CAN LIFE STYLE CHANGE PREVENT DIABETES?

Mounica Mandapati, Oklahoma State University, Stillwater, Oklahoma

ABSTRACT

Food defines our life. Healthy food enhances human life and enables people to avoid some preventable chronic diseases. Diabetes being one among them. According to CDC, diabetes rates have exponentially increased by a staggering 650% during the past 50 years. Diabetes is usually blamed on people's genetics but not on their food environment and lifestyle. So, the main objective of this study is to find out the important food environment factors and lifestyle factors responsible for causing diabetes and predict the risks of the increase in diabetes population.

For the purpose of this study, the data has been collected from the United States Department of Agriculture. At present, the data set contains 275 variables and 3,144 observations. To explain, predict and visualize the data SAS Enterprise Miner will be used. According to the results, there is an increase of diabetes rates in 85% counties with a maximum increase of 81% of the population from Roosevelt County, New Mexico. Along with known predictors like obesity, race and age, it is observed that counties with more SNAP Participants are more prone to diabetes. Also, taxes on unhealthy foods is helping reduce diabetes rate. Increase in direct farm sales and physical fitness are aiding in decrease of diabetes rate.

1. INTRODUCTION

Our Health is majorly influenced by the food choices we make. Current trends of food choices like overconsumption of calories, added sugars; under consumption of fruits, vegetables; and health issues such as obesity, high blood pressure and high cholesterol are rising rapidly posing an enormous socioeconomic and health challenges. One among such health issues is diabetes, which occurs due to high blood glucose. Diabetes is one of the major non-communicable and fastest growing public health problems in the world; it is a condition difficult to treat and expensive to manage. It has been estimated that the number of diabetes sufferers in the world will double from the current value of about 190 million to 325 million people during the next 25 years. So, immediate corrective actions are required to wipe out diabetes.

It is frequently seen that there is a relationship between food-environment factors like store/restaurant/farms proximity, food prices/taxes, food and nutrition assistance programs, and community characteristics – and diet quality. This project aimed to study the relationship between diet quality and diabetes and to understand the role of different lifestyle choices that contribute in reducing the diabetes rate significantly.

2. DATA BACKGROUND

The data was obtained from the United States Department of Agriculture Economic Research Service. 9 different excel sheets are used from USDA official website which contain data about Stores, Restaurants, Health, Socioeconomic data for each county in the US. These excels are imported into SAS Enterprise Miner for preliminary analysis. The data contained percentage counts of population in each county in 2009 and 2013 for each of the factors.

