Babatunde Atolagbe

CIEG 642 – Advanced Data Analysis Term Paper

Identifying Key Factors for Predicting prevalence of heart diseases in the United States.

1. Introduction

The aim of this project is to demonstrate the application of variable subset selection for improving model performance. Specifically, the suitability of machine learning technique like Lasso and LassoCV regularization for simplifying models is explored. Principal component analysis and multiple linear regression were also conducted on the dataset. The dataset used is the heart disease dataset from the center for diseases control website [1] and the problem of interest in this project is to identify the most important variables for developing a simple, yet very good, model for predicting the prevalence of heart diseases in the United States. This dataset has been used for different applications in literature. The dataset contains segment of health The problem is that there is that has attempted to integrate all the components of the data for research purpose. It is worth mentioning that this project is defined following a Kaggle challenge [2] requiring the development of a model for predicting the likelihood of a patient to have heart disease. However, the difference is that in this project, the likelihood of patients to have heart disease is not explored, rather predicting (i.e., estimating) the prevalence of heart diseases in the cities of the USA is considered. This can be very helpful for city stakeholders to make decisions on how to improve on some influencing factors of heart diseases, where possible.

2. Dataset & Data Preparation

The dataset was downloaded from the CDC website as ".csv" files with each file representing one variable. For each file, there are 3226 rows representing all the counties in the USA and there are 4 columns: county code, county name, state and the variable that is represented by that .csv file. 38 such files were downloaded, and a python script was written to combine them all while filtering multiple instances of county code, county name, and state to avoid repetitions. The resulting dataset is one that combines all the variables, having 40 columns and 3226 rows.

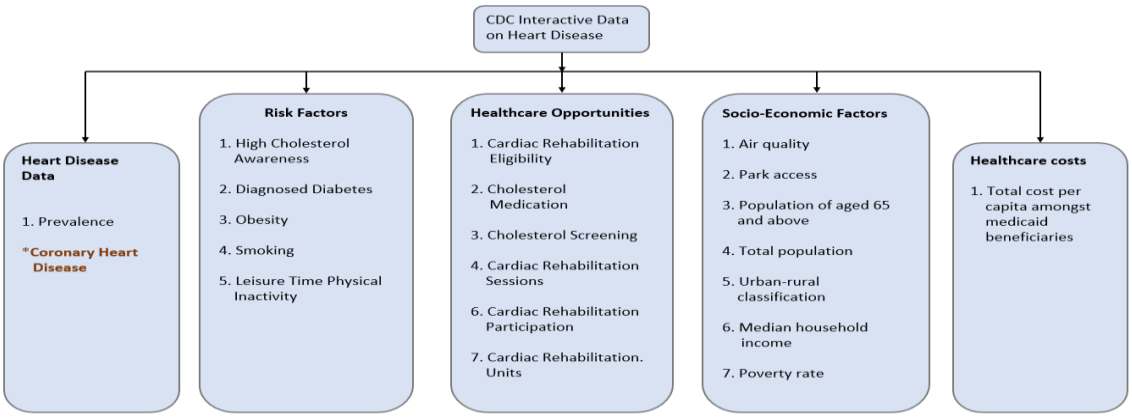

Figure 1. Categories of variables in the dataset.

The variables in the dataset can be broadly categorized into five classes as shown in Figure 1. Data wrangling exercises were conducted. These include renaming columns with shorter names, ensuring consistency in the format of each variable, dealing with missing data and dropping variables with more than 50 percent missing data. After wrangling, the dataset had 2310 rows and 36 columns.

3. Exploratory Data Analysis

With the cleaned dataset, exploratory data analysis was conducted in two phases and was for risk factors, healthcare opportunities and socio-economic factors separately. The idea was to determine the most important variables in each category and ensure that each category is represented in the predictive model. In the first step, correlations between all the variables were computed and their statistical significance were computed by comparing their p values with α significant value of 0.05. Where the p values are greater than α, the correlation values computed were considered as insignificant. Figures 2 and 3 present the correlation plots of the variables and the correlation values and their statistical significance, which indicates whether the variables are truly correlated or not.

