases		2777	

Diabetes * Medium_Intensity_Sport_4_hours

Crosstab

			Medium_Intensity_Sport_4_hours		
			1	2	Total
Diabetes	1	Count	56	41	97
		Expected Count	46,0	51,0	97,0
	2	Count	1276	1433	2709
		Expected Count	1286,0	1423,0	2709,0
Total		Count	1332	1474	2806
		Expected Count	1332,0	1474,0	2806,0

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	4,243[a]	1	,039		
Continuity Correction[b]	3,828	1	,050		
Likelihood Ratio	4,244	1	,039		
Fisher's Exact Test				,049	,025
Linear-by-Linear Association	4,242	1	,039		
N of Valid Cases	2806				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 46,05.

b. Computed only for a 2x2 table

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	,039	,039
	Cramer's V	,039	,039
N of Valid Cases		2806	

Diabetes * Low_Intensity_Sport_4_hours

Crosstab

			Low_Intensity_Sport_4_hours		Total
			1	2	
Diabetes	1	Count	86	35	121
		Expected Count	70,8	50,2	121,0
	2	Count	1652	1196	2848
		Expected Count	1667,2	1180,8	2848,0
Total		Count	1738	1231	2969
		Expected Count	1738,0	1231,0	2969,0

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	8,168[a]	1	,004		
Continuity Correction[b]	7,638	1	,006		
Likelihood Ratio	8,505	1	,004		
Fisher's Exact Test				,005	,002
Linear-by-Linear Association	8,165	1	,004		
N of Valid Cases	2969				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 50,17.

b. Computed only for a 2x2 table

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	,052	,004
	Cramer's V	,052	,004
N of Valid Cases		2969	

APPENDIX C

Crosstabulation and Chi-Square for Illnesses, Chapter 4.2 + 4.3

Diabetes * Hyptertension

Crosstab

			Hyptertension		
			1	2	Total
Diabetes	1	Count	630	487	1117
		Expected Count	230,9	886,1	1117,0
	2	Count	3704	16146	19850
		Expected Count	4103,1	15746,9	19850,0
Total		Count	4334	16633	20967
		Expected Count	4334,0	16633,0	20967,0

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	918,587[a]	1	,000		
Continuity Correction[b]	916,287	1	,000		
Likelihood Ratio	731,922	1	,000		
Fisher's Exact Test				,000	,000
Linear-by-Linear Association	918,543	1	,000		
N of Valid Cases	20967				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 230,89.

b. Computed only for a 2x2 table

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	,209	,000
	Cramer's V	,209	,000
N of Valid Cases		20967	

Diabetes * Dyslipedimia

Crosstab

			Dyslipedimia 1	2	Total
Diabetes	1	Count	509	608	1117
		Expected Count	98,8	1018,2	1117,0
	2	Count	1346	18504	19850
		Expected Count	1756,2	18093,8	19850,0
Total		Count	1855	19112	20967
		Expected Count	1855,0	19112,0	20967,0

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1972,815[a]	1	,000		
Continuity Correction[b]	1968,009	1	,000		
Likelihood Ratio	1155,159	1	,000		
Fisher's Exact Test				,000	,000
Linear-by-Linear Association	1972,721	1	,000		
N of Valid Cases	20967				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 98,82.

b. Computed only for a 2x2 table

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	,307	,000
	Cramer's V	,307	,000
N of Valid Cases		20967	

Diabetes * Cancer

Crosstab

			Cancer		
			1	2	Total
Diabetes	1	Count	26	1091	1117
		Expected Count	11,2	1105,8	1117,0
	2	Count	184	19666	19850
		Expected Count	198,8	19651,2	19850,0
Total		Count	210	20757	20967
		Expected Count	210,0	20757,0	20967,0

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	20,925[a]	1	,000		
Continuity Correction[b]	19,536	1	,000		
Likelihood Ratio	15,569	1	,000		
Fisher's Exact Test				,000	,000
Linear-by-Linear Association	20,924	1	,000		
N of Valid Cases	20967				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 11,19.

b. Computed only for a 2x2 table

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	,032	,000
	Cramer's V	,032	,000
N of Valid Cases		20967	

Diabetes * Chronic_Lung_Disease

Crosstab

			Chronic_Lung_Disease		
			1	2	Total
Diabetes	1	Count	162	955	1117
		Expected Count	105,3	1011,7	1117,0
	2	Count	1814	18036	19850
		Expected Count	1870,7	17979,3	19850,0
Total		Count	1976	18991	20967
		Expected Count	1976,0	18991,0	20967,0

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	35,652[a]	1	,000		
Continuity Correction[b]	35,027	1	,000		
Likelihood Ratio	31,365	1	,000		
Fisher's Exact Test				,000	,000
Linear-by-Linear Association	35,651	1	,000		
N of Valid Cases	20967				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 105,27.

b. Computed only for a 2x2 table

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	,041	,000
	Cramer's V	,041	,000
N of Valid Cases		20967	

Diabetes * Liver_disease

Crosstab

			Liver_disease		
			1	2	Total
Diabetes	1	Count	94	1023	1117
		Expected Count	43,1	1073,9	1117,0
	2	Count	715	19135	19850
		Expected Count	765,9	19084,1	19850,0
Total		Count	809	20158	20967
		Expected Count	809,0	20158,0	20967,0

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	66,047[a]	1	,000		
Continuity Correction[b]	64,756	1	,000		
Likelihood Ratio	50,848	1	,000		
Fisher's Exact Test				,000	,000
Linear-by-Linear Association	66,044	1	,000		
N of Valid Cases	20967				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 43,10.

