Mr. Caduceus: Predicting Diseases Using Patient Data for Health-Adjusted Taxation

Aapo Asp, Keith Davis, Sasu Mäkinen, and Tri Nguyen

University of Helsinki, Helsinki 00560, FI
firstname.lastname@helsinki.fi

Abstract. Finland's healthcare system is primarily supported by a healthcare tax, based on annual income, that is collected from working Finns. Currently, this method of taxation does not reward individuals who live a healthier lifestyle than their less active, less healthy peers. Mr. Caduceus is a proposed solution that seeks to adjust the taxation system so that individuals' perceived levels of health will also be taken into consideration. This report explains how Mr. Caduceus was built, methods used to implement fairness-aware machine learning, and the legality of the system.

Keywords: healthcare, health tech, fairness aware ai, machine learning

1 Introduction

Given an array of health data for a patient, including the a list of diseases or medical conditions they are suffering from, it is possible to predict the likelihood of diseases for other patients for which similar data is available. While the morality of taxing the sick and the elderly *more* than their healthier and (often) younger counterparts is worth seriously questioning, such a discussion is outside the scope of this paper. The intention of this project was to confirm our hypothesis that, given enough medical data, one can simply train a classifier to predict medical conditions. Additionally, could we construct such a classifier so that it does not unfairly discriminate against healthy individuals based on protected variables or attributes. All figures for this report are available in the appendix.

2 Data

The dataset used comes from the 2014 National Health and Nutrition Examination Survey, as conducted by the National Center for Health Statistics in the United States. The survey data includes interviews regarding patient demographics, nutritional profile, health history, as well as the results of physical examinations, laboratory tests, and medication profiles, linked to a unique patient ID. It contains over 10K unique patient IDs.

First, the data had to be transformed to suit the purposes of this project. The different inputs (demography, medication, labs, diet) were joined together, grouped into a patient profile (by ID). Structured in this way, each row represents a unique patient, and each column represents a variable related to some medical information. Disease status was inferred using the responses of certain questions, as in Fig. 1.

Approximately 200 columns were used to encode for 169 different diseases, with "Yes" responses indicating a patient had a given disease, and all other responses indicating the opposite. The indicator columns were then dropped from the dataset. The resulting data was saved as our "disease indicator" dataset. The pipeline is depicted in Fig. 2. Next, the raw inputs were again combined, and columns with more than 35% blank were removed. After this, rows with more than 50% blank were dropped. For the remaining observations, we replaced blank observations with the column mean (for integer columns) and "no category" for category columns. This data was saved as our "patient data" dataset. This process is depicted in Fig. 3.

An exploratory analysis of the data was then conducted. We found that a majority of patients suffered from at least one illness or medical condition (see Fig. 4), and that the most common ailments were chronic conditions (see Fig. 5). Patients were generally older and female; the patient dataset contained approximately 750 more women than men (see Fig. 6). As is expected, the prevalences of diseases over time increases as patients age (see Fig. 7).

3 Initial Models

A variety of different methods were used to attempt to predict medical conditions: one-vs-all, K-nearest neighbors, random forest, and a densely connected feedforward neural network. The initial performance is shown in Fig. 8. We experimented with removing age and gender variables from the training data, and training performance improved slightly, as in Fig. 9. An ensemble of random forest classifiers, each predicting for a single medical conditions, performed similarly to the neural network. The neural network gave the best performance, typically with an F1-score between 0.50 and 0.60. However, no sensitive variables were removed when training the neural network.

4 Legal Justification - Test-Aschats

After reviewing the legal literature and our dataset, we identified and removed sensitive variables. Sensitive variables include the following:

- Race
- Country of birth
- Language of SP interview
- Marital Status

Additionally, variables that were initially classified as sensitive, but were later determined to not be sensitive, are as follows:

- Age in years at screening
- Gender of the participant
- Citizenship status
- Education level
- Annual household income

Despite our initial assumptions, age and gender, in the context of medical care and medical costs, are not necessarily protected variables. Referencing the Test-Aschats ruling, Case C-236/09 (Grand Chamber decision), paragraph 12 states the following: *Direct discrimination occurs only when one person is treated less favourably, on grounds of sex, than another person in a comparable situation. Accordingly, for example, differences between men and women in the provision of healthcare services, which result from the physical differences between men and women, do not relate to comparable situations and therefore, do not constitute discrimination.* In this manner, we are free to use age and gender in our models, as it directly influences the nature of medical diagnosis and treatment.