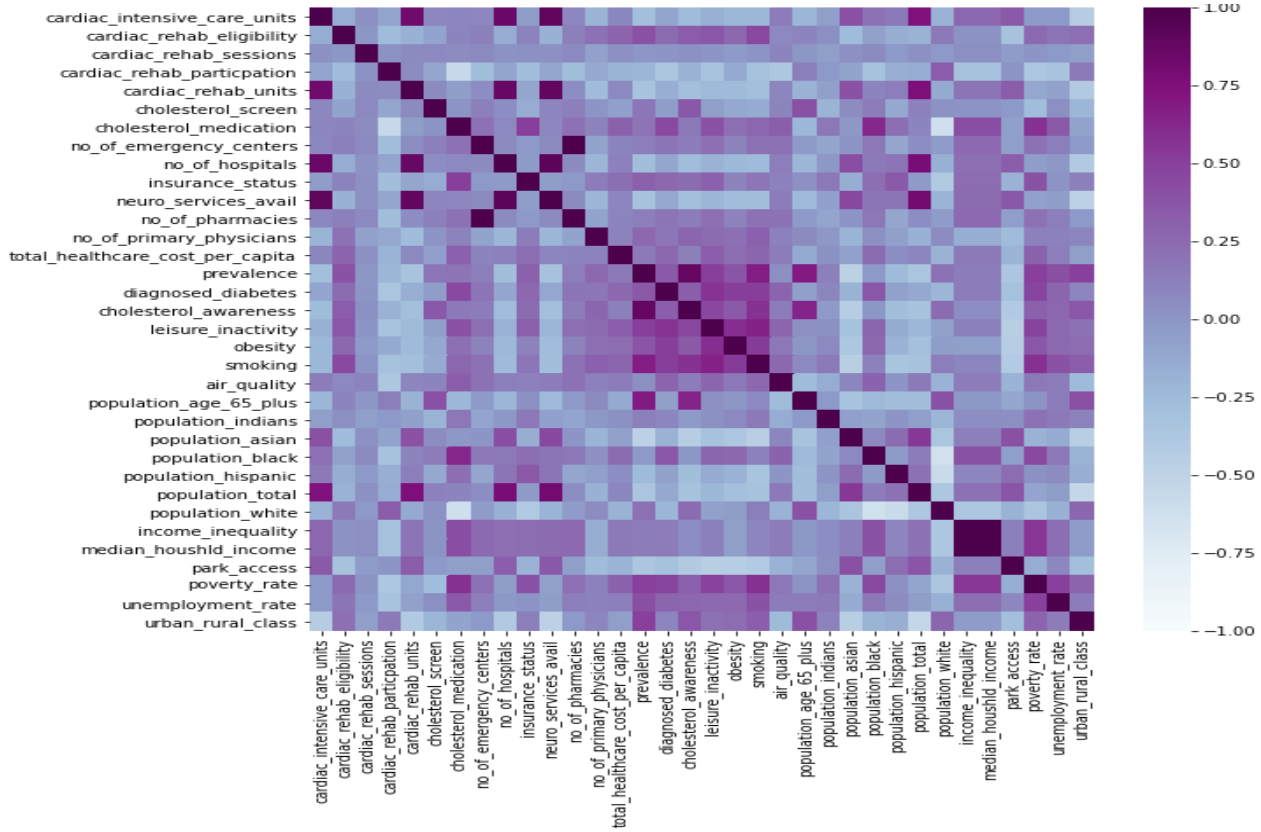

Figure 2. Correlation plots of predictor variables.

