MACHINE LEARNING TO SUMMON EFFECTIVE MEDICATION PRACTICE

Rajath Akshay Vanikul

TABLE OF CONTENTS

1. PROJECT DESCRIPTION

1.1 PROJECT MOTIVATION

Medication is the art of curing or preventing illness. The evolution of modern medicine represents a tremendous positive change in the approach to treating illness and disease. The rise of modern medicine has contributed immensely towards identifying the procedures to treat several diseases and prevent the spread of infectious diseases. Most medicines are prepared and targeted towards an average individual suffering the illness and are recently perceived as an ineffective process to treat an individual. Medication errors are equally recurrent in general practice which harms the patient and creates an unsafe treating environment. However, current technologies are enhancing the field of precision medicines which targets the individualized treatment based on genetic, phenotypic, biomarker, or psychosocial characteristics. (Jameson, 2015)

Every drug uniquely interacts with each human body, so what works well for one individual may not help the another. Precision medicines are often referred to as the future of healthcare as it aims to tailor the medication as per the condition of the individual. According to "Trends in Precision Medicine Adoption" report by Oracle states (Issa, 2014) a majority of healthcare organizations in the world have planned to actively participate in precision medicine studies and invest on developing the technology in the next two years.

1.2 PROJECT DEFINITION

Considering the advantages of precision medicine and the issue of recurring medication errors, this project proposes a model to gather personalised data which can be processed individually using machine learning tools to address the treatment. This model aims to constantly analyse the patterns in symptoms and effect of medication during the treatment.

Project addresses the following issues and suggests definitive solution.

- This model promises to design an interactive platform to harvest personalised data from smart monitoring devices which can be suggested by healthcare right from the first consultation which diagnosis the effects of medication.
- Analyse individual body reactions to the medicines at each stage of the medication and find patterns to make better informed decisions by doctors.
- Build a machine learning algorithm that tracks the patterns in the electronic medical records and suggests the possible outcome of the situation to help develop precision medicine at the earliest.

This model also shows the scope in harvesting enormous amount data from the users which will eventually help us get efficient at approximating the chances which promises

faster and better medication decisions. This can help practice an effective process to diagnose the patient at the early stages of illness.

1.3 ROLES AND RESPONSIBILITIES

The following is recognised to be the major data science roles and their responsibilities.

- **Data Engineer:** Integration of smart monitoring devices needs a consolidated single repository with specific data format. This will be engineered and the requirement set is generated by data engineers to facilitate data management.
- **Data Archivist/Curator:** To ensure the data from different integrated sources are monitored and stored in the right location. Data curator is necessary for this project to gather data from different sources which also involves storage and maintenance of the data.
- **Database Developer:** A designated personnel is necessary to focus on improving databases and modifying legacy applications to work with a database setup. Database Developers is also necessary to design database systems to handle raw and unorganized data for research.
- **Machine Learning Engineer:** The platform is based on the machine learning algorithm built to recognise patterns in the drug interactions. Machine learning engineer is vital in developing better training algorithms to result in better prediction results.
- **Data Scientist:** Monitoring the algorithm and developing informative insights will be performed by a data scientist. Visualising the data and the progress of the project will be profoundly showcased by a data scientist.
- **Data Security Administrator:** As the model deals with personal data about individuals, we will need a dedicated security administrator to ensure data security.

2. BUSINESS MODEL

The project focuses on complete lifecycle of medical treatment. This model fortifies the evolution of the modern diagnose and medicine which ensures to embrace technology to stand out from the existing practices.

- Smart monitoring devices to be developed to detect vital variables like heart rate, body temperature and advice the patient to use them during the course of medication. This will help in collecting data on the effects of drug on the body and help us identify the potential risks from the data monitored during the time between the consultations.

- The data from the smart monitoring devices can be gathered and analysed using the sophisticated machine learning tools to identify behaviours and patterns to make better informed decisions/prescriptions.
- Precision medicine can help us deliver better medicine tailored to the individual's condition which is analysed by our algorithm.

