# Breast Cancer in Washington State (2007-2017)

Precious Stowers

# Topic and Background

Cancer is a known disease that has taken the lives of too many loved ones for too many years. As cancer has taken the lives of both young and old, I have not seen as much coverage on the concurrent topic - and I wanted to explore it.

In the Level 5 Data Analysis, I will be viewing the Breast Cancer Data from the Center of Diseases Control & Prevention (CDC) from 2007-2017 in Washington State.

# Level 5 Analysis Questions

Can I predict a categorical value from a data set?

Should I use a deepnet or ensemble to make my predictions?

How does my chosen model perform?

What factors are most influential in predicting the categorical variable?

According to my scenario, what are my predictions and what should I recommend?

# Data Source + Data Gathering

**Source Name:** The CDC (Center of Disease Control)

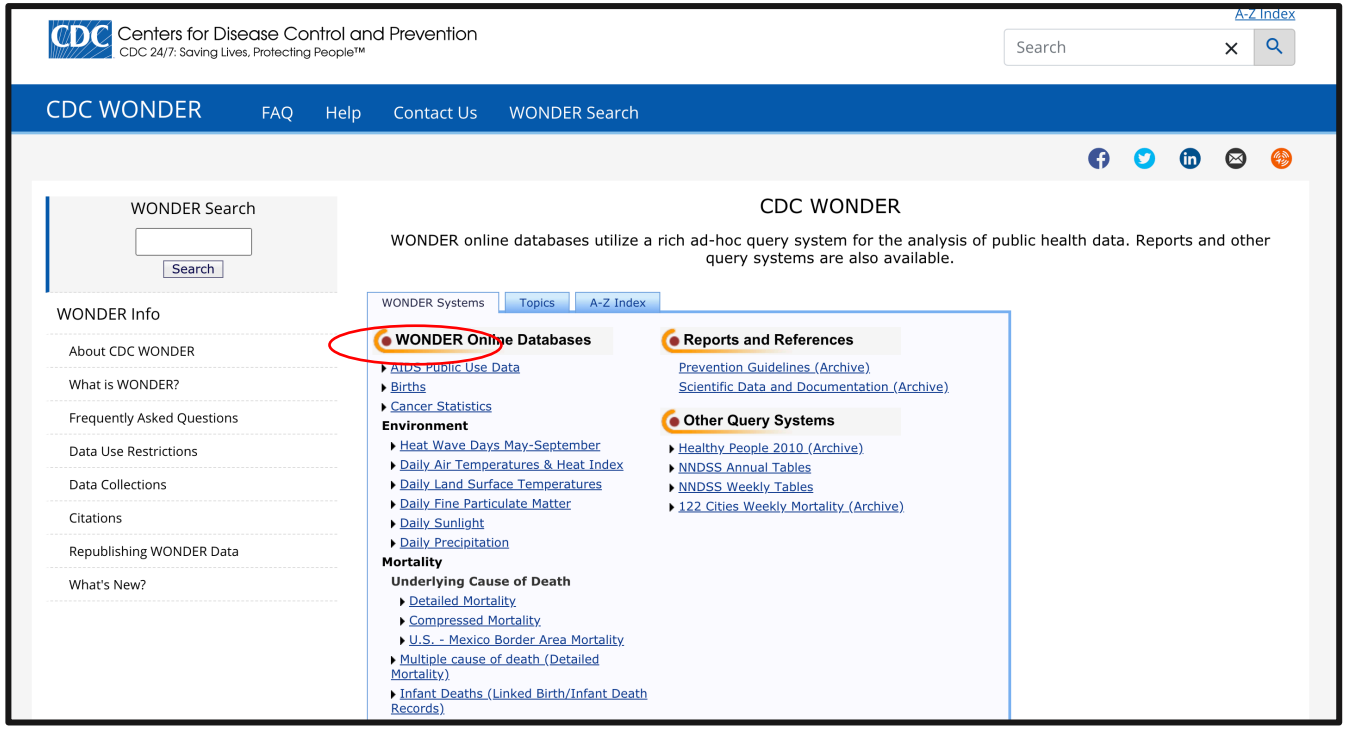
**Data Type:** Official

Gathering the data – I used the CDC Wonder – which allowed me to pick and create a data set from their larger data base