	prevalence	p_value	significance
population_asian	-0.482293	0.0	True
park_access	-0.358428	0.0	True
population_total	-0.347519	0.0	True
cardiac_rehab_units	-0.321223	0.0	True
neuro_services_avail	-0.317675	0.0	True
cardiac_intensive_care_units	-0.278313	0.0	True
no_of_hospitals	-0.273078	0.0	True
population_hispanic	-0.246380	0.0	True
cardiac_rehab_particpation	-0.206498	0.0	True
population_black	-0.021265	0.0	True
cardiac_rehab_sessions	-0.004531	0.0	True
air_quality	0.041882	0.0	True
population_indians	0.127955	0.0	True
no_of_emergency_centers	0.139953	0.0	True
no_of_pharmacies	0.139953	0.0	True
median_houshld_income	0.153208	0.0	True
income_inequality	0.153208	0.0	True
total_healthcare_cost_per_capita	0.160032	0.0	True
cholesterol_screen	0.185265	0.0	True
cholesterol_medication	0.199996	0.0	True
population_white	0.211037	0.0	True
no_of_primary_physicians	0.265981	0.0	True
insurance_status	0.311594	0.0	True
diagnosed_diabetes	0.357811	0.0	True
obesity	0.377841	0.0	True
cardiac_rehab_eligibility	0.403965	0.0	True
unemployment_rate	0.414303	0.0	True
poverty_rate	0.500473	0.0	True
urban_rural_class	0.501646	0.0	True
leisure_inactivity	0.512535	0.0	True
smoking	0.676678	0.0	True
population_age_65_plus	0.689180	0.0	True
cholesterol_awareness	0.877170	0.0	True

Figure 3. Statistical significance of correlation values of predictors with prevalence of heart diseases

In Figure 3, one can see that “park_access” is negatively correlated with “prevalence” thus suggesting that when there are more relaxation centers like parks in a city for people to relief stress, there is a high tendency for heart disease occurrence to be reduced. Conversely, “smoking” and “cholesterol_awareness” are positively correlated thus suggesting that these variables influence the occurrence of heart diseases. Other variables in Figure 3 can be interpreted as such. On population groups listed in the table, there is a positive correlation between “population_age_65_plus” and “prevalence” as one would have expected that older people should be more susceptible to having heart disease. However, surprisingly, the

relationship between “population_total” and “prevalence” is negative, suggesting that the more the population, the lower the prevalent of heart disease. Both variables capture contrasting effect of population (demographics). As such, we will let our model reflect both variables. All other population groups are not considered in the model for simplicity.

The “urban_rural_class” variable was further explored (results in Jupyter notebook), and it was observed that there is a tendency for heart diseases to be prevalent in nonmetro areas, while the reverse is the case for metro areas. Perhaps, this may be due to decreased activities in nonmetro area which can translate into more leisure inactivity, and vice-versa for metro areas. However, investigation of the relationship between nonmetro, metro areas and leisure_inactivity, as shown below, indicates that this can be true. Similar observations can be made of park_access except that the behavior is opposite to leisure_inactivity as one would have expected.

Finally, the relationship between "income_inequality", "median_houshld_income", "poverty_ rate", "unemployment_rate" and “prevalence” was studied since these variables all seem to be indicators for economic status. Both “income_inequality” and “median_houshld_income” were found to be 100% correlated with each other and that means either of them can be representative of the other. “unemployment_rate” and “poverty_rate” is fairly correlated, but it is assumed that “poverty_rate” is a better variable considering that it is more correlated with “prevalence”. Of all these four variables, “poverty_rate” correlates best with prevalence and as such is a variable of interest.