Figure 1: Influence diagram of the business proposal

This business model specifically addresses the evolution of the healthcare sector. The following domains could earn the most benefit from the proposed project:

- **Clinical Research and Development**: Machine learning technology can help in accelerating clinical trial procedures which usually take several years. The algorithms can predict the outcomes of most of the clinical trials in a shorter duration.
- **Hospitals**: The proposed procedure will definitely improve the quality of treatment by utilizing the effective diagnosing method in hospitals. Successful treatments will impact the industry financially and can result in higher market share.
- **Pharmaceutical Industry**: The huge data gathered over the period of time will ensure better-tailored medicine production. This project also moulds the exposure on precision medicine area of research which can show snowball effect on custom medication.

3. CHARACTERISING AND PROCESSING DATA

Characterization is an effective description of the characteristics and behaviour of a particular observed data. Characterising data helps unsupervised learning algorithms to find trends and patterns without any bias. Below are the different parameters that are characterised in our project:

- **Volume**: All the smart devices are equipped to communicate and exchange data with servers and computers. Usage of smart devices to monitor health has increased exponentially over the past decade. According to statistics from EMC, the total digital data created by smart devices in 2013 was 4.4 zettabytes, 21 zettabytes in 2017 and is estimated to be 44 zettabytes by 2020 (also equal to 44 trillion gigabytes). It also estimated that about 0.2% of 44 zettabytes of data is related to the data generated by smart devices that classifies itself as health data which accounts to around **0.088 zettabytes** of data by 2020. (Kanellos, 2016)

- **Velocity**: IBM claims the world creates 2.5 quintillion bytes of data every day. Considering the factor of 0.2% as discussed earlier, Smart devices can be estimated to generate up to **0.005 quintillion bytes of data per day.** (Bresnick, 2017)

- **Variety**: We have various health monitoring devices like Libre reader, Qardiocode, fitness trackers to measure and share various health measured data which has different ranges and measures that can be standardised.

- **Variability**: A smart device measures different values with each given moment. This model will receive a wide spread of variable data from several different devices in a given time. A unified platform needs to be developed to handle the diversity in data.

- **Veracity**: Measures and data generated by non-invasive smart devices can be contaminated by external factors like temperature, humidity and electronic support. Issues created by external factors need to be handled to arrive at a meaningful data. Data scientists spend about 60% of the time cleaning the data. Health care usually employs **data governance and information governance** to ensure their data is trustworthy, complete, clean and standardized.

- **Visualization**: The data needs to be presented using plots and summaries which can be conveyed to practitioners and public. Installations of the **interactive dashboard** at consultation site to enable a better understanding of the data and patterns during diagnosing.

- **Value**: Size, complexity and critical health records of individuals can be used to treat individuals with **precision medicine**. The collective data can further be used to accelerate the process of diagnosing and will potentially be used at **clinical research**.

Data Processing is referred to collecting innumerable data from different sources and arrange them in a way that is practically beneficial for the purpose.

- We need a unified platform with **API (Application programming interface)** to establish a communication between all the smart devices and our database to gather real-time measured data.
- Deploying an **in-house development team** to build a customized API with ensure tailored product and will ensure better security.
- Hiring an outsourced cloud computing platform to assist us with higher scalability with respect to storage