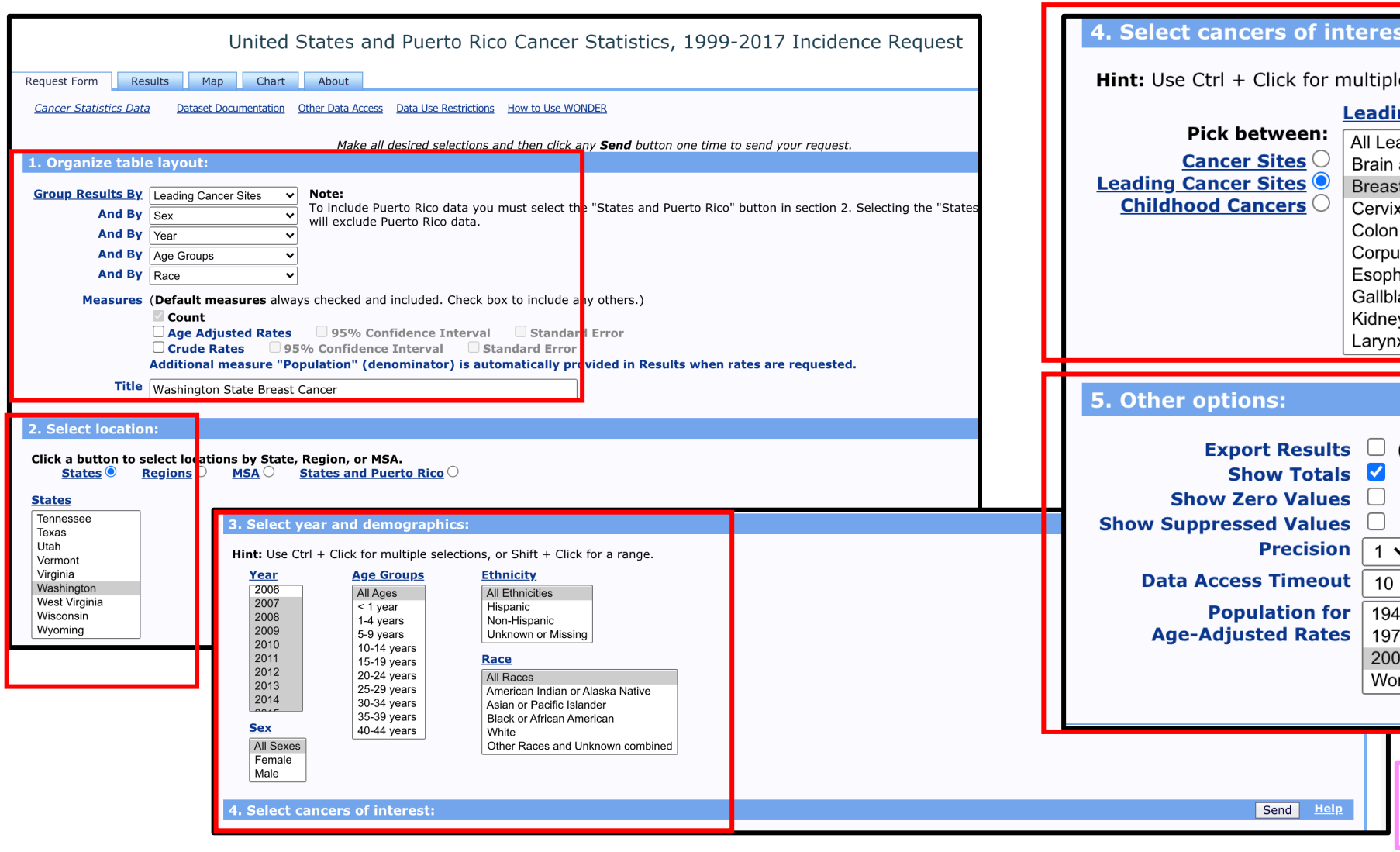
**CDC Wonder Website:** https://wonder.cdc.gov/

**CDC Wonder Cancer Statistics Request form:** https://wonder.cdc.gov/cancer-v2017.html

**Data Summary Descriptions:** https://wonder.cdc.gov/DataSets.html



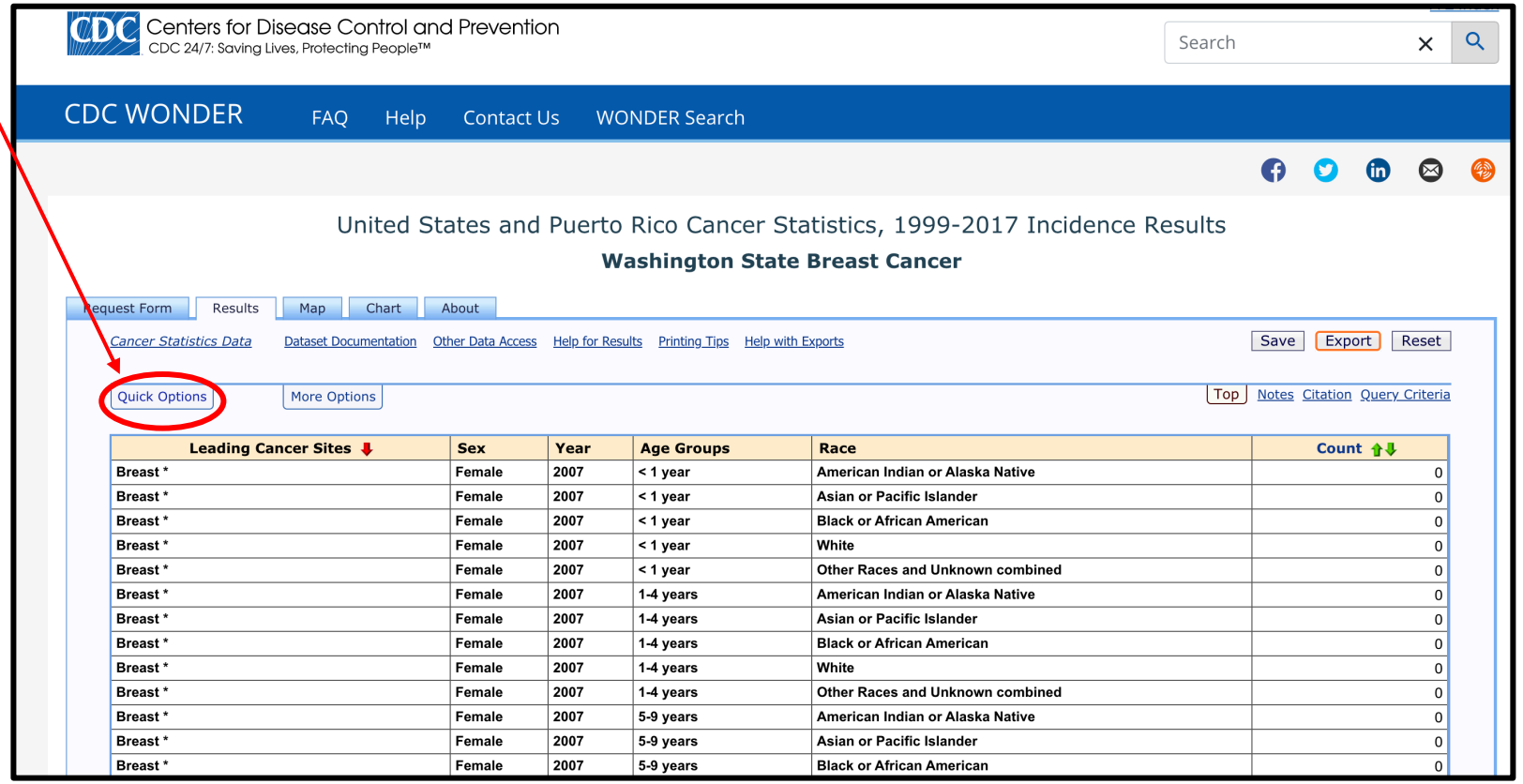
# Data Gathering Continued: the variables selected