The Supreme Court of Estonia has held that only arbitrary differential treatment of two persons or groups of persons in essentially similar situations is considered as a violation of the general right to equality. In this manner, Mr. Caduceus does not violate the general right to equality, as it can be argued that predicting medical conditions from a large array of medical data (including laboratory tests) is rational and non-arbitrary.

For General Data Protection Regulation requirements, it is not immediately apparent whether or not this violates the regulation. Participants willingly gave their medical information for this study, and personally identifying information has been removed. Nonetheless, if Mr. Caduceus were implemented by the Finnish government, GDPR requirements would be subject to further debate.

5 Fairness-Aware Models

In the second round of model development, an ensemble of random forests and a feedforward neural network were trained and tested. Removing all sensitive variables, even those that were considered "legal" by the legal review, produced reasonable results (see Fig. 10 and Fig. 11).

Without removing sensitive variables, the neural network's F1 performance was as follows:

- min: 0.45
- max: 0.51
- avg: 0.47

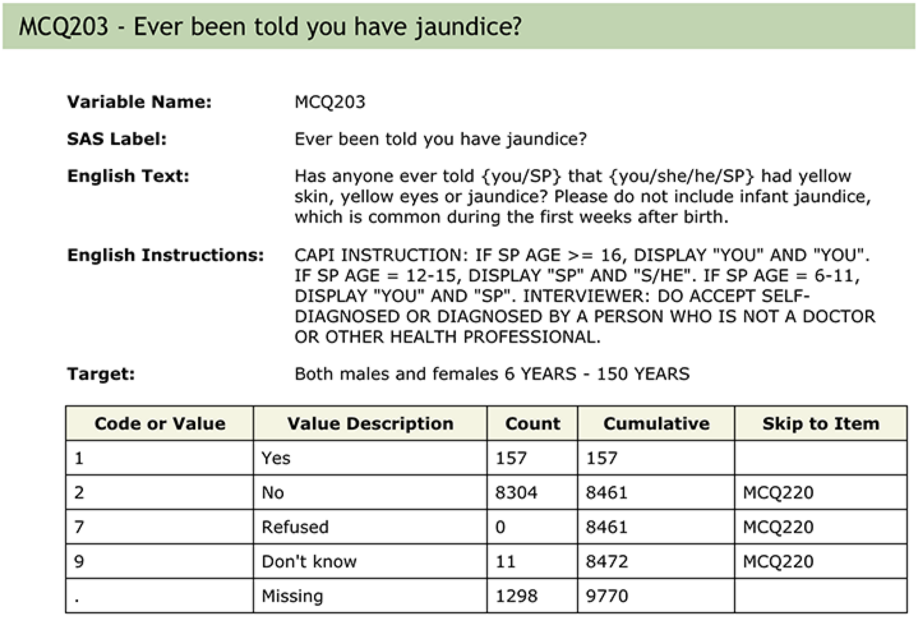
Rsensitive variables, the neural network's F1 performance did not change much:

- min: 0.42
- max: 0.50
- avg: 0.46

6 Conclusion

Despite using information for only approximately 3K patients, Mr. Caduceus performed surprisingly well in its prediction task. Due to the sheer number of medically-relevant variables available, removing sensitive variables did not significantly impact model performance. In some instances, such as in age and gender, removing sensitive variables improved the performance of our models. The models could be improved further by using additional years from the NHANES dataset, as this would expand our dataset by several thousand patients per year.

A Figures



MCQ203 - Ever been told you have jaundice?

Variable Name: MCQ203

SAS Label: Ever been told you have jaundice?

English Text: Has anyone ever told {you/SP} that {you/she/he/SP} had yellow skin, yellow eyes or jaundice? Please do not include infant jaundice, which is common during the first weeks after birth.

English Instructions: CAPI INSTRUCTION: IF SP AGE >= 16, DISPLAY "YOU" AND "YOU". IF SP AGE = 12-15, DISPLAY "SP" AND "S/HE". IF SP AGE = 6-11, DISPLAY "YOU" AND "SP". INTERVIEWER: DO ACCEPT SELF-DIAGNOSED OR DIAGNOSED BY A PERSON WHO IS NOT A DOCTOR OR OTHER HEALTH PROFESSIONAL.