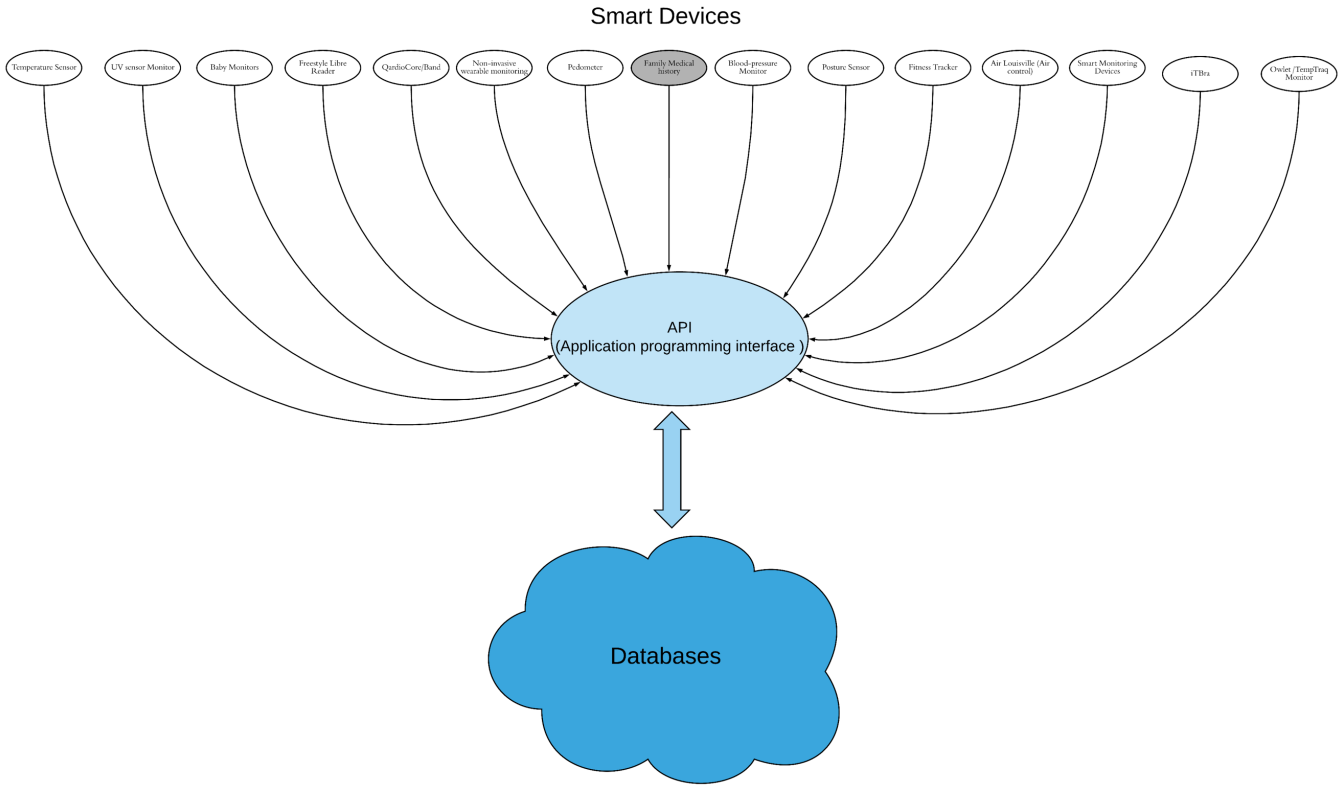

Figure 2: Diagram to depict API model

4. RESOURCES

Potential data sources for the project and expected data processing tools are explained below:

- **Data Sources**: This project accepts most of the smart devices that is used to track health behaviour. Major sources and their functions are listed as follows:
  - ⇨ **Freestyle libre reader**: Device used to track glucose levels in the body.
  - ⇨ **QardioCore/KardiaBand**: Wearable device around the chest that tracks ECG and EKG of an individual.
  - ⇨ **Pedometer**: Non-invasive wearable device on the wrist/ankle used to track pulse, fitness activities and sleep cycles.
  - ⇨ **Tempe Sensor**: A device that transmits ambient temperature data wirelessly to a designated machine.
  - ⇨ **Blood Pressure Monitor**: Wrist wearable device that can display BP of an individual and transmit the data to the synchronised computer.

- ⇨ **Posture Sensor**: A device that tracks postures of spine, leg, arms and neck. This can help us address the issues specific to an individual's lifestyle.
- ⇨ **Air Louisville**: This monitors the quality of air around the individual.
- ⇨ **iTBra**: Keeps a track of unusual developments that can lead to breast cancer.
- ⇨ **Baby Monitor**: Device that tracks metabolic and sleep cycles in babies.

- **System Compute**: Rackspace Cloud is the suggested cloud computing platform as it is highly customizable, secure and easy to scale.
- **Software**: Python, R, Hadoop, Teradata, Hive, MongoDB.
- **Analytics**: Standardisation of incoming data, general data analysis and visualisation for data. Sisense would be an appropriate tool for the project to support medical intelligence solutions.
- **Processing**: Knime is an analytics platform that can help us in analytic processing, handling the volume.
- **Storage/security**: Rackspace Cloud can be used to store and is known for privacy and confidentiality.