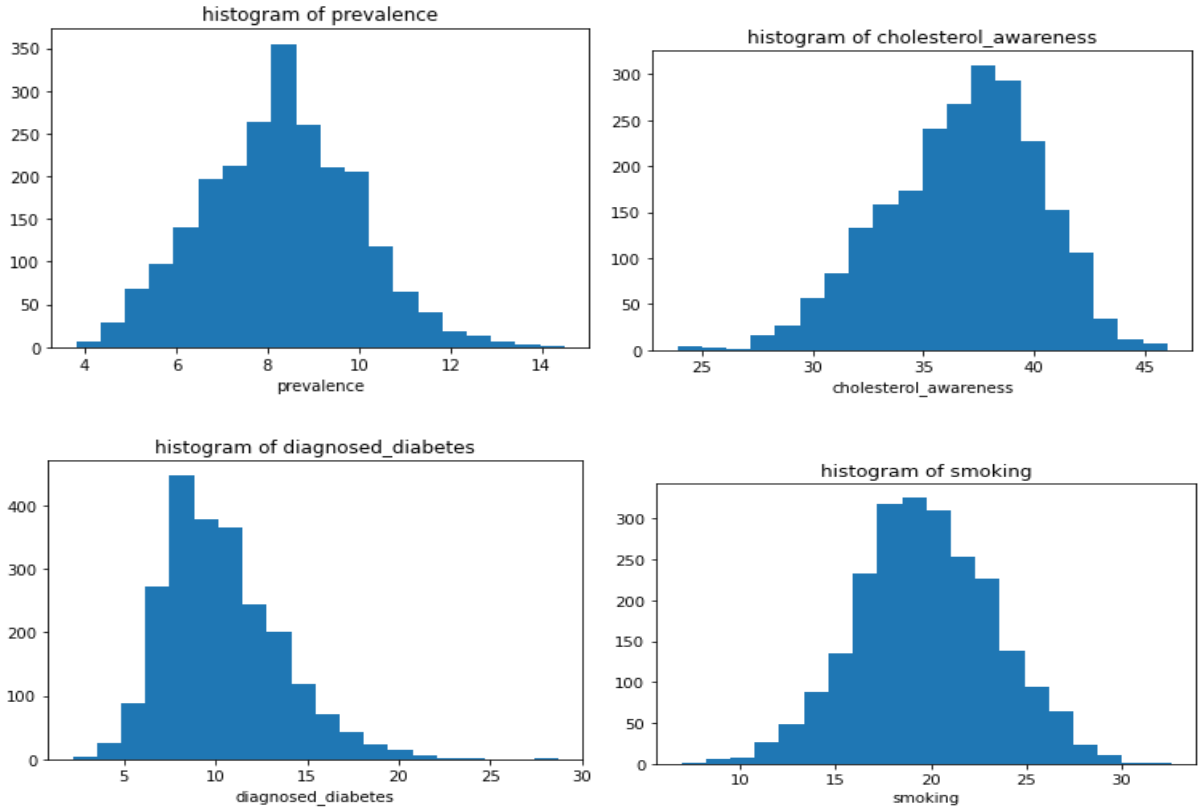

Figure 4. Histograms of some variables

At the end of the first phase of the exploratory data analysis, 18 variables were considered as variables of interest for modeling purpose. Moving forward, the second phase of exploratory data analysis was conducted. Essentially, in this phase, univariate distributions of the selected 18 variables and their visualizations were studied for any outliers and skewness in the distribution. For most of the variables, the distributions appear normal with no outlier detected. This is likely because most of the variables had been normalized from the source. Figure 4 presents the histograms of some of the variables. Figure 5 is another correlation plot for the selected 18 variables.