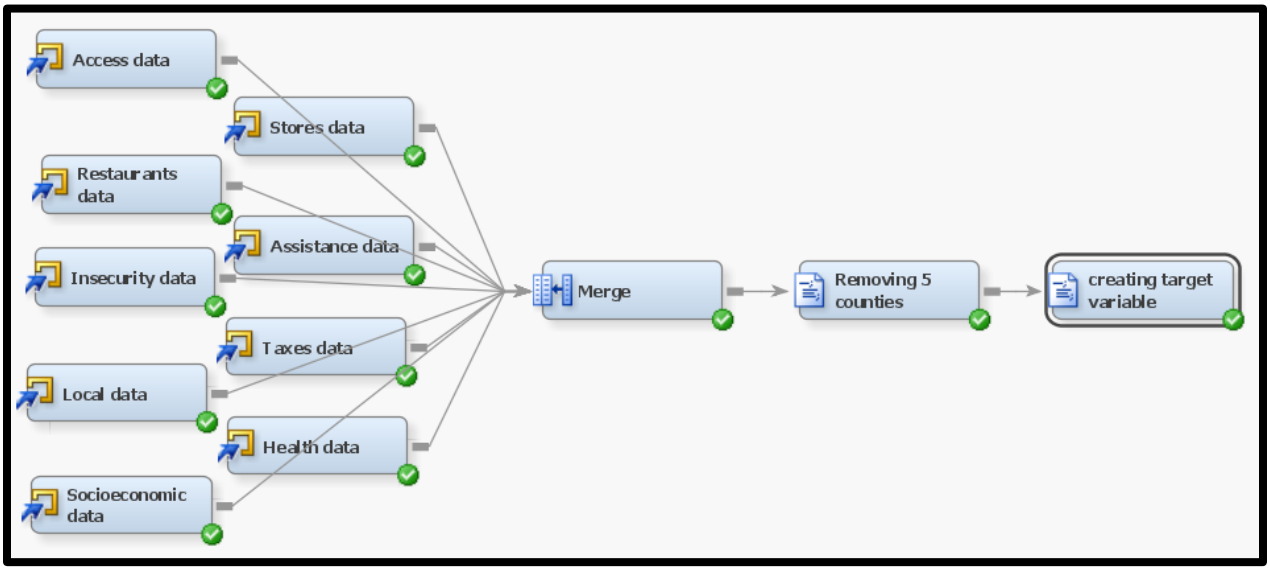

Figure 1: SAS EM Process Flow Diagram for Data Preparation

3. DATA PREPARATION

To understand the common characteristics at county levels these, excel sheets are merged based on the FIPS county code which is Federal Information Processing Standard Code which uniquely identifies counties and county equivalents in the United States. On observing the diabetes change percentage counts, there are 5 counties which had missing values. Since, this is not a systematic error, these counties are deleted from the dataset.

Also, to understand the trends from 2009 to 2013 and how the diabetes percentage changed, which is our target variable, the formula used to calculate the variable is:

$$\text{Diabetes\_Change} = \frac{(\text{PCT\_DIABETES\_ADULTS13}) - (\text{PCT\_DIABETES\_ADULTS08})}{(\text{PCT\_DIABETES\_ADULTS08})} \times 100\%$$

This variable gives the percentage change in diabetes from 2009 to 2013 for each county.

Few other variables calculated for further analysis are:

a) Age_Classification (which says if a county has elder population or younger population)

b) Program Recipients (which says the total percent of the population in a county participating in the National School Lunch Program and Summer Food Service Program)

Thus, a cleaned dataset is prepared for analysis.

4. DATA DICTIONARY

This dataset has 275 variables, which shows the county level information. The following are important variables among them:

Variable	Data Type	Description
FIPS	ID	Federal Information Processing Standard Code for unique identification of counties
PCT_Diabetes_Adults13	Interval	% Population diabetes population, 2013
PCT_Diabetes_Adults08	Interval	% Population diabetes population, 2008
PCT_NHASIAN10	Interval	% Population Asians, 2010
PCT_NHNA10	Interval	% Population American Indians, 2010
PCT_NHWHITE10	Interval	% Population Whites in 2010
PCT_NHBLACK10	Interval	% Population African Americans,2010
PCT_65OLDER10	Interval	% Population 65 years or older, 2010
PCT_18YOUNGER10	Interval	% Population under age 18, 2010
MILK_PRICE10	Interval	Price of low-fat milk/national average, 2010
SODA_PRICE10	Interval	Price of sodas/national average, 2010
MILK_SODA_PRICE10	Interval	Price of low-fat milk/price of sodas, 2010
SODATAX_STORES14	Interval	Soda sales tax, retail stores, 2014
SODATAX_VENDM14	Interval	Soda sales tax, vending, 2014
CHIPSTAX_STORES14	Interval	Chip & pretzel sales tax, retail stores, 2014
CHIPSTAX_VENDM14	Interval	Chip & pretzel sales tax, vending, 2014
FOOD_TAX14	Interval	General food sales tax, retail stores, 2014
PCH_RECFAC_09_14	Interval	Recreation & fitness facilities (% change), 2009-14
PCH_NSLP_09_15	Interval	National School Lunch Program participants (change % pop), 2009-15
PCH_SFSP_09_15	Interval	Summer Food Program participants (change % pop), 2009-15
SNAP_PART_RATE13	Interval	SNAP participants (% eligible pop), 2013
PC_DIRSALES12	Interval	Direct farm sales per capita, 2012
PC_FFRSALES12	Interval	Expenditures per capita, fast food, 2012

5. DATA INSIGHTS:

The cleaned dataset having diabetes percentage change and several other food environment factors are examined. Here are the descriptive analysis results:

i. Before looking into the factors that are responsible for diabetes change, it is important to understand if there is significant change in diabetes change from 2009 to 2013. On conducting a paired t-test for the percentage of diabetes population from 2009 and 2013, it yielded that there is a significant difference between these two variables.

ii. On observing the trends of diabetes change from 2009 to 2013, we see that, on an average, there is a 13.9% increase in diabetes rate across all counties with maximum increase of 80.8% in Roosevelt County, New Mexico.

