Medical Insurance Cost Prediction

Amir Khabbab Ahammed
Department of Computer Science & Engineering
East West University
Dhaka, Bangladesh
2019-2-60-092@std.ewubd.edu

Syada Tasfia Rahman
Department of Computer Science & Engineering
East West University
Dhaka, Bangladesh
2019-2-60-006@std.ewubd.edu

Maisha Mahajabin
Department of Computer Science & Engineering
East West University
Dhaka, Bangladesh
2019-2-60-005@std.ewubd.edu

Shuvo Kumar Das
Department of Computer Science & Engineering
East West University
Dhaka, Bangladesh
2019-1-60-022@std.ewubd.edu

Abstract—**Insurance is a policy that eliminates or decreases loss costs incurred by various risks. Various factors influence the cost of insurance. These considerations contribute to the insurance policy formulation. Machine learning (ML) for the insurance industry sector can make the wording of insurance policies more efficient. This study demonstrates how different models of regression can forecast insurance costs.**

Keywords—machine learning, artificial intelligence, random forest, linear regression, models analysis, health insurance, insurance cost, prediction.

I. Introduction

We live in a world that is filled with dangers and uncertainties. However, because risks cannot always be avoided, the financial sector has devised a number of products to protect individuals and organizations from them by utilizing financial resources to compensate them [1]. Medical insurance is a type of health care coverage that pays for unexpected medical expenses incurred as a result of a disease. These costs could include hospital stays, medication, or doctor consultations. Medical insurance is determined by estimating a population's collective medical bills and then dividing that risk among policy subscribers.Bangladesh has so far maintained about 90,000 life insurance policies out of a total population of over 160 million, and of those, less than 10% of life insurance subscribers have any type of health coverage.

Healthcare cost forecasting is now a valuable tool for improving healthcare accountability. The healthcare sector produces a very large amount of data related to patients, diseases, and diagnosis, but since it has not been analyzed properly, it does not provide the significance that it holds along with the patient's healthcare costs [2]. The cost of health insurance can be determined by a variety of factors such as age, sex, BMI, children, smoker region etc. The goal of this project is to allow a person to get an idea about the necessary amount required according to their own health status. Later they can comply with any health insurance company and their schemes & benefits keeping in mind the predicted amount from our project [3].

In Machine Learning(ML), we work with large amounts of data that must be processed simply in order to be fully understood. One can analyze, classify, or predict future outcomes using that solution. In the insurance sector, ML can help enhance the efficiency of policy wording. In healthcare, ML algorithms are particularly good at predicting high-cost, high-need patient expenditures [4]. We used supervised machine learning models in our project to demonstrate and compare the accuracy of two regression models, Linear Regression (LR) and Random Forest Regressor (RFR). This project focuses on the LR and RFR models, which thoroughly compare regression in estimating total healthcare expenditures. Estimating the health expenses associated with obesity in the population is the primary emphasis of this study.

The component of this paper is structured as follows: The numerous approach methods for evaluating healthcare expenses are discussed in Section 2. Results and discussion are outlined in depth in Section 3. Section 4 brings the study to a close.

II. Methodology

In this project, we used the Python programming language to implement and train a machine learning-based model for predicting medical insurance costs. The dataset in discussion is a compilation of medical expense personal data, which includes depersonalized data on individuals. These data will serve as a tool for technique learning and will produce useful information. Following **Fig-1** [5] shows the project's working process.

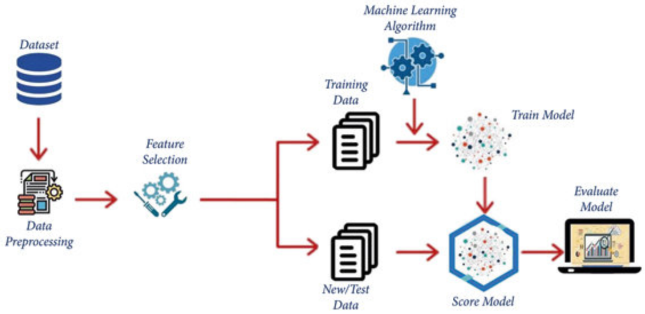
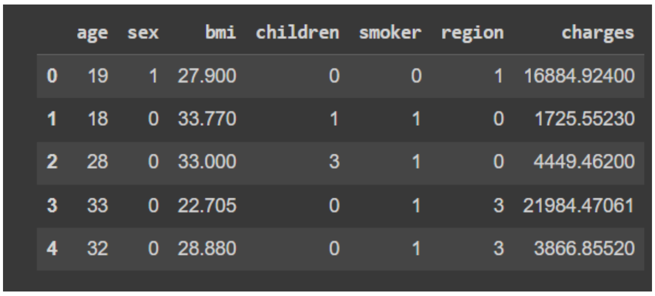