Target: Both males and females 6 YEARS - 150 YEARS

Code or Value	Value Description	Count	Cumulative	Skip to Item
1	Yes	157	157	
2	No	8304	8461	MCQ220
7	Refused	0	8461	MCQ220
9	Don't know	11	8472	MCQ220
.	Missing	1298	9770	

Fig. 1. Example from NHANES data dictionary

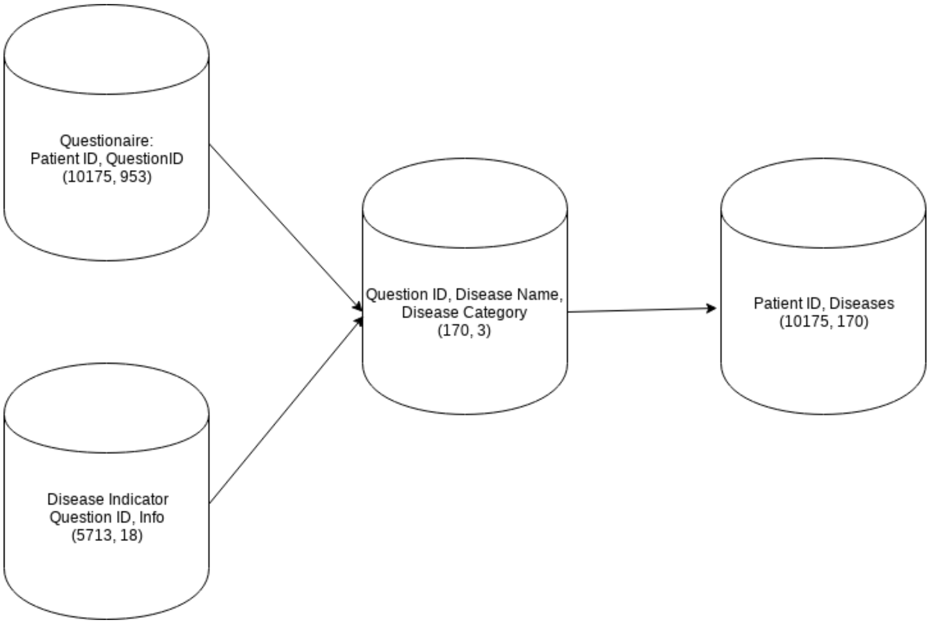


Fig. 2. Data featurization pipeline: patient diseases