Variable Name	Minimum	Maximum	Mean	Percent Missing
Diabetes_change	-39.1304	80.76923	13.87036	0

Figure 2: Summary Statistics of Target Variable

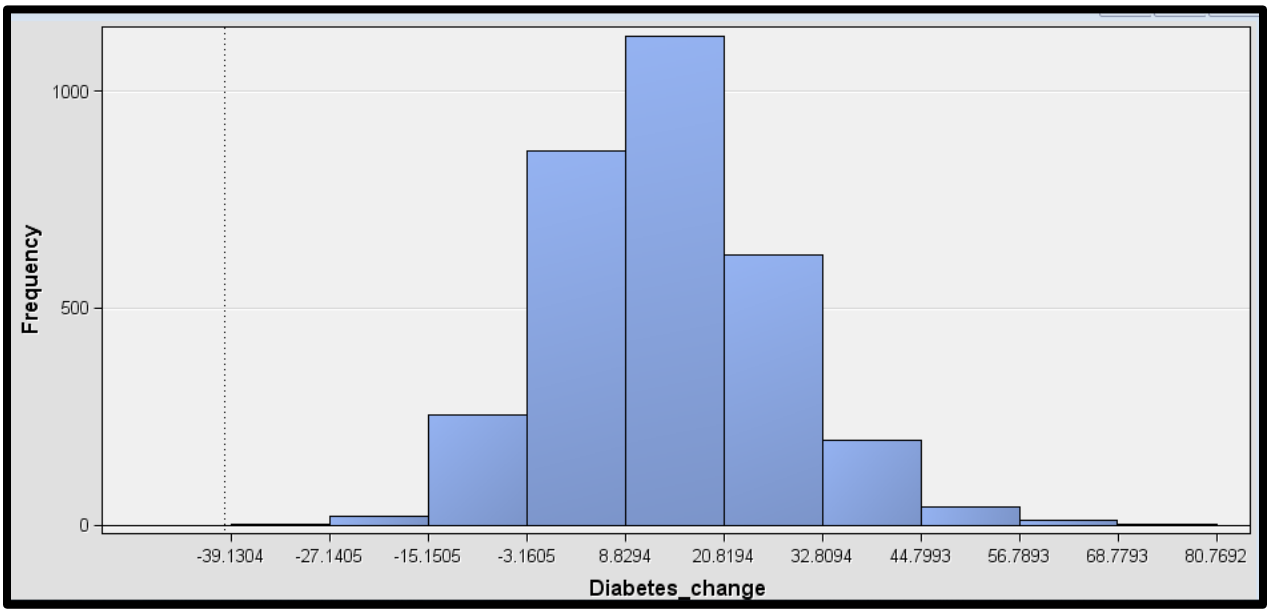

Figure 3: Distribution of Target variable

The negative values in this variable shows the counties that have decreased in diabetes from 2009 to 2013.

iii. There are 85% of counties with increase in diabetes rate while, there are 15% of counties with decrease in diabetes rate.

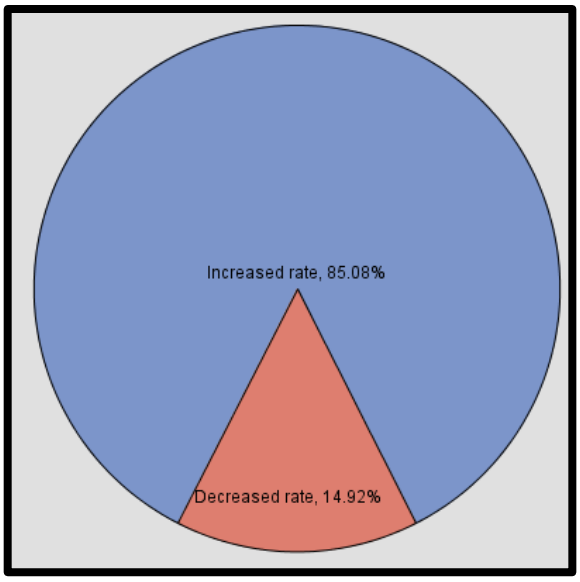

Figure 4: Pie diagram showing frequency of diabetes change indicator

iv. The diabetes rate has relationship with the obesity rate. On comparing the increase/decrease in obesity rate with diabetes rate, it is observed that the means of diabetes rate is significantly different between these two categories.