Fig-1: Working methodology

A. Dataset description

We preprocessed the data after obtaining the medical insurance cost dataset from KAGGLE's repository. The dataset contains 1338 rows and 7 columns. The columns present in the dataset are 'age', 'sex', 'BMI', 'children', 'smoker', 'region' and 'charges'. The charges column is the target column and the rest others are independent columns. Independent columns are those which will predict the outcome.



	age	sex	bmi	children	smoker	region	charges
0	19	1	27.900	0	0	1	16884.92400
1	18	0	33.770	1	1	0	1725.55230
2	28	0	33.000	3	1	0	4449.46200
3	33	0	22.705	0	1	3	21984.47061
4	32	0	28.880	0	1	3	3866.85520

Fig-2: Sample dataset

The first column is Age. Age is an important factor for predicting medical expenses because young people are generally more healthy than old ones and the medical expenses for Young People will be quite less as compared to old people.

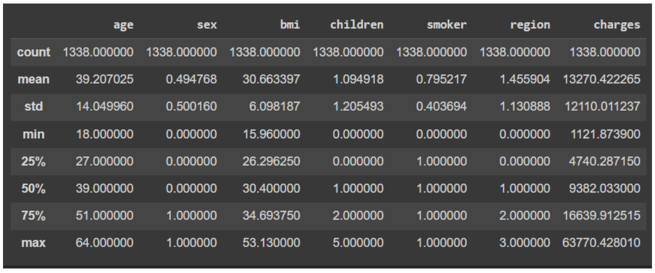
The Next column is sex, which has two Categories in this column: Male and Female. The sex of the person can also play a vital role in predicting the medical expenses of a subject.

After that, you have the ‘BMI’ column, then BMI is Body Mass Index. For most adults, an ideal BMI is in the 18.5 to 24.9 range. For children and young people aged 2 to 18, the BMI calculation takes into account age and gender as well as height and weight. If your BMI is less than 18.5, you are considered underweight. People with very low or very high ‘BMI’ are more likely to require medical assistance, resulting in higher costs.

The fourth column is the ‘children’ column, which contains information on how many children your patients have. Persons who have children are under more pressure because of their children’s education, and other needs than people who do not have children.

The fifth is the ‘smoker’ column. The Smoking factor is also considered to be one of the Most Important factors as the people who smoke are always at risk when their age reaches 50 to 64.

Next is the ‘region’ column. Some Regions are Hygienic, Clean, Neat, and Prosperous, But some Regions are not, and this information affects health which is related to medical expenses.



	age	sex	bmi	children	smoker	region	charges
count	1338.000000	1338.000000	1338.000000	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	0.494768	30.663397	1.094918	0.795217	1.455904	13270.422265
std	14.049960	0.500160	6.098187	1.205493	0.403694	1.130888	12110.011237
min	18.000000	0.000000	15.960000	0.000000	0.000000	0.000000	1121.873900
25%	27.000000	0.000000	26.296250	0.000000	1.000000	0.000000	4740.287150
50%	39.000000	0.000000	30.400000	1.000000	1.000000	1.000000	9382.033000
75%	51.000000	1.000000	34.693750	2.000000	1.000000	2.000000	16639.912515
max	64.000000	1.000000	53.130000	5.000000	1.000000	3.000000	63770.428010