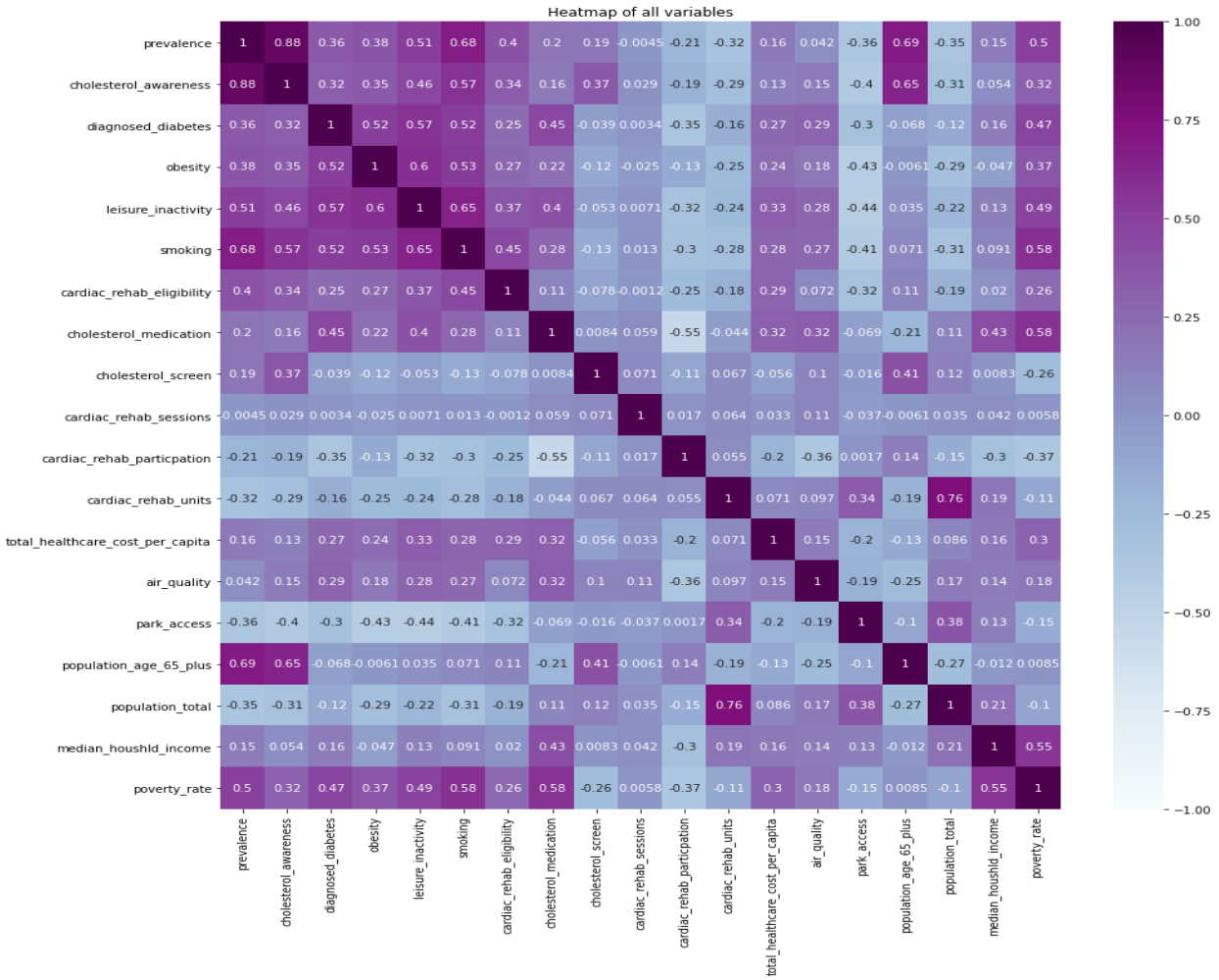

Figure 4. Correlation plots of selected 18 variables of interest.

4. Modeling

Principal Component Regression

Principal component analysis (PCA) is a method used to reduce a large data set of variables to a smaller set of variables, called principal components, through a linear combination of the original variables. The principal components capture unique proportion of the variance in the dataset. This way, the subsequent modeling, such as prediction or classification, is simplified as

there are now fewer variables to build the model with. In this project, PCA was conducted on the original dataset and the proportion of explained variance captured by the principal components is presented in Figure 5a. Figure 5b presents the cumulative explained variance captured by all the principal components.

	percent explained variance
PC1	27.303296
PC2	15.731304
PC3	9.326657
PC4	7.733268
PC5	5.811532
PC6	5.344981
PC7	4.668847
PC8	3.960581
PC9	3.436981
PC10	2.734057
PC11	2.321666
PC12	2.226600
PC13	1.975374
PC14	1.806317
PC15	1.747824
PC16	1.507503
PC17	0.964908
PC18	0.773548
PC19	0.624757