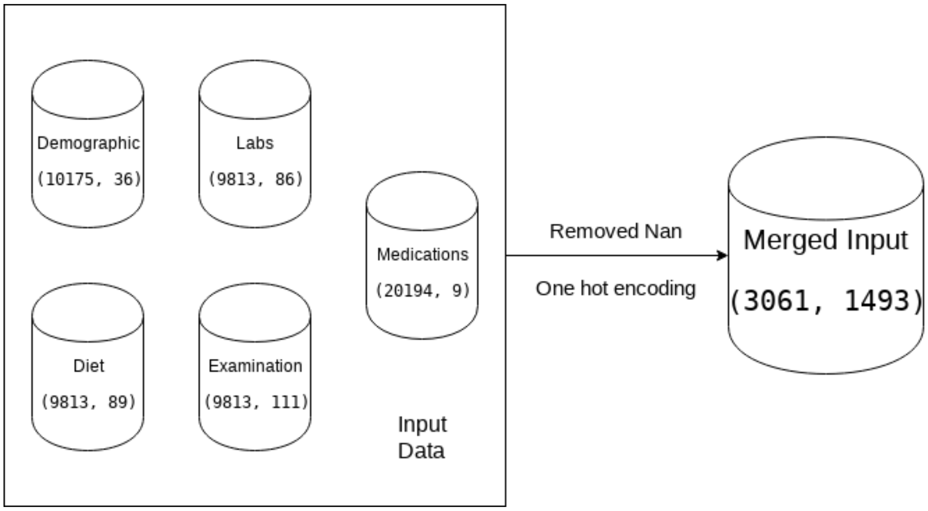


Fig. 3. Data featurization pipeline: filtering for patients with complete data

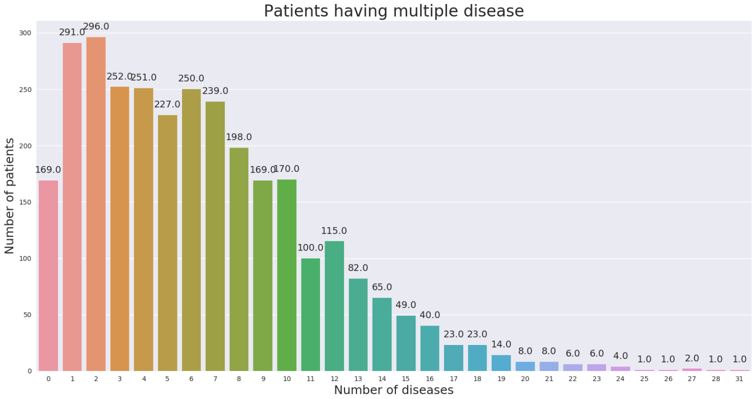


Fig. 4. Most patients had one or more medical condition

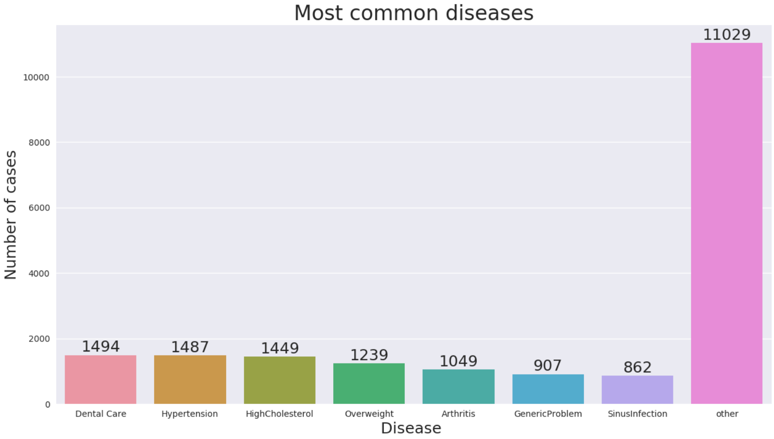


Fig. 5. The most common conditions were chronic ailments