Dependent Variable: Diabetes_change

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	19033.4429	19033.4429	113.54	<.0001
Error	3135	525521.0326	167.6303		
Corrected Total	3136	544554.4754			

R-Square	Coeff Var	Root MSE	Diabetes_change Mean
0.034952	93.34443	12.94721	13.87036

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Obesity_change_indic	1	19033.44286	19033.44286	113.54	<.0001

Figure 5: ANOVA results for diabetes and obesity rate

v. Diabetes rate also has correlation with the proportion of races in each county. If a county has high African American population, then the diabetes percent is also high in that county. In contrast, if a county has high Asians, whites or Hispanics, then the diabetes rates are low.

Pearson Correlation Coefficients, N = 3137
Prob > |r| under H0: Rho=0

	PCT_DIABETES_ ADULTS08	PCT_DIABETES_ ADULTS13	PCT_ NHASIAN10	PCT_ NHBLACK10	PCT_ NHNA10	PCT_ NHPI10	PCT_ NHWHITE10	PCT_ HISP10
PCT_DIABETES_ADULTS08	1.00000	0.85767	-0.25498	0.54991	0.01574	-0.02717	-0.20521	-0.23828
PCT_DIABETES_ADULTS08		<.0001	<.0001	<.0001	0.3781	0.1282	<.0001	<.0001
PCT_DIABETES_ADULTS13	0.85767	1.00000	-0.29132	0.46164	0.02380	-0.04354	-0.10004	-0.29585
PCT_DIABETES_ADULTS13	<.0001		<.0001	<.0001	0.1827	0.0147	<.0001	<.0001
PCT_NHASIAN10	-0.25498	-0.29132	1.00000	0.02021	-0.01712	0.24091	-0.27095	0.14539
PCT_NHASIAN10	<.0001	<.0001		0.2579	0.3377	<.0001	<.0001	<.0001
PCT_NHBLACK10	0.54991	0.46164	0.02021	1.00000	-0.09812	-0.02431	-0.61565	-0.10423
PCT_NHBLACK10	<.0001	<.0001	0.2579		<.0001	0.1734	<.0001	<.0001
PCT_NHNA10	0.01574	0.02380	-0.01712	-0.09812	1.00000	0.00341	-0.29903	-0.04369
PCT_NHNA10	0.3781	0.1827	0.3377	<.0001		0.8487	<.0001	0.0144
PCT_NHPI10	-0.02717	-0.04354	0.24091	-0.02431	0.00341	1.00000	-0.09169	0.00528
PCT_NHPI10	0.1282	0.0147	<.0001	0.1734	0.8487		<.0001	0.7675
PCT_NHWHITE10	-0.20521	-0.10004	-0.27095	-0.61565	-0.29903	-0.09169	1.00000	-0.58683
PCT_NHWHITE10	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001
PCT_HISP10	-0.23828	-0.29585	0.14539	-0.10423	-0.04369	0.00528	-0.58683	1.00000
PCT_HISP10	<.0001	<.0001	<.0001	<.0001	0.0144	0.7675	<.0001	

Figure 6: Correlation Matrix of diabetes change with race variables

vi. The diabetes rate is also influenced by the percent of older or younger population in the county. A county with higher population above 65 years' age is having high diabetes rate. A county with younger population (below 18 years) is having low diabetes rate.

Pearson Correlation Coefficients, N = 3137
Prob > |r| under H0: Rho=0

	Diabetes_ change	PCT_ 65OLDER10	PCT_18YOUNGER10	AGE_COMBO
Diabetes_change	1.00000	0.12107	-0.06291	0.04262
		<.0001	0.0004	0.0170
PCT_65OLDER10	0.12107	1.00000	-0.53090	0.63958
PCT_65OLDER10	<.0001		<.0001	<.0001
PCT_18YOUNGER10	-0.06291	-0.53090	1.00000	-0.44427
PCT_18YOUNGER10	0.0004	<.0001		<.0001
AGE_COMBO	0.04262	0.63958	-0.44427	1.00000
	0.0170	<.0001	<.0001	

Figure 7: Correlation matrix of diabetes rate with Age variables

vii. On observing the prices of milk, soda, chips, we can see that they are correlated with the diabetes rate. Furthermore, taxes on soda and chips at vending machines has correlation with diabetes rate, while at retail stores are not.