Figure 5a. Percent explained variance by principal components

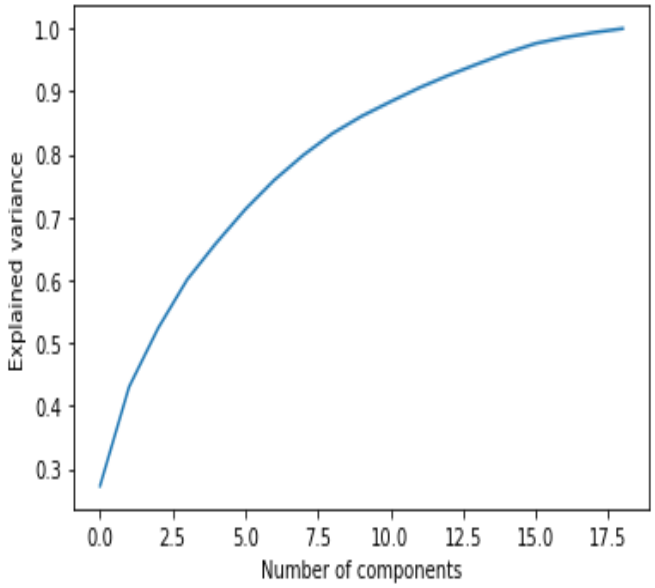

Figure 5b. Cumulative percent explained number of principal components

In Figure 5a, it is observed that the first principal component captures only 27% of the variance in the dataset, second principal component captures 15%, third principal component captures 9% and so on and so forth. In Figure 5b, it is observed that the first 12 principal components capture 90% of the variance. As such, from this analysis, it may suffice to use only the first 12 principal components for developing a multiple linear regression model.

To build a principal components regression model with the 12 principal components, the coefficients of the principal components are mapped to both the training and test sets and the regression model is applied on this transformed data. The resulting model yielded an accuracy score of 88.89% and the mean squared error on the testing set is 0.501%.

Multiple Linear Regression

In addition to the PCA regression model, three different multiple linear regression (MLR) models were developed in this study. These are named MLR_SIMPLE, MLR_LASSO, and MLR_LASSOCV. These models were selected because the goal of this project is to identify the best predictors that produces the best model. Before building the MLR_SIMPLE, an exploratory

regression analysis was first conducted. This procedure is called exploratory regression analysis because it was conducted to identify the predictors that can potentially improve the model rather than it being a predictive model. This was conducted using the “statsmodels” package in Python. Since this model was not intended to predict, the entire data was used as opposed to the case where data is split into train and test sets for predictive models. Figure 6 summarizes the attributes for this exploratory model.

OLS Regression Results

Dep. Variable:	prevalence	R-squared:	0.937
Model:	OLS	Adj. R-squared:	0.937
Method:	Least Squares	F-statistic:	1794.
Date:	Fri, 04 Mar 2022	Prob (F-statistic):	0.00
Time:	23:49:08	Log-Likelihood:	-1250.3
No. Observations:	2310	AIC:	2541.
Df Residuals:	2290	BIC:	2656.
Df Model:	19		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-3.5240	0.354	-9.950	0.000	-4.219	-2.829
cholesterol_awareness	0.1471	0.005	27.034	0.000	0.136	0.158
diagnosed_diabetes	0.0002	0.004	0.050	0.960	-0.007	0.007
obesity	0.0033	0.002	1.466	0.143	-0.001	0.008
leisure_inactivity	0.0076	0.002	3.070	0.002	0.003	0.012
smoking	0.1496	0.004	33.700	0.000	0.141	0.158
cardiac_rehab_eligibility	0.0119	0.002	5.403	0.000	0.008	0.016
cholesterol_medication	0.0373	0.005	7.902	0.000	0.028	0.047
cholesterol_screen	-0.0235	0.004	-5.672	0.000	-0.032	-0.015
cardiac_rehab_sessions	-0.0039	0.002	-2.131	0.033	-0.008	-0.000
cardiac_rehab_particpation	-0.0027	0.001	-3.735	0.000	-0.004	-0.001
cardiac_rehab_units	-0.0379	0.010	-3.929	0.000	-0.057	-0.019
total_healthcare_cost_per_capita	5.903e-06	4.51e-06	1.308	0.191	-2.95e-06	1.48e-05
air_quality	-0.0239	0.007	-3.491	0.000	-0.037	-0.010
park_access	-0.0004	0.001	-0.745	0.456	-0.001	0.001
population_age_65_plus	0.1816	0.004	49.022	0.000	0.174	0.189
population_total	2.668e-07	4.84e-08	5.513	0.000	1.72e-07	3.62e-07
urban_rural_class	0.0923	0.015	6.083	0.000	0.063	0.122
median_houshld_income	0.8267	0.364	2.269	0.023	0.112	1.541
poverty_rate	0.0357	0.003	10.347	0.000	0.029	0.042

Figure 6. Summary of exploratory regression analysis.