5. DATA ANALYSIS

The project deals with transforming the raw data gathered from different sources to a relatable structured data which can assist in building predictive models to develop precision medicine and a strong visualisation to help doctors and patients understand the developments during the treatment.

5.1 DATA WRANGLING

Wrangling is a process of mapping and transforming data from raw data source to a standardized structured data record. In the project wrangling plays a major role in maintaining and evaluating the following points:

- Add data and integrate various systems with caution
- Make sure that the right devices are communicating with the right end system
- Identify and link incoming data with right time and history
- Maintain an accurate timestamp for each data point and source

Thus wrangling will ensure the constant structuration and flow of data. We recommend the use of Trifacta Wrangler along with common scripting languages like Python and Perl. Data wrangling using Trifacta wrangler will be advantageous to combine and leverage new data sources in the existing data model.

5.2 UNSUPERVISED MACHINE LEARNING

We now have data retrieved from various different interest sources from a given individual, we deploy unsupervised learning technique in parallel to supervised learning technique to find naturally occurring groupings and patterns within the data

which wouldn't be possible to visible to human eye. These patterns are scaled and matched with the existing medical data for similar medicines for a faster and accurate diagnosing process. Thus, the unsupervised algorithm will ensure to enhance the accuracy of diagnosing and prescription process during the consultation.

5.3 SUPERVISED MACHINE LEARNING

Supervised learning technique is performed to predict a known outcome. During diagnosing, a physician/medical expert suspects a likely risk that needs careful analysis, we employ supervised learning to predict the development of the condition and confirm the risk occurrence. Supervised learning algorithms are often practised in medicine to recognize patterns in electrocardiogram (ECG), automate detection of a lung nodule from a chest X-ray and perform Framingham Risk Score to detect coronary heart disease (CHD). The supervised model incorporated in this project would include all the algorithms that can help us to analyse data from all the above-mentioned sources infer and predict conclusions on the diagnosing.

5.4 PREDICTIVE ANALYTICS

Advanced analysis will be performed on the existing data to make predictions about unknown future developments of the situation. A dedicated advanced analytics is included in the project to support and prepare precision medicine production and prescription. Predictive Analysis will play a major role in anticipating the demand and likelihood of a certain class of medicine and to study its effects.

5.5 CAUSAL ANALYSIS

Major motivation of this analysis is to find causes that can treat a condition rather than treating the symptoms. All the above-mentioned algorithms analyse the effects and predict the outcome. However, causal analysis ensures to:

- Identify the specific cause of the issue that will be explored in detail
- Provide evidence for the causal claims
- Demonstrate an understanding between the relations and provide explanations for the claim.

Involvement of causal analysis in medical practices always increases the quality of the analysis conducted. Thus, The project aims for a better quality of output from the analysis performed.

6. SUMMARY

Implementation of the proposed data science project will ensure an efficient approach for diagnosing and treatment. Combination of unsupervised learning, supervised learning, predictive analysis and causal analysis in right proportions to arrive at data-driven conclusions will promise a strong evidence to make informed decisions to produce precision medicine and transform the complexion of treating an individual. The project also fortifies a potential to generate huge volumes of data from various different sources that can also support clinical research for better medicine and to predict the future health risks of a community.

7. REFERENCES

- Bresnick, J. (2017, June 05). *Healthcare Big Data Analytics*. Retrieved from HealthITAnalytics: https://healthitanalytics.com/news/understanding-the-many-vs-of-healthcare-big-data-analytics
- Issa, N. T. (2014). *Big data: the next frontier for innovation in therapeutics and healthcare.* Expert review of clinical pharmacology.
- Jameson, J. L. (2015). *Precision medicine—personalized, problematic, and promising.* Obstetrical & Gynecological Survey.
- Kanellos, M. (2016, March 3). *IDC Outlines The Future of Smart Things*. Retrieved from Forbes: https://www.forbes.com/sites/michaelkanellos/2016/03/03/152000-smart-devices-every-minute-in-2025-idc-outlines-the-future-of-smart-things/#8bd640c4b63e
- Stoneburner, G. G. (2009). *Risk Management Guide for Information Technology Systems.* National Institute of Standards and Technology.