**CDC Wonder Cancer Statistics Request form:**
https://wonder.cdc.gov/cancer-v2017.html

# Data Set View:

From “Quick Options” – I selected the “show zeros” and “show suppressed” values for the data set



Centers for Disease Control and Prevention
CDC 24/7: Saving Lives, Protecting People™

Search

CDC WONDER FAQ Help Contact Us WONDER Search

United States and Puerto Rico Cancer Statistics, 1999-2017 Incidence Results

**Washington State Breast Cancer**

Request Form | Results | Map | Chart | About

*Cancer Statistics Data* | Dataset Documentation | Other Data Access | Help for Results | Printing Tips | Help with Exports

Save | Export | Reset

Quick Options | More Options

Top | Notes | Citation | Query Criteria

| Leading Cancer Sites | Sex | Year | Age Groups | Race | Count |
|---|---|---|---|---|---|
| Breast * | Female | 2007 | < 1 year | American Indian or Alaska Native | 0 |
| Breast * | Female | 2007 | < 1 year | Asian or Pacific Islander | 0 |
| Breast * | Female | 2007 | < 1 year | Black or African American | 0 |
| Breast * | Female | 2007 | < 1 year | White | 0 |
| Breast * | Female | 2007 | < 1 year | Other Races and Unknown combined | 0 |
| Breast * | Female | 2007 | 1-4 years | American Indian or Alaska Native | 0 |
| Breast * | Female | 2007 | 1-4 years | Asian or Pacific Islander | 0 |
| Breast * | Female | 2007 | 1-4 years | Black or African American | 0 |
| Breast * | Female | 2007 | 1-4 years | White | 0 |
| Breast * | Female | 2007 | 1-4 years | Other Races and Unknown combined | 0 |
| Breast * | Female | 2007 | 5-9 years | American Indian or Alaska Native | 0 |
| Breast * | Female | 2007 | 5-9 years | Asian or Pacific Islander | 0 |
| Breast * | Female | 2007 | 5-9 years | Black or African American | 0 |

**About Data:**

- Data contained information from the years 2007 – 2017
- Columns included:
  - Leading cancer site
  - Sex
  - Year
  - Age Groups
  - Race
  - Count