In Figure 6, we see that the R^2 value is 0.937 suggesting that 93.7% of the variations in the target variable, "prevalence", can be explained by the variations in the predictor variables. Furthermore, the fact that the F-statistic is far greater than the p-value (0.00) suggests that we cannot but conclude that there exists some relationship between prevalence and the predictor variables. Here the R^2 value is not considered as a predicting power of the model since the model is only intended to be used for exploring the predictors and select the best ones. As such this value will not be considered for comparing the performance of the MLR models.

By inspecting the p-values of the coefficients of the predictors, it is observed that "diagnosed_ diabetes", "obesity", "total_healthcare_cost_per_capita" and "park_access" with p-values 0.960, 0.143, 0.191 and 0.456 respectively, all have larger p-values than α, considering an α significant

level of 0.05 thus suggesting that these coefficients are statistically insignificant and may not have any impact in the model.

Based on these findings, the four variables were penalized (i.e., dropped) and a predictive MLR model was built with the remaining variables. Building a predictive MLR involves, first splitting the actual data to training and testing sets. Then, the training set is used to train a linear regression model while the testing set is used to check how well the model generalizes to new datasets. This model, denoted as MLR_SIMPLE, yielded an accuracy score of 91.73% and the root mean squared error on the testing set is 0.439%.

Next, a Lasso regression model was developed. This is just another multiple linear regression but one that has lasso regularization term embedded. The lasso regularization is a shrinkage method that forces parameters in the model to be penalized. It involves application of the L_2 norm (called lasso penalty) as opposed to the L_1 norm which is used in ridge regression [3]. Since the goal of this project is to identify best predictor variables, Lasso regression seems to be a perfect technique to explore. As such, a model based on Lasso called MLR_LASSO was developed. This model, i.e., MLR_LASSO, yielded an accuracy score of 91.99% and the root mean squared error on the testing set is 0.432%.

An extension to the Lasso regression is the LassoCV regression which has a cross-validation step embedded in its implementation. As such, another model based on LassoCV called MLR_ LASSOCV was also built with 5-fold cross validation step. This model, i.e., MLR_LASSOCV, yielded an accuracy score of 91.93% and the root mean squared error on the testing set is 0.434%.

5. Results and Discussion

The discussion of the results focuses on two aspects namely: factor identification and model performance. On identification of key factors influencing the prevalence of heart diseases in the United States, the parameters selected by the three MLR models are compared and presented in Figure 7. In Figure 7, it is observed that four variables have been penalized in the simple exploratory regression analysis. These include obesity, healthcare_cost, diagnosed_diabetes, and park_access. While one may have been motivated, apriori, that these variables, especially diabetes and obesity, would likely be more important for building a predictive model on heart diseases, the penalization observed here suggests that it is possible that some other variables, which are correlated with these four variables such as cholesterol are more correlated with prevalence and thus are accounting for more variations in the model that the effects of these four variables become insignificant.

Similarly, it is observed in Figure 7 that 12 variables were penalized in the Lasso model while three were penalized in the LassoCV model. It is observed however that diagnosed_diabetes and park_access are penalized in all three models, while obesity, total healthcare cost and urban_rural_class are penalized in more than one model suggesting that they can be completely

disregarded for model building in this study. Other techniques may be required to justify the removal of the remaining seven other variables penalized in the Lasso model. However, for the scope of this study, the simplest model is the lasso model which has fewer non-penalized variables and the variables selected by the model are considered as the best for predicting the prevalence of heart diseases in the United States. These are 'cholesterol_awareness', “leisure_ inactivity”, “smoking”, “cardiac_rehab_eligibility”, “cholesterol_medication”, “cardiac_rehab _particpation”, “population_age_65_plus”, “poverty_rate”, and “urban_rural_class_nonmetro”.