b. Computed only for a 2x2 table

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	,056	,000
	Cramer's V	,056	,000
N of Valid Cases		20967	

Diabetes * Heart_Attack

Crosstab

			Heart_Attack 1	2	Total
Diabetes	1	Count	312	805	1117
		Expected Count	118,3	998,7	1117,0
	2	Count	1909	17941	19850
		Expected Count	2102,7	17747,3	19850,0
Total		Count	2221	18746	20967
		Expected Count	2221,0	18746,0	20967,0

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	374,540[a]	1	,000		
Continuity Correction[b]	372,609	1	,000		
Likelihood Ratio	278,446	1	,000		
Fisher's Exact Test				,000	,000
Linear-by-Linear Association	374,523	1	,000		
N of Valid Cases	20967				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 118,32.

b. Computed only for a 2x2 table

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	,134	,000
	Cramer's V	,134	,000
N of Valid Cases		20967	

Diabetes * Stroke

Crosstab

			Stroke 1	Stroke 2	Total
Diabetes	1	Count	72	1045	1117
		Expected Count	23,5	1093,5	1117,0
	2	Count	369	19481	19850
		Expected Count	417,5	19432,5	19850,0
Total		Count	441	20526	20967
		Expected Count	441,0	20526,0	20967,0

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	108,055[a]	1	,000		
Continuity Correction[b]	105,839	1	,000		
Likelihood Ratio	72,429	1	,000		
Fisher's Exact Test				,000	,000
Linear-by-Linear Association	108,050	1	,000		
N of Valid Cases	20967				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 23,49.

b. Computed only for a 2x2 table

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	,072	,000
	Cramer's V	,072	,000
N of Valid Cases		20967	

Diabetes * Kidney_disease

Crosstab

			Kidney_disease 1	Kidney_disease 2	Total
Diabetes	1	Count	151	966	1117
		Expected Count	67,4	1049,6	1117,0
	2	Count	1114	18736	19850
		Expected Count	1197,6	18652,4	19850,0
Total		Count	1265	19702	20967
		Expected Count	1265,0	19702,0	20967,0

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	116,598[a]	1	,000		
Continuity Correction[b]	115,208	1	,000		
Likelihood Ratio	89,620	1	,000		
Fisher's Exact Test				,000	,000
Linear-by-Linear Association	116,593	1	,000		
N of Valid Cases	20967				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 67,39.

b. Computed only for a 2x2 table

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	,075	,000
	Cramer's V	,075	,000
N of Valid Cases		20967	

Diabetes * Stomage_disease

Crosstab

			Stomage_disease		
			1	2	Total
Diabetes	1	Count	288	829	1117
		Expected Count	243,9	873,1	1117,0
	2	Count	4291	15559	19850
		Expected Count	4335,1	15514,9	19850,0
Total		Count	4579	16388	20967
		Expected Count	4579,0	16388,0	20967,0

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	10,753[a]	1	,001		
Continuity Correction[b]	10,511	1	,001		
Likelihood Ratio	10,353	1	,001		
Fisher's Exact Test				,001	,001
Linear-by-Linear Association	10,753	1	,001		
N of Valid Cases	20967				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 243,94.

b. Computed only for a 2x2 table

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	,023	,001
	Cramer's V	,023	,001
N of Valid Cases		20967	

Diabetes * Other_Medical_Diseases

Crosstab

			Other_Medical_Diseases		
			1	2	Total
Diabetes	1	Count	73	1041	1114
		Expected Count	88,3	1025,7	1114,0
	2	Count	1584	18218	19802
		Expected Count	1568,7	18233,3	19802,0
Total		Count	1657	19259	20916
		Expected Count	1657,0	19259,0	20916,0

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	3,024[a]	1	,082		
Continuity Correction[b]	2,829	1	,093		
Likelihood Ratio	3,189	1	,074		
Fisher's Exact Test				,090	,046
Linear-by-Linear Association	3,024	1	,082		
N of Valid Cases	20916				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 88,25.

b. Computed only for a 2x2 table

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	-,012	,082
	Cramer's V	,012	,082
N of Valid Cases		20916	

APPENDIX D

Crosstabulation and Chi-Square for Drinking and Smoking, Chapter 4.2 + 4.3

Diabetes * Gender

Crosstab

			Gender		
			1	2	Total
Diabetes	1	Count	483	634	1117
		Expected Count	531,5	585,5	1117,0
	2	Count	9489	10350	19839
		Expected Count	9440,5	10398,5	19839,0
Total		Count	9972	10984	20956
		Expected Count	9972,0	10984,0	20956,0

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	8,929[a]	1	,003		
Continuity Correction[b]	8,746	1	,003		
Likelihood Ratio	8,964	1	,003		
Fisher's Exact Test				,003	,002
Linear-by-Linear Association	8,929	1	,003		
N of Valid Cases	20956				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 531,53.

b. Computed only for a 2x2 table

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	-,021	,003
	Cramer's V	,021	,003
N of Valid Cases		20956	

Diabetes * Drink

Crosstab

			Drink			
			1	2	3	Total
Diabetes	1	Count	217	73	823	1113
		Expected Count	296,1	98,1	718,8	1113,0
	2	Count	5346	1770	12684	19800
		Expected Count	5266,9	1744,9	12788,2	19800,0
Total		Count	5563	1843	13507	20913
		Expected Count	5563,0	1843,0	13507,0	20913,0

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	45,016[a]	2	,000
Likelihood Ratio	47,140	2	,000
Linear-by-Linear Association	41,499	1	,000
N of Valid Cases	20913		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 98,09.

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	,046	,000
	Cramer's V	,046	,000
N of Valid Cases		20913	