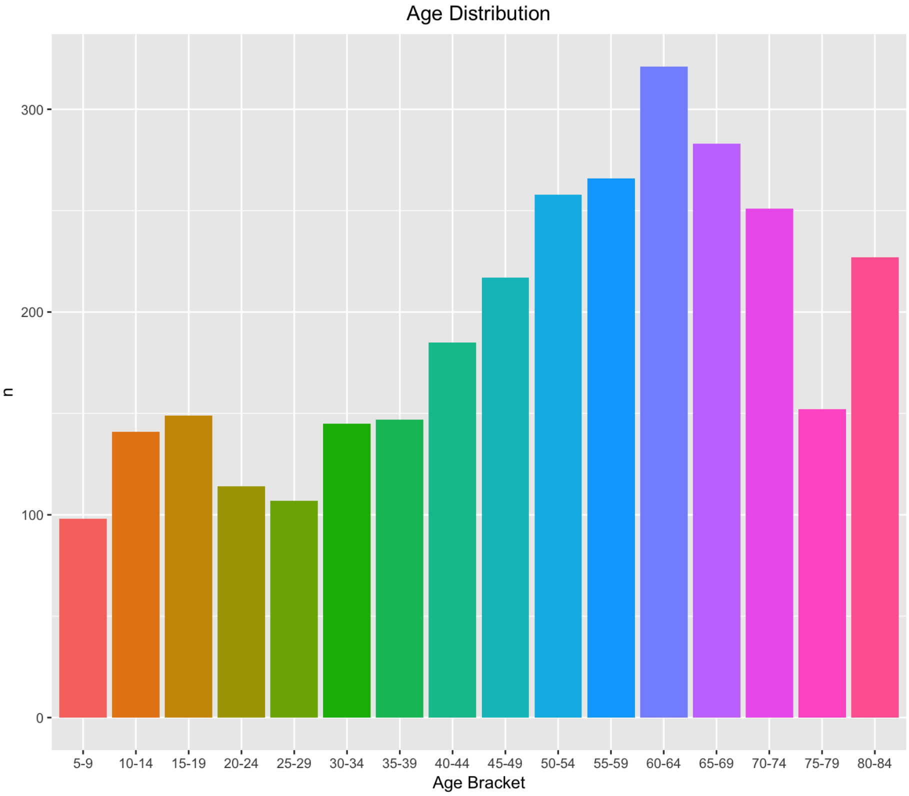


Fig. 6. The most common patient type is female, over the age of 60

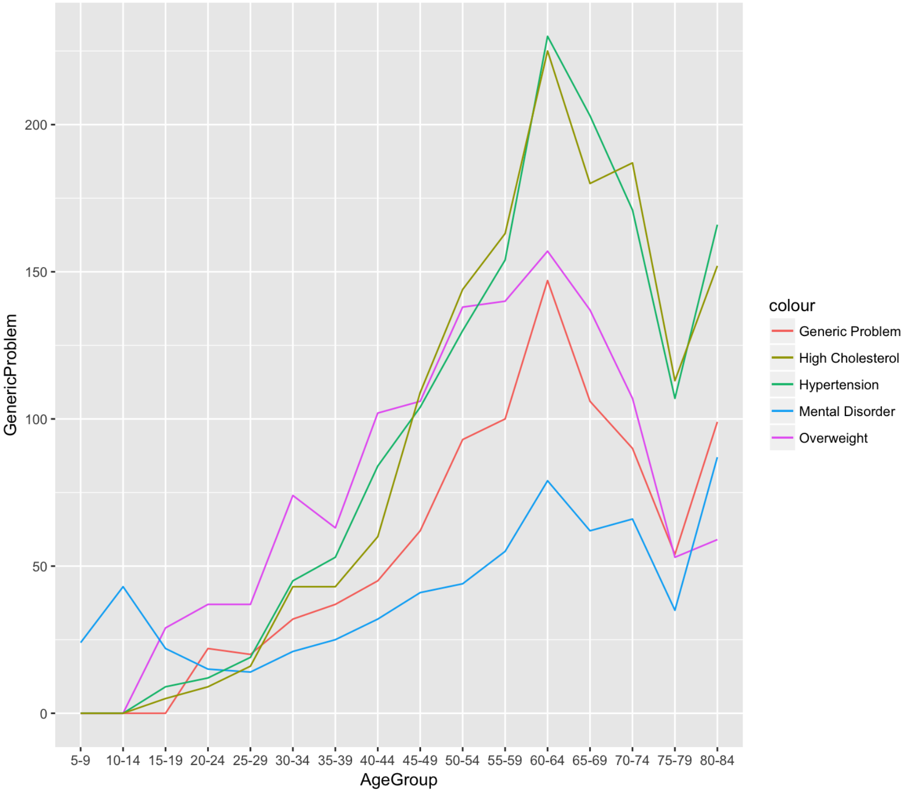


Fig. 7. The frequency of the most common diseases over time

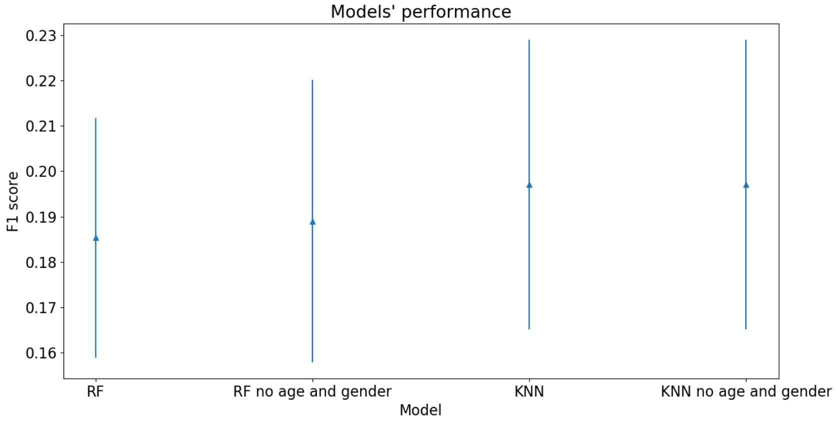


Fig. 8. Initial model performance wasn't very good

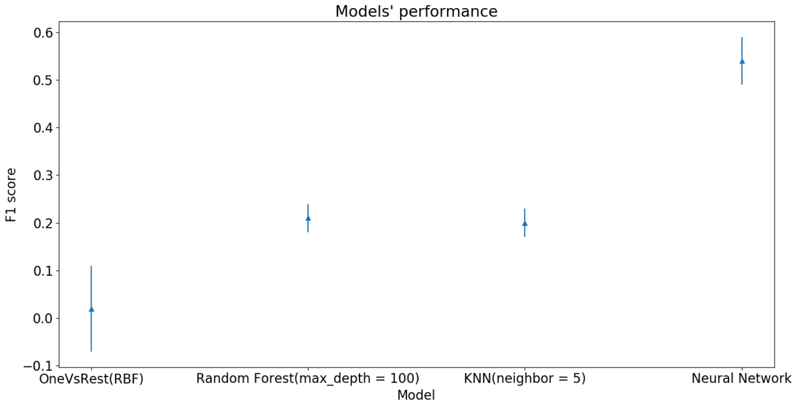