Fig-3: Statistics Measures of Dataset

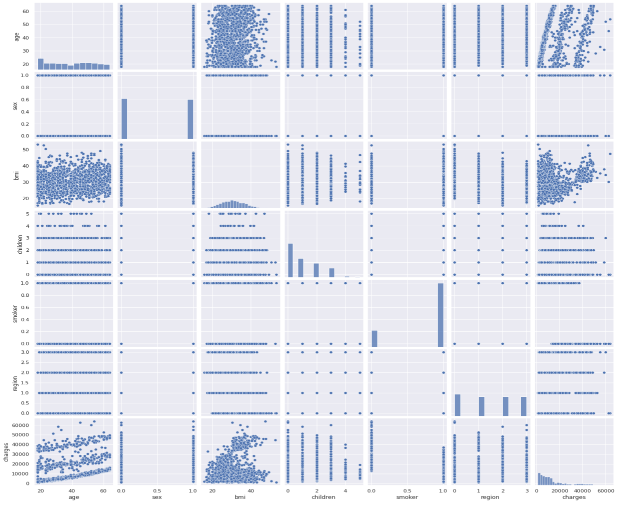


Fig-4: Visualization of dataset

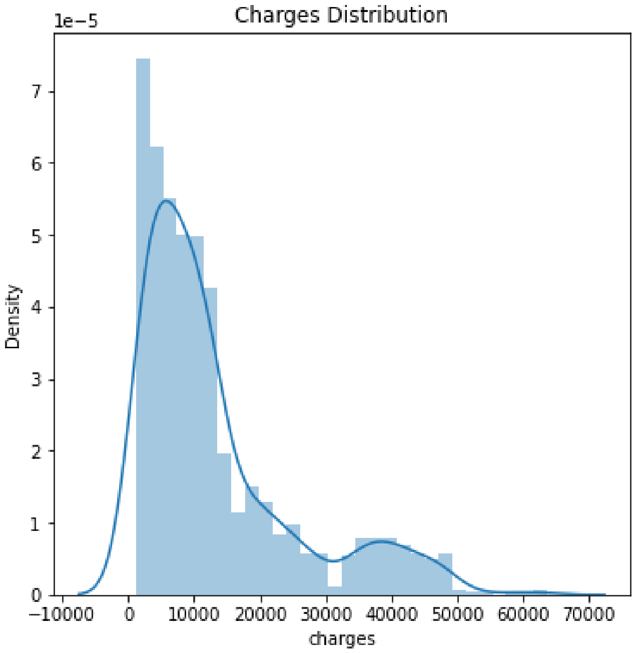


Fig-5: distribution of charges values

B. Data preprocessing

The following steps illustrate the steps of data pre-processing:

➢ Columns sex, region, smoker are converted to categorical variables first and then were converted to numerical variables to be compatible with the model building.

➢ Missing values are removed, and the data is cleaned for analysis and model building.

➢ Some columns are scaled for the model building.

Table 1 The description of the dataset is described and conversion of categorical feature values to numerical values [5]

Seria l No.	Feature Name	Description	Value

1	Age	One of the most important aspects of health care is age	It has an integer value
2	Sex	Gender	(Male = 0, female = 1)
3	BMI	Understanding the human body: weights that are exceptionally high or low in relation to height	An objective body weight index (kg/m^2) based on the height-to-weight ratio, ideally 18.5 − 25
4	Children	Number of children/dependents	It has an integer value
5	Smoker	Smoking state	(Smoker = 1, nonsmoker = 0)
6	Region	Area of residence	(Southeast = 0, southwest = 1, northwest = 2, northeast=3)
7	Charges	Medical costs paid by healthcare insurance	It has an integer value

C. Model Specification

Multi Linear Regression

Linear regression is one of the most common supervisory machine learning statistical analysis techniques [6]. It is frequently used to ascertain linear correlations between two or more responses and predictive variables. This report used multiple linear regression[Fig-1] to look into the connection between the concept of total and other properties in datasets in order to determine the properties most affected by total cost of maintenance. 80% of the data in the dataset was trained, while 20% was tested. **Fig-6** [7] shows the process of linear regression for our proposed model.

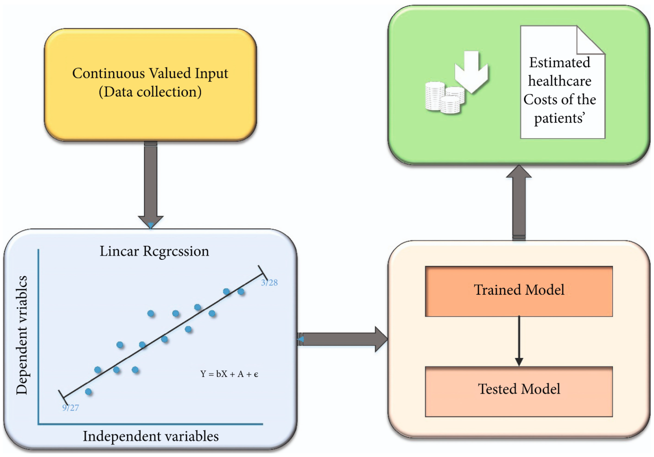


Fig-6: Diagram of Linear Regression for proposed model

In multi linear regression the dependent variable is determined from independent variables and it's intercepts as follows,