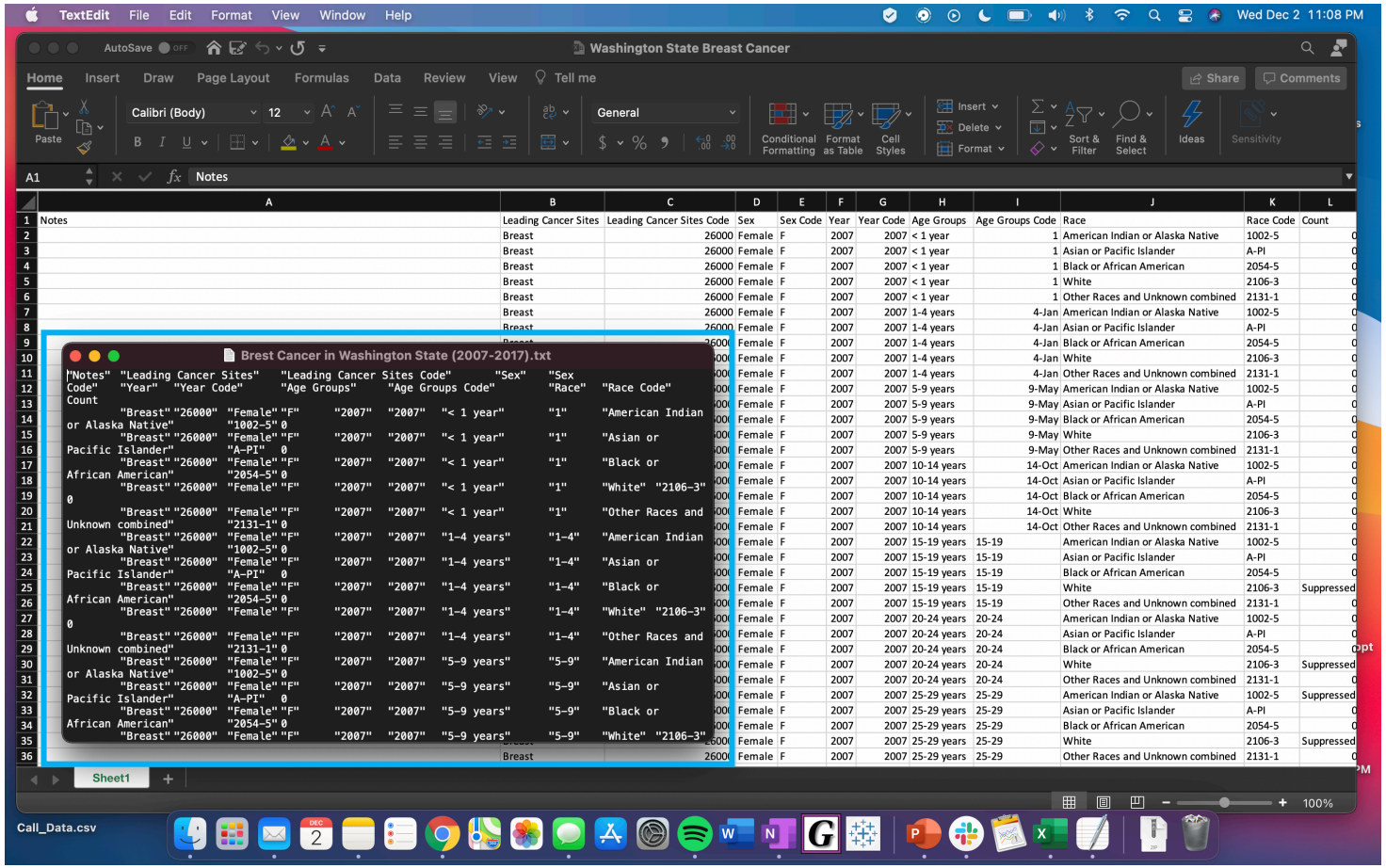
**Data set citation:** United States Cancer Statistics - Incidence: 1999 - 2017, WONDER Online Database. United States Department of Health and Human Services, Centers for Disease Control and Prevention and National Cancer Institute; 2020. Accessed at http://wonder.cdc.gov/cancer-v2017.html on Dec 5, 2020 10:26:31 PM

# Data Source Gathering Continued

When one downloads their dataset from the CDC Wonder interactive database, they must download it as a txt file (light blue box) and then insert it into excel to then download the data set as a csv file to input into BigML

**Caveats:**

1. If there was more than ~99,000~ rows wanted from the data variables chosen, it would not return a data set and one would have to change their request
2. One had to select a certain number of variables and could not select every variable/option listed
    - This forces the user to limit the data they can see at once from the website



# Data Integrity: Med-High

1. **Source cleanliness: High** - The source overall was clean, especially since I had the opportunity to pick the data for my wanted data set
2. **Appropriate collection methods: Med/High/Unknown** -The source appears to be clean with its collection and CDC is also known for gathering health information about different topics
3. **Collection ethics: Med/High** – Once the data was collective from WONDER, the data appeared to be clean and all the selected data was on the data set
4. **Source credibility: Med/High** – the Center of disease control is known for collecting data about health over a long period of time, as they are also gathering information about the Coronavirus Pandemic right now
5. **Appropriate provenance and curation: Med/High** - The data set overall is a little older, however the data appears to cover the large amount of selected (2007-2017) without large gaps
6. **Data bias: Med/Unknown** – the data from the CDC themselves may not be biased when they received the data from other hospital/clinic—like organizations. The organizations themselves when record the data from their patients may be biased because there are multiple trials of testing when testing for cancer and mistakes/faulty recordings may happen which could create bias.
7. **Data fairly collected: Med/Unknown** – the data from the CDC appears to be collected fairly because of the lack of blanks within the data. The data also states reasons for why they suppressed them.

# Data Integrity Continued

**8. Algorithmic bias: Med/Unknown –**Historically, White people were treated the best and the most represented within the health industry. This could affect algorithms how algorithms record data today and its affects on marginalized people.

**9. Algorithmic transparency: Unknown** – From the CDC they don't state specifically how they are transparent with their algorithm. The CDC also state they partner with many National organizations including: North American Association of Central Cancer Registries (NAACCR) & United States Department of Health and Human Services (US DHHS)

- The national organizations can make this data set reputable – although I would question the transparency due to the amount of local, state and national systems this data had to go through

**10. Data weaponization potential: Possibly Yes/Unknown** – health records can help determine political stances, such as the patients who receives health care and how much federal/state funding hospitals should receive. Furthering, if there are breast cancer patients and they are not recorded within the hospital system, these patients might receive less care and/or less financial support due to the lack of representation within the system. Health records are is also very personal to a user/patient in comparison to other recorded data.

# Data Set Note + Caveat

**Caveats:**
The “suppressed data” disproportionally suppresses minority groups from being represented in the data, because there is an overall smaller population of minorities groups in comparison to White people.