Fig. 9. Neural Network gave the best performance

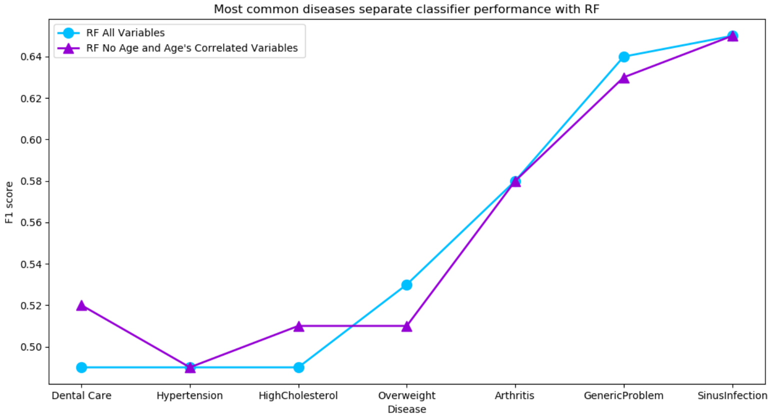


Fig. 10. Fairness Aware Random Forest Performance 1

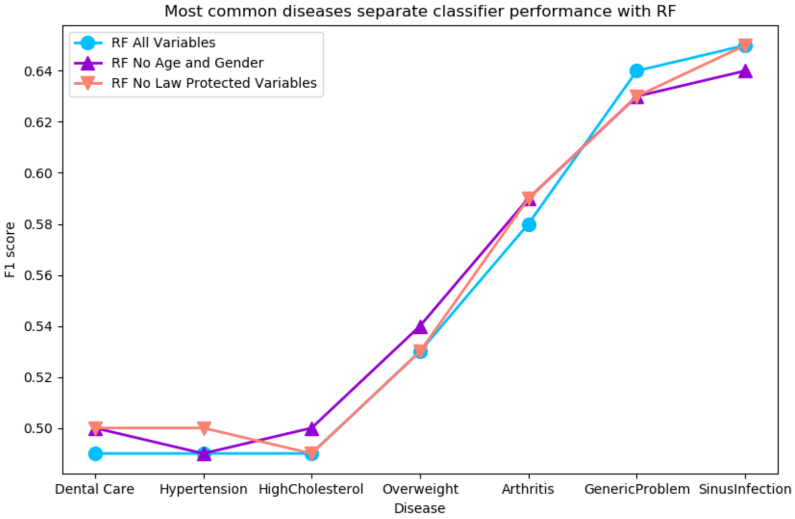


Fig. 11. Fairness Aware Random Forest Performance 2

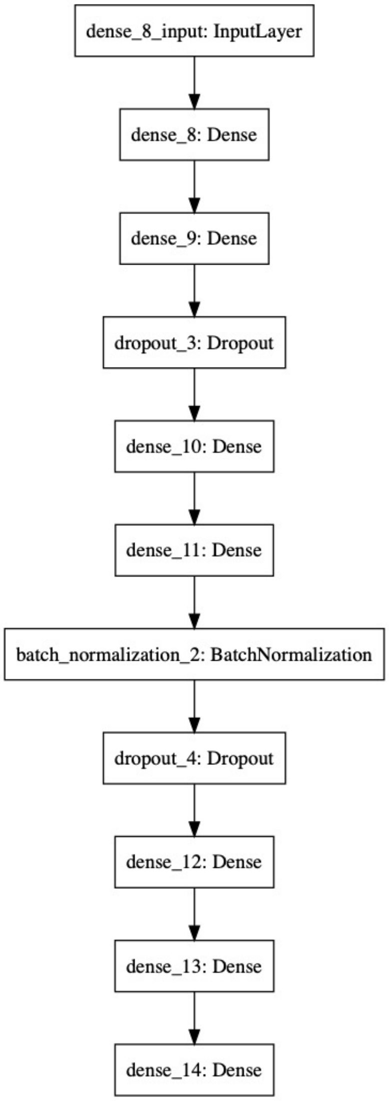


Fig. 12. Neural Network Architecture