$$Y =\beta 0 +\beta 1X1 +\beta 2X2 +...........+\beta kXk+\varepsilon$$

In this equation Y is the dependent variable which is the cost of the health insurance in this project and X1, X2, X3, ……. Xk are the independent variables which are age, sex, bmi, children, smoker, region in this project. β0, β1, β2 are the intercepts.

Ridge Regression

Any data that exhibits multicollinearity can be analyzed using the model tuning technique known as ridge regression. This technique carries out L2 regularization. Predicted values differ much from real values when the problem of multicollinearity arises, least-squares are unbiased, and variances are significant.

Support Vector Machine

A support vector machine (SVM) could be an administered machine that shows employment classification calculations for two-group classification issues. After giving an SVM demonstrate sets of labeled training data for each category, they're able to classify unused content. To be sure, it's most commonly employed in the context of classifying data. According to this method (where n is the number of features you have), we plot each data item as an individual point in space with the value of each feature being a coordinate.

III. Results and Discussion

We see that the accuracy of the predicted amount was seen best i.e. 83% in Ridge Regression and Support Vector Machine. We can choose any of those two. Other model also gave good accuracy.

Table 2 shows the accuracy percentage of models.

Name of Algorithms	Accuracy Score (in Percentage)
Linear Regression	74%

Support Vector Machine(SVM)	83%
Ridge Regression	83%

Table 3 shows the comparison between RMSE,R2_score(test_training),R2_score(test),Cross validation of models

Model	RMSE	R2_score (training)	R2_score (test)	Cross validation
Linear Regression	0.480	0.741	0.783	0.745
Support Vector Machine	0.359	0.857	0.871	0.813
Ridge Regressor	0.465	0.741	0.784	0.826

After processing the data, a model has been made by using multiple linear regression in which basic health details need to be input and you will get the future insurance cost of that individual. As per the dataset we received promising results.