Privacy is important – but equal representation is unattainable through this data set

**Notes:**

**Caveats:** 'Suppressed' is displayed for data values when they must not be provided in order to protect personal privacy. Data is suppressed if fewer than 16 cases are reported in any one line of the results table. Data for the "Asian / Pacific Islander" and "American Indian or Alaska Native" race categories are suppressed at the Metropolitan Statistical Area level for populations less than 50,000 persons. Data are suppressed at the state and Metropolitan Statistical Area level for certain race and ethnicity groups: 1) "American Indian or Alaska Native" race data are suppressed at the state and MSA level for Delaware, Illinois, Kansas, Kentucky, New Jersey, and New York; 2) "Asian or Pacific Islander" race data are suppressed at the state and MSA level for Delaware, Illinois, Kansas, and Kentucky; 3) "Hispanic" ethnicity data are supressed at the state and MSA level for Delaware, and Kentucky. 4) All combinations of race by ethnicity are supressed at the state and MSA level for Delaware, Kansas, Kentucky, Pennsylvania, and Massachusetts. More information.

The "Show Totals" checkbox is disabled when data are grouped by Cancer Sites or by Childhood Cancer, because aggregate values are displayed in the table. Also be aware that charts and maps containing both aggregate and detail data could be misleading.

Age-adjusted rates are not available when results are grouped by age.

Data are from selected statewide and metropolitan area cancer registries that meet data quality criteria. Incidence data shown for the United States are diagnoses and populations from included registries only.

Information on primary site, behavior, and histology was coded according to the International Classification of Diseases for Oncology, Third Edition (ICD-O-3) and categorized according to the revised SEER recodes dated January 27, 2003, which define standard groupings of primary cancer sites.

For the 2005 year, the Census Bureau estimates that 203,937 persons were displaced from Alabama, Louisiana, Mississippi and Texas due to Hurricanes Katrina and Rita. CDC WONDER does not include the displaced persons in the 2005 population counts for these states, nor are these counts included in the summary populations for the affected division, regions or national population. However, the USCS web site does include these displaced persons in the national population figures for 2005.

**Help:** See United States and Puerto Rico Cancer Statistics, 1999-2017 Incidence Documentation for more information.

**Query Date:** Dec 5, 2020 10:26:31 PM

# Scenario – Before Analysis

**Scenario Background:** There has been many cases of Breast Cancer in the past in Washington State, with different attributes of patients including age group, race, sex and the year that happened. Washington State is deciding if they should increase state funding to help Breast Cancer research and what demographic of the population should they be expecting in Cancer patients (based on age group, race and sex)?

**Scenario Question(s):**

1. Based on age group, race and sex - what race will most likely have the most amount of reported Breast Cancer incidences?
2. Can this prediction be targeted for future years of Breast Cancer Incidences?
3. Should Washington State increase funding?

**Task(s) to do:** Create and Provide an Algorithm based on the amount of Cancer Incidences in Washington State and their demographics – to decide if funding should be increased

**Predicted Outcome:** Out of age group, race, sex and year – White people of any age will have the most reported Breast Cancer incidences. This outcome will be able to be targeted for future years. (This will be our focus throughout the analysis)

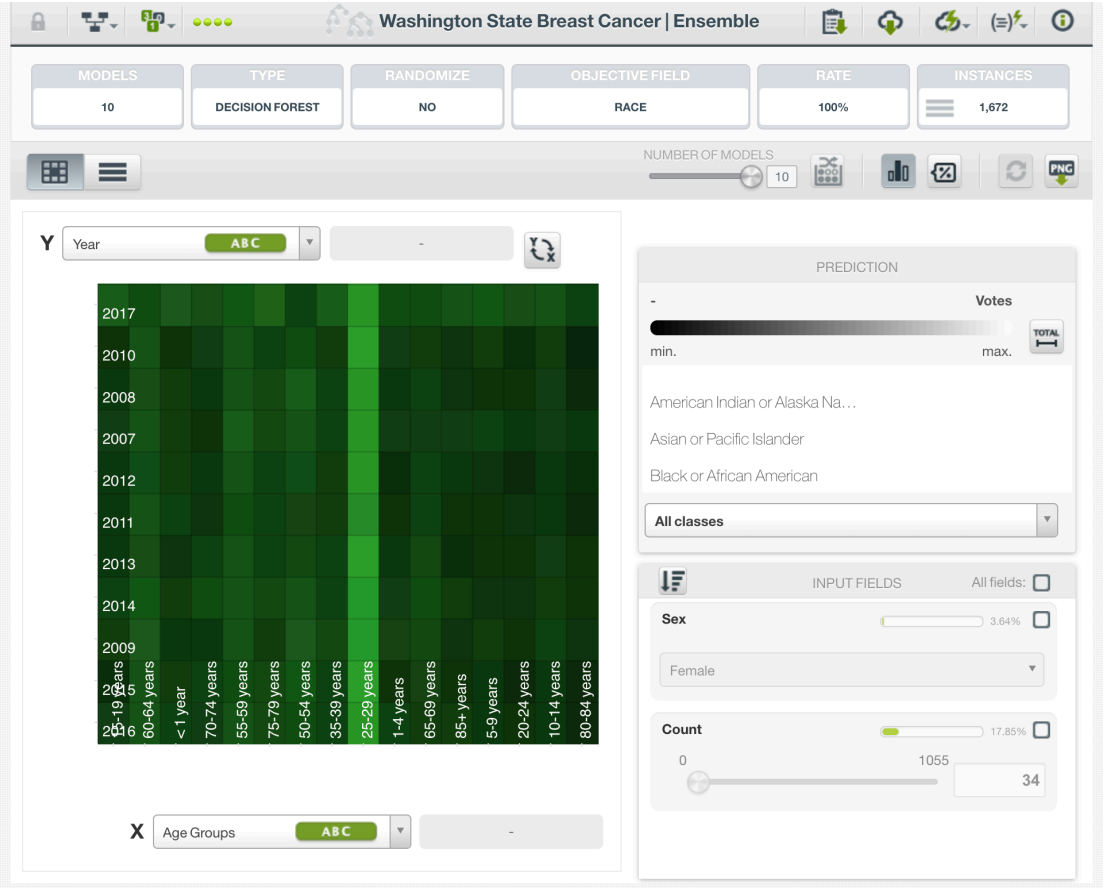
**Background for prediction:** White people currently have the largest population in the U.S, especially in Washington State.