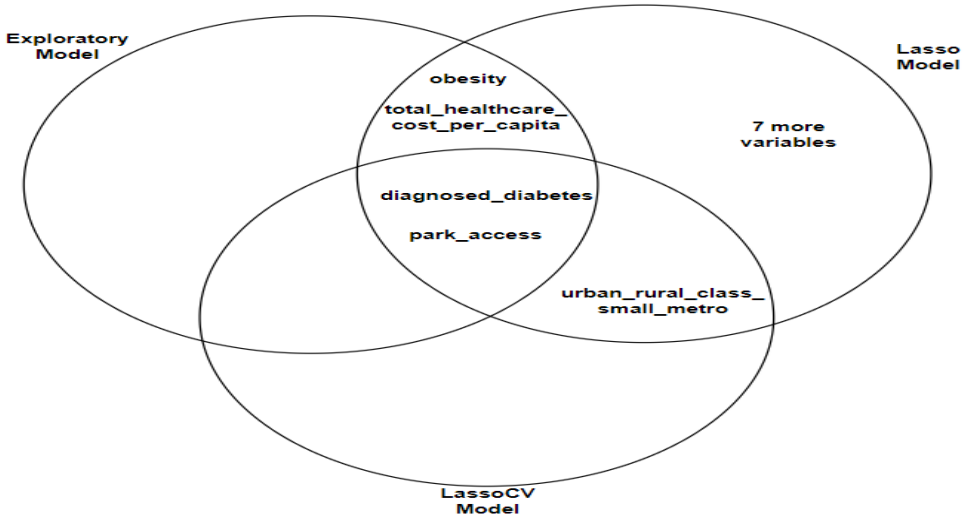

Figure 7. Summary of penalized variables in the simple MLR and Lasso MLR models.

The results of the regression models trained in this project are summarized in Table 1. R^2 and Root Mean Squared Error (RMSE) are the model evaluation metrics. Although, apart from the MLR_PCA that performed worst, the performance of the MLR models were close, but the MLR_LASSO model performs best with R^2 value of 91.99% and root mean squared error of 0.432%. This means that apart from the fact that the lasso model is simple as it uses fewer variables, it also predicted the prevalence by the least error margin of just 0.432% while explaining 91.99% variance between the prevalence of heart diseases and the selected variables. This also supports the popular opinion that simple models are preferrable because they often tend to produce the best performance.

Table 1. Summary of model performance evaluations

Model	**R^2 (%)**	**RMSE (%)**
MLR_PCA	88.89	0.501
MLR_SIMPLE	91.73	0.439
MLR_LASSO	91.99	0.432
MLR_LASSOCV	91.93	0.434

6. Conclusion

In this project, the causal relationship between the prevalence of heart diseases in the United States and the risk factors, healthcare opportunities and socio-economic factors has been explored using the heart diseases dataset from the center for diseases control. Although many works have used this data to answer various questions, few or none have considered integrating these different categories, i.e., risk factors, healthcare opportunities and socio-economic factors in their analyses. Focusing on the selection of best factors for predicting the prevalence of heart diseases as it is done in this project can be beneficial to healthcare stakeholders in every city by helping them identify the most crucial influencing factors that should be focused on. In this study, models based on principal components regression, ordinary linear regression, Lasso regression and LassoCV regression were developed. It was found that the Lasso regression model performed best in predicting the prevalence of heart disease with high R^2 and least error margin. Also, the results from this analysis show that, of all the 40 variables considered in this study, only 12 are important to produce the best predictive model. The study suggests, for example, that amongst other health risk factors, cholesterol, rather than diabetes and obesity, is the most influencing factor of prevalence of heart diseases and that both diabetes and obesity may not be considered for building the best model thus highlighting the danger of holding perceptions about variables before exploring data and underscoring the importance of exploratory analysis of the data before models are built. It is important to mention, however, that these findings are based on the data explored in this project and that different findings may be obtained if another dataset is used.

References

1. Center for Diseases Control (n.d.). Interactive Atlas of Heart Disease and Stroke. Available: https://nccd.cdc.gov/DHDSPAtlas/Reports.aspx. Assessed 05/24/2022.
2. Kaggle Datasets (n.d.). Heart Disease Prediction Dataset. Available: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction. Assessed 05/24/2022.
3. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.