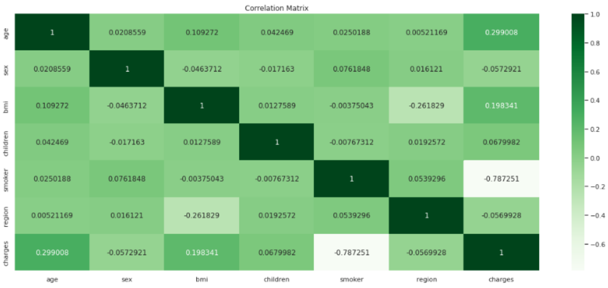


Fig-7: Correlation plot of feature

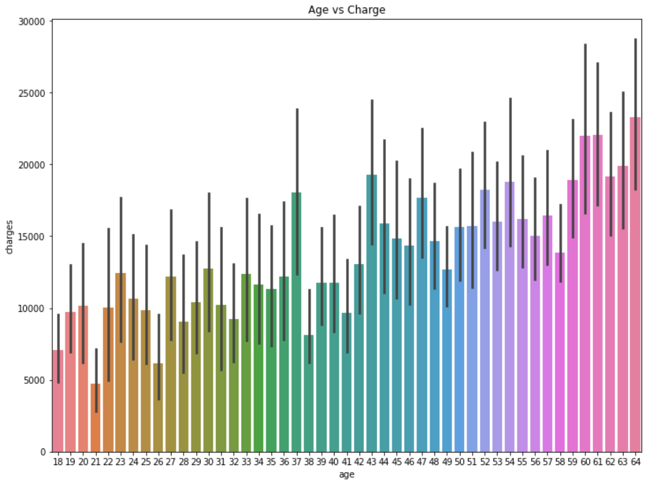


Fig-8(a): age vs charge

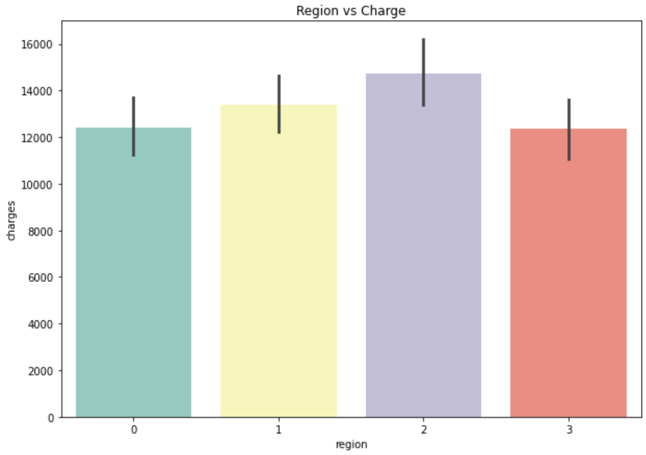


Fig-8(b): region vs charges

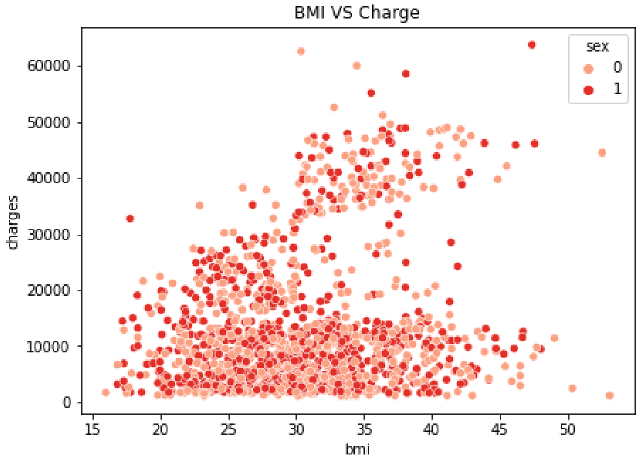


Fig-8(c): BMI vs charges

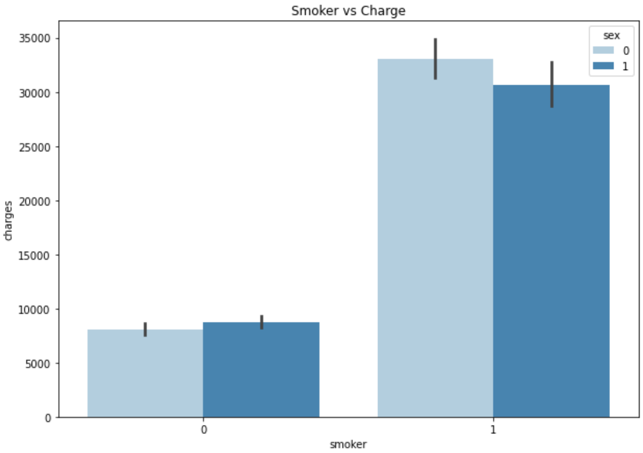


Fig-8(d): smoker vs charges

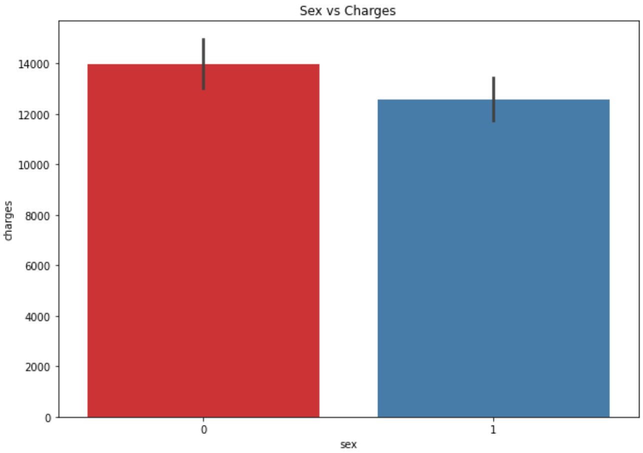


Fig-8(e): sex vs charges

Fig-8(a-e): Distribution of Feature

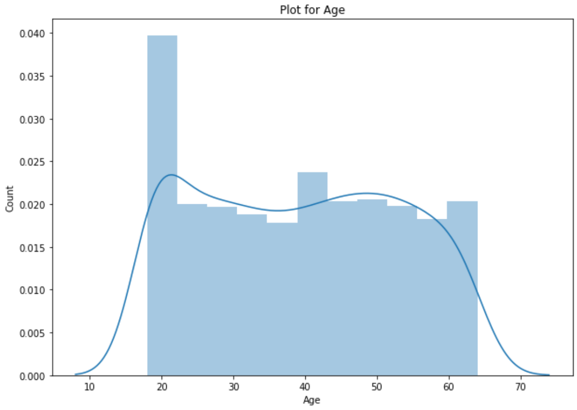


Fig-9(a): plot for age

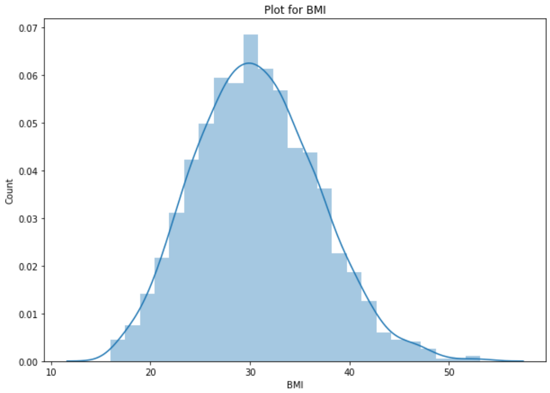


Fig-9(b): plot for BMI

In the future we can implement more models. If there are more variables data available in the research such as income, drinking, marital status, diabetes, profession and past major illness then we will get more accurate results in predicting medical insurance for an individual. With these kinds of data, we can also predict an individual's health for upcoming time.

IV. Conclusion

When it comes to utilizing historical data, machine learning (ML) is one component of computational intelligence that can use a variety of applications and systems to tackle a variety of problems. In the healthcare sector, predicting medical insurance costs is a challenge that has to be looked into and resolved. In this study, computational intelligence is used to forecast healthcare insurance expenditures using a machine learning system. The KAGGLE repository was used to collect the medical insurance dataset, which was then used to train and test the linear regression methods. This dataset underwent preprocessing, feature engineering, data splitting, regression, and evaluation processes before being subjected to regression analysis. The accuracy of our model is 74%. This model is working accurately in the testing period.

Future research will adjust the parameters of machine learning and deep learning approaches using nature-inspired and meta-heuristic algorithms on several medical health-related datasets.

References

[1] Kaushik K, Bhardwaj A, Dwivedi A D, Singh R. Machine Learning-Based Regression Framework to Predict Health Insurance Premiums. Int. J. Environ. Res. Public Health 2022, 19, 7898.

[2] B. D. Sommers, "Health insurance coverage: what comes after the ACA?" Health Affairs, vol. 39, no. 3, pp. 502–508, 2020.

[3] Bhardwaj N.; Anand R. Health Insurance Amount Prediction. International Journal of Engineering Research and Technology (IJERT), 2020.

[4] C. Delcher, E. Shenkman, and S. Ranka, "Machine learning approaches for predicting high cost high need patient expenditures in health care," BioMedical Engineering Online, vol. 17, no. 1, pp. 131–220, 2018.

[5] Hassan U A Ch, Iqbal J, Hussain S, Mosleh M, Ullah S S. A Computational Intelligence Approach for Predicting Medical Insurance Cost. Mathematical Problems in Engineering. 2021. 1-13.

[6] H. N. Alhazmi, A. Alghamdi, F. Alajlani, S. Abuayied, and F. M. Aldosari, "Care cost prediction model for orphanage organizations in Saudi Arabia," *IJCSNS*, vol. 21, no. 4, p. 84, 2021.

[7] Ahmed I. Taloba, Rasha M. Abd El-Aziz, Huda M. Alshanbari, Abdal-Aziz H. El-Bagoury, "Estimation and Prediction of Hospitalization and Medical Care Costs Using Regression in Machine Learning", *Journal of Healthcare Engineering*, vol. 2022, Article ID 7969220, 10 pages, 2022.

[8] Tike, Anuja. A MEDICAL PRICE PREDICTION SYSTEM. 2018. Master's Projects. 619.

[9] Bertsimas, Dimitris & Bjarnadóttir, Margrét & Kane, Michael & Kryder, J. & Pandey, Rudra & Vempala, Santosh & Wang, Grant. (2007). Algorithmic Prediction of Health Care Costs and Discovery of Medical Knowledge. Operations Research. 56.