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	Diabetes_ change	MILK_ PRICE10	SODA_ PRICE10	MILK_ SODA_ PRICE10	SODATAX_ VENDM14	SODATAX_ STORES14	CHIPSTAX_ STORES14	CHIPSTAX_ VENDM14	FOOD_ TAX14
Diabetes_change	1.00000	-0.15362	0.06119	-0.16143	-0.04570	0.01935	0.00174	-0.08787	0.00174
		<.0001	0.0006	<.0001	0.0105	0.2787	0.9222	<.0001	0.9222
	3137	3108	3108	3108	3137	3137	3137	3137	3137
MILK_PRICE10	-0.15362	1.00000	-0.03419	0.91639	0.14537	0.03747	0.05449	0.30828	0.05449
MILK_PRICE10	<.0001		0.0567	<.0001	<.0001	0.0367	0.0024	<.0001	0.0024
	3108	3108	3108	3108	3108	3108	3108	3108	3108
SODA_PRICE10	0.06119	-0.03419	1.00000	-0.42528	-0.11860	-0.07148	0.02424	0.00106	0.02424
SODA_PRICE10	0.0006	0.0567		<.0001	<.0001	<.0001	0.1767	0.9530	0.1767
	3108	3108	3108	3108	3108	3108	3108	3108	3108
MILK_SODA_PRICE10	-0.16143	0.91639	-0.42528	1.00000	0.17194	0.05231	0.03334	0.27881	0.03334
MILK_SODA_PRICE10	<.0001	<.0001	<.0001		<.0001	0.0035	0.0631	<.0001	0.0631
	3108	3108	3108	3108	3108	3108	3108	3108	3108
SODATAX_VENDM14	-0.04570	0.14537	-0.11860	0.17194	1.00000	0.77517	0.16086	0.70352	0.16086
SODATAX_VENDM14	0.0105	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001
	3137	3108	3108	3108	3137	3137	3137	3137	3137
SODATAX_STORES14	0.01935	0.03747	-0.07148	0.05231	0.77517	1.00000	0.20450	0.45548	0.20450
SODATAX_STORES14	0.2787	0.0367	<.0001	0.0035	<.0001		<.0001	<.0001	<.0001
	3137	3108	3108	3108	3137	3137	3137	3137	3137
CHIPSTAX_STORES14	0.00174	0.05449	0.02424	0.03334	0.16086	0.20450	1.00000	0.26525	1.00000
CHIPSTAX_STORES14	0.9222	0.0024	0.1767	0.0631	<.0001	<.0001		<.0001	<.0001
	3137	3108	3108	3108	3137	3137	3137	3137	3137
CHIPSTAX_VENDM14	-0.08787	0.30828	0.00106	0.27881	0.70352	0.45548	0.26525	1.00000	0.26525
CHIPSTAX_VENDM14	<.0001	<.0001	0.9530	<.0001	<.0001	<.0001	<.0001		<.0001
	3137	3108	3108	3108	3137	3137	3137	3137	3137
FOOD_TAX14	0.00174	0.05449	0.02424	0.03334	0.16086	0.20450	1.00000	0.26525	1.00000
FOOD_TAX14	0.9222	0.0024	0.1767	0.0631	<.0001	<.0001	<.0001	<.0001	
	3137	3108	3108	3108	3137	3137	3137	3137	3137

Figure 8: Correlation matrix of diabetes change with taxes and prices

viii. The recreation and fitness facilities in a county are also having significant relation with the diabetes rate in a county. As the number of recreation and fitness facilities are more the diabetes rate is low in a county.

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	Diabetes_ change	PCH_RECFAC_ 09_14
Diabetes_change	1.00000	-0.07702
		<.0001
	3137	3014
PCH_RECFAC_09_14	-0.07702	1.00000
PCH_RECFAC_09_14	<.0001	
	3014	3014

Figure 9: Correlation matrix of diabetes change with fitness facilities

ix. The government programs like farm to school, Summer school service and National school lunch program are also having significant relation with diabetes. As the number of people participating in these programs increases in a county, there is a chance that the diabetes rate also increasing. Also, observing if this is due to the increase in obesity due to these programs, we can see that there is a correlation with obesity and each program participants as well.