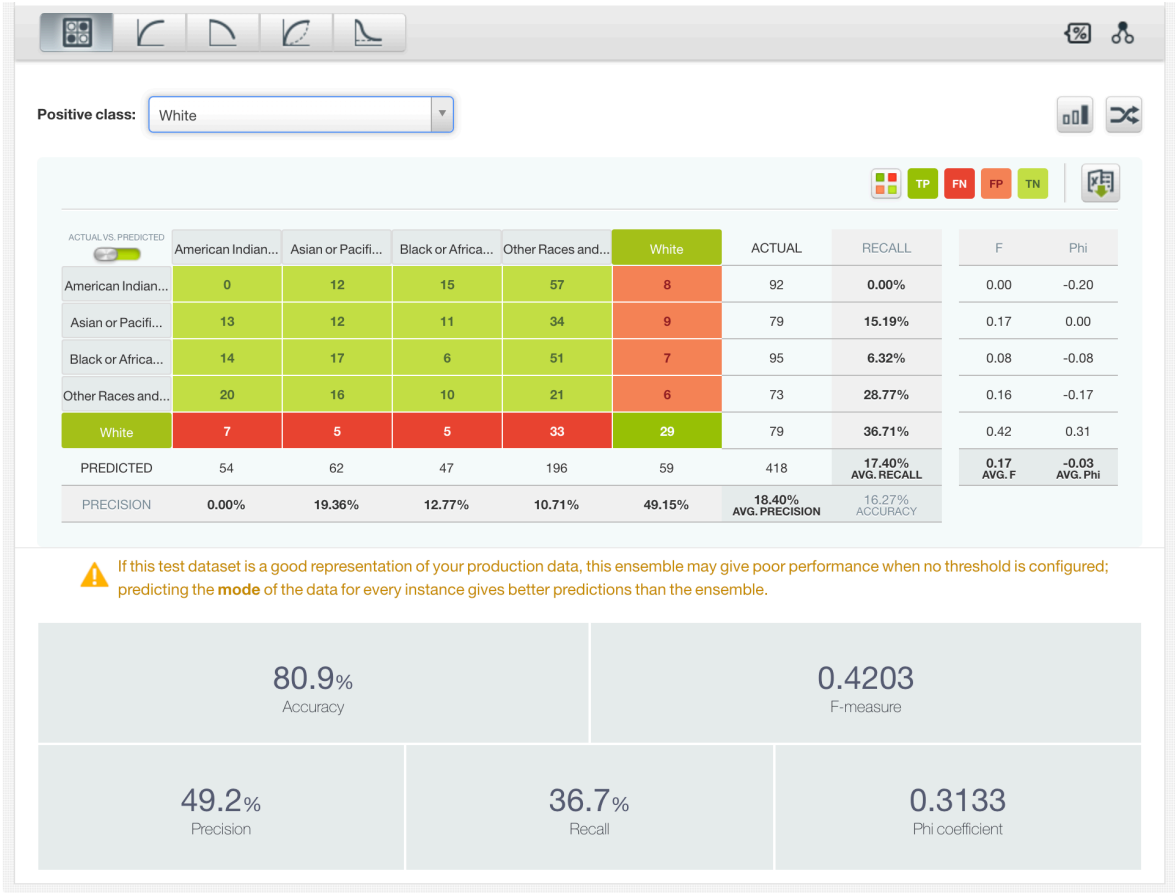
**Predicted Variable (objective Variable): Race**

Can I predict a categorical value from a data set?

# Ensemble

## Objective Variable: Race



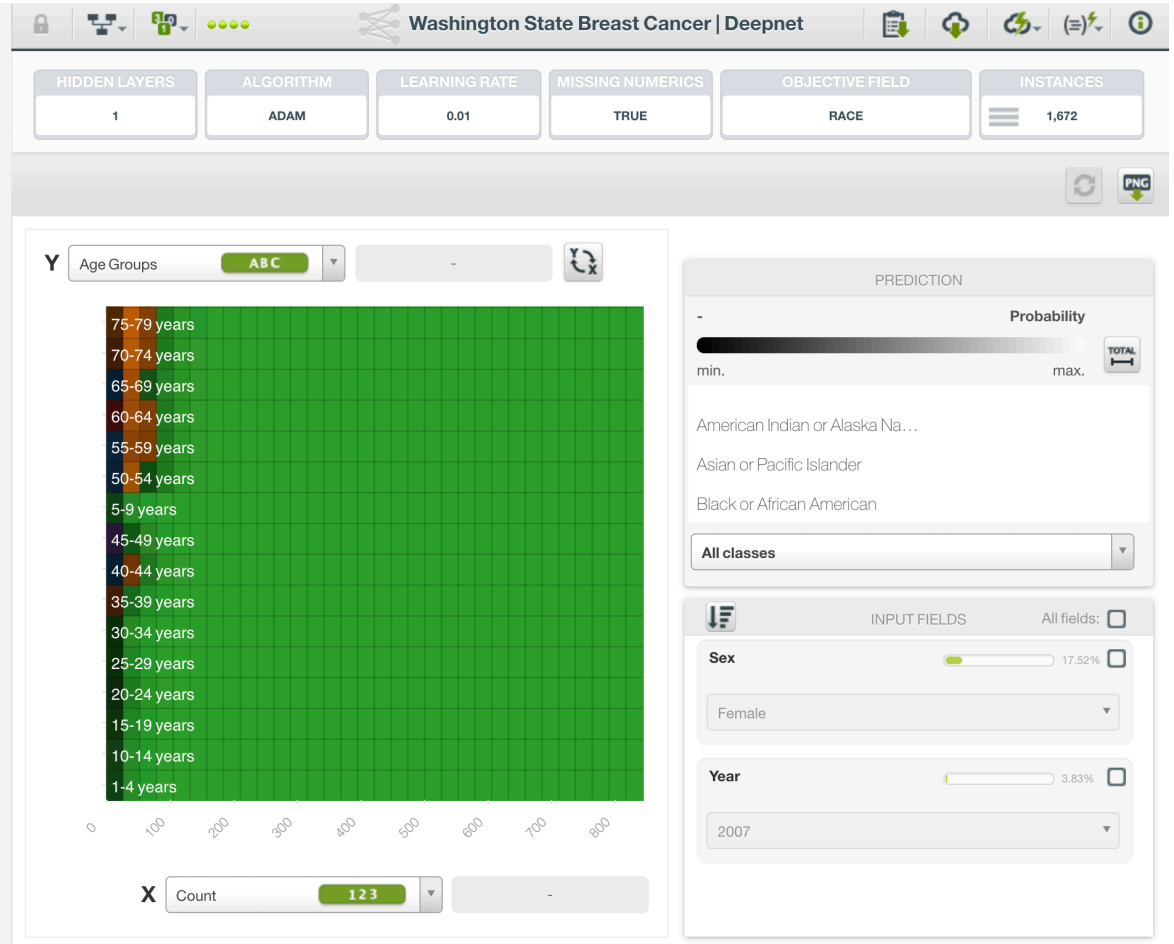



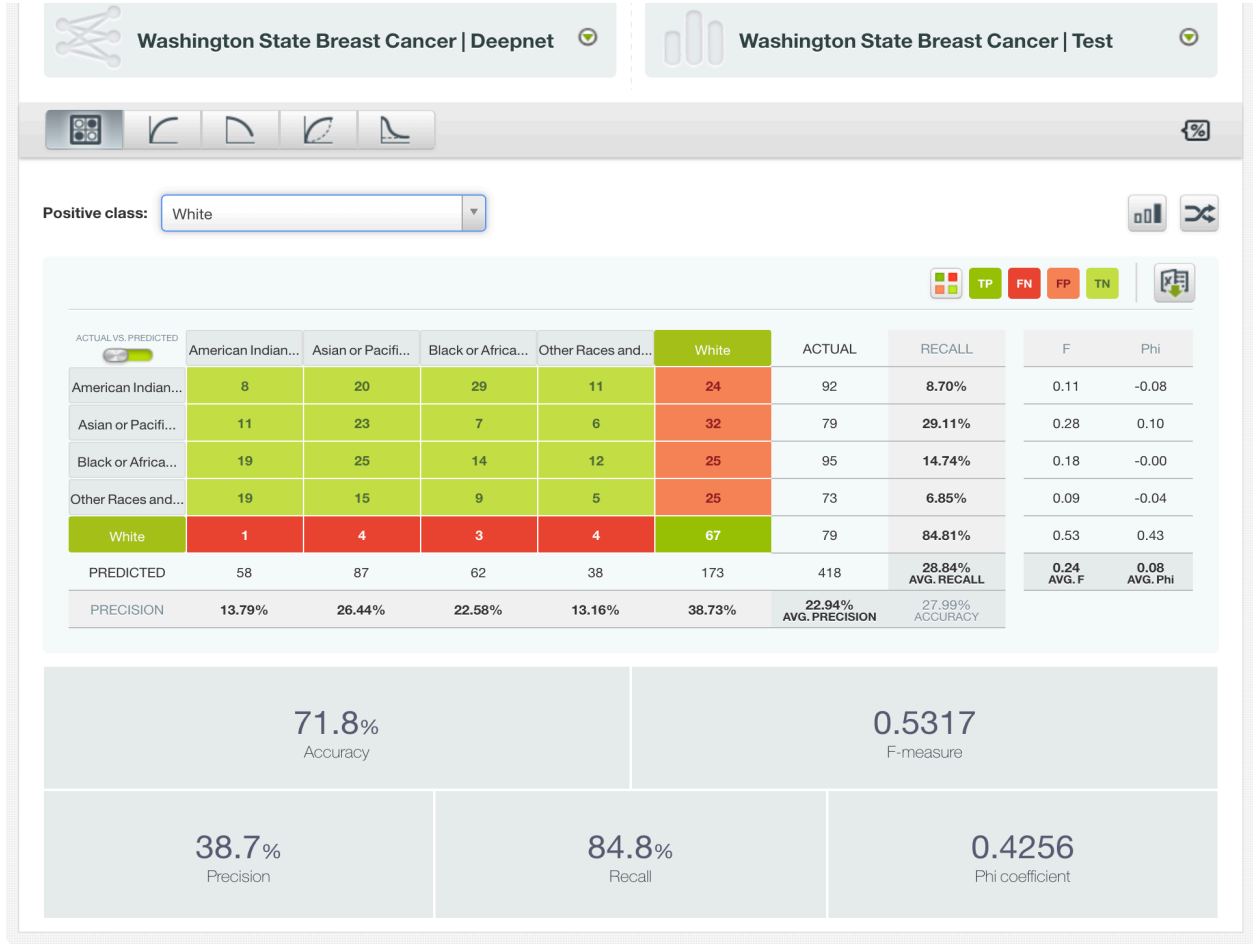

[Click Here to View in BigML]

[Click Here to View in BigML]

# Deepnet

Objective Variable: Race




[Click Here to View in BigML]

[Click Here to View in BigML]

# Can I predict a categorical value from a data set?

## Findings:

**Yes** – I can predict a categorical value from a data set

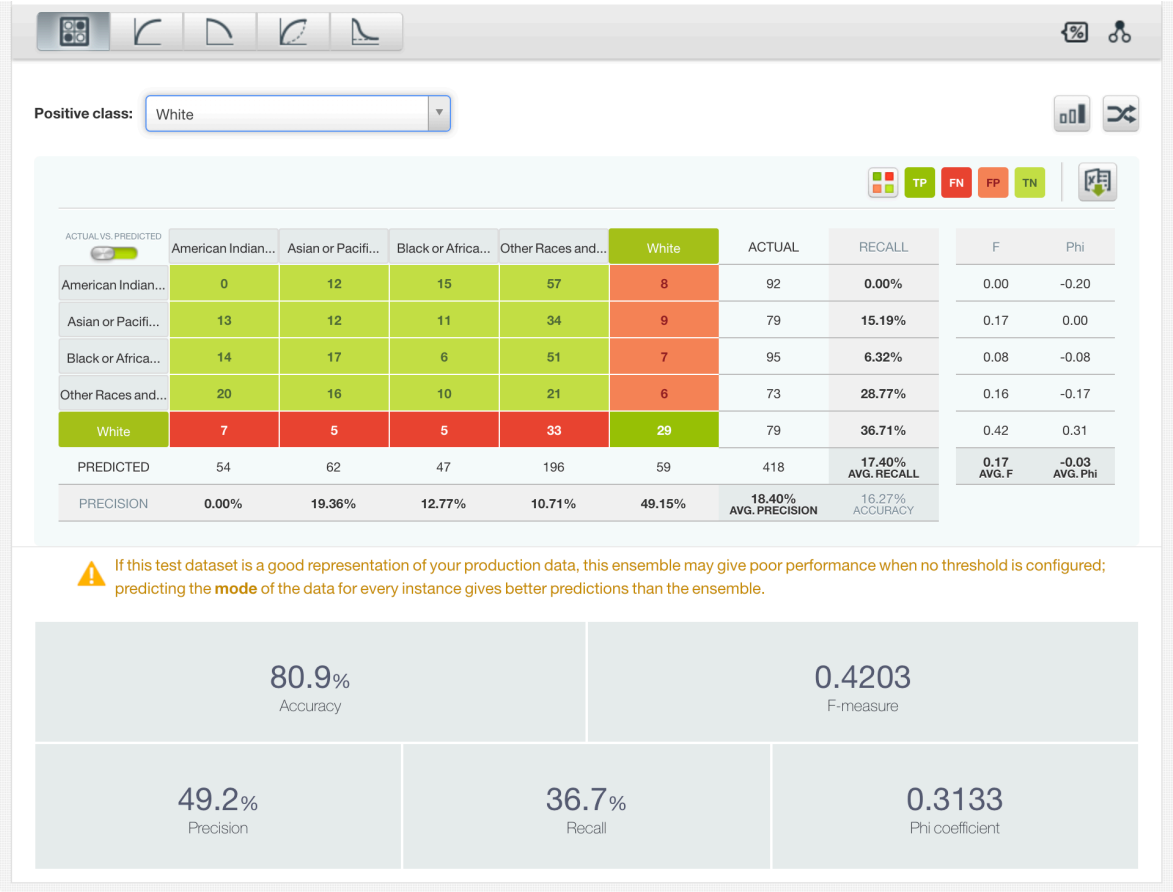
- Both the Ensemble and the Deepnet was able to predict the race of a Breast cancer patient based on the data set

## Caveats:

From previously stated before, we must be aware that not everyone is represented as numbers within this data set and could skew further predictions

Should I use a Deepnet or Ensemble to make my predictions?

# Ensemble & Deepnet Eval Comparison



Positive class: White