Pearson Correlation Coefficients, N = 3137
Prob > |r| under H0: Rho=0

	Diabetes_ change	Obesity_ change	Program_ Recipients	PCH_ NSLP_ 09_15	PCH_ SFSP_ 09_15
Diabetes_change	1.00000	0.26144 <.0001	0.04899 0.0061	0.05799 0.0012	0.05829 0.0011
Obesity_change	0.26144 <.0001	1.00000	-0.01603 0.3693	-0.06623 0.0002	0.06204 0.0005
Program_Recipients	0.04899 0.0061	-0.01603 0.3693	1.00000	0.75394 <.0001	0.57485 <.0001
PCH_NSLP_09_15	0.05799	-0.06623	0.75394	1.00000	-0.00180
PCH_NSLP_09_15	0.0012	0.0002	<.0001		0.9195
PCH_SFSP_09_15	0.05829	0.06204	0.57485	-0.00180	1.00000
PCH_SFSP_09_15	0.0011	0.0005	<.0001	0.9195	

Figure 10: Correlation Matrix of diabetes change with Programs

x. While, all the above factors are showing significant relation with diabetes rate, there are few other variables that are not showing any relation. They are proximity to store, metro/non-metro County and population loss in a county, poverty rate etc.

6. MODEL BUILDING

With these insights from the descriptive statistics, data was feed to the models. Models are built to see how well the trends in diabetes rates can be captured by food environment and lifestyle factors. The data is partitioned into 70-30 split to evaluate the performance of our model on validation dataset. Different predictive models are used to capture the trends in the data like Neural Net, Decision tree, Linear Regression, Ensemble models. The selection criterion is Average square error since we are trying to reduce the error between the actual and predicted value.

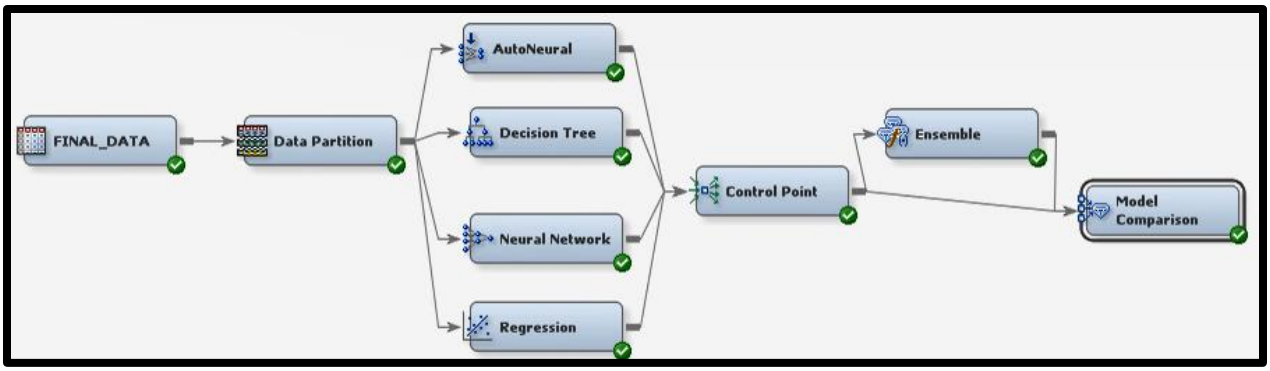

Figure 11: SAS EM Process Flow diagram for Modelling

On studying the results from different models, we can say that Decision tree is the best model, because it has the lowest Average square error.

Selected Model	Predecessor Node	Model Node	Model Description	Train: Target Variable	Selection Criterion: Valid: Average Squared Error	Train: Average Squared Error
Y	Tree	Tree	Decision Tr...	Diabetes_change	169.9934	134.6803
	AutoNeural	AutoNeural	AutoNeural	Diabetes_change	187.754	165.6618
	Neural	Neural	Neural Net...	Diabetes_change	188.6325	165.1209
	Ensmbl	Ensmbl	Ensemble	Diabetes_change	507.3969	377.0847
	Reg	Reg	Regression	Diabetes_change	5767.387	3949.46

Figure 12: Model comparison results

From the variable importance table, we can say that diabetes change is mostly influenced by obesity change variable followed by SNAP participants, percent of direct sales in each county.

Variable Name	Importance	Validation Importance	Ratio of Validation to Training Importance
Obesity_change	1.0000	1.0000	1.0000
SNAP_PART_RATE13	0.7291	0.4345	0.5960
PC_DIRSALES12	0.6910	0.2969	0.4297
PC_FFRSALES12	0.4354	0.4367	1.0030
pct_p2009	0.3986	0.5071	1.2722
LACCESS_HHNV10	0.3920	0.0000	0.0000
PCT_NHNA10	0.3895	0.1538	0.3948
PCT_LOCLSALE12	0.3798	0.0000	0.0000
PCT_HSPA15	0.3680	0.5379	1.4617
PCT_REDUCED_LUNCH14	0.3180	0.1855	0.5834
PCT_LACCESS_NHASIAN15	0.3155	0.3439	1.0902
SNAP_BBCE16	0.2940	0.3849	1.3091
PCH_VEG_FARMS_07_12	0.2625	0.3137	1.1949
PC_DIRSALES07	0.2140	0.2569	1.2003

Figure 13: Variable importance table

The highest predicted diabetes rate change of 21%, from the above decision tree, is when SNAP participants are greater than 86% and obesity rate greater than 20%.

7. RESULTS

- It is seen that 85% counties have increased in diabetes rate, with a maximum increase of 80.8% in Roosevelt County, New Mexico.
- Diabetes is seen to be dependent on population percentage of race. The counties with high African American percentages are having more diabetes rates.
- Diabetes rate is higher in counties having higher older population who are older than 65 years' age, while counties with higher younger population, has lower diabetes rate.
- Counties with higher prices on milk, chips, sodas have lower diabetes rate. Also, if taxes for Soda and Chips at vending machines are high then diabetes rate is relatively low.
- On observing the participants of physical fitness activities, recreational activities in a county, it is seen that as participant rate increases, the diabetes rate decreases.
- With increase of participation in government programs such as farm to school, Summer school service and National school lunch programs, the diabetes rate is also increases. Another important fact is that; these counties have higher obesity.
- The counties having SNAP participants greater than 86% and obesity rate greater than 20% are predicted to have the highest increase in diabetes change of about 21%.
- The counties with SNAP participants less than 86%, percent of direct sales less than 4% and obesity rate less than 5% are predicted to have the highest decrease in diabetes rate of about 5%.

8. CONCLUSIONS

- Food environment factors such as SNAP participation, direct farm sales, taxes on junk foods affect the change in diabetes rate.
- Lifestyle factors like physical fitness levels, fast food expenditures in a county affect the diabetes rate.
- Traditional factors such as ethnicity, age, obesity also have an effect on diabetes rate.

9. RECOMMENDATIONS

- Taxes act on consumer behavior by changing the cost of different choices relative to one another. If unhealthy foods like sodas and chips are cheap to buy, then raising their price through taxation provides a price signal — although without removing choice altogether. A 2012 review of health-related food taxes found that, if carefully designed, these could be effective in shifting patterns of consumption towards healthier foods, with a 20% tax suggested as the minimum rate for effectiveness.
- SNAP participation was associated with obesity and USDA report published last year found that 20 cents of every SNAP dollar was spent on sweetened drinks, desserts, salty snacks, candy, and sugar. As, people who enroll in SNAP are the kinds of people—stressed-out, poor, less educated—who are

more likely to be obese for unrelated reasons. Government needs to create awareness program for people about the causes of obesity and diabetes alongside with SNAP program.

10. ACKNOWLEDGEMENTS

We thank Dr. Goutam Chakraborty, SAS® Professor of Marketing Analytics and Dr. Miriam McGaugh, Clinical Professor at Oklahoma State University for their constant support and guidance throughout this project

12. REFERENCES

- Diabetes definition and statistics-

  https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes

- SNAP Program issues – https://www.theatlantic.com/health/archive/2017/05/the-messy-relationship-between-food-stamps-and-health/527820/
- Diabetes statistics -

  https://www.cdc.gov/diabetes/statistics/slides/long_term_trends.pdf

CONTACT INFORMATION:

Your comments, feedback and questions are valued and encouraged. You can contact the author at:

Mounica Mandapati

Oklahoma State University, Stillwater OK

Email: mounica.mandapati@okstate.edu / mounicaraaj.mandapati@gmail.com

Phone no.: 817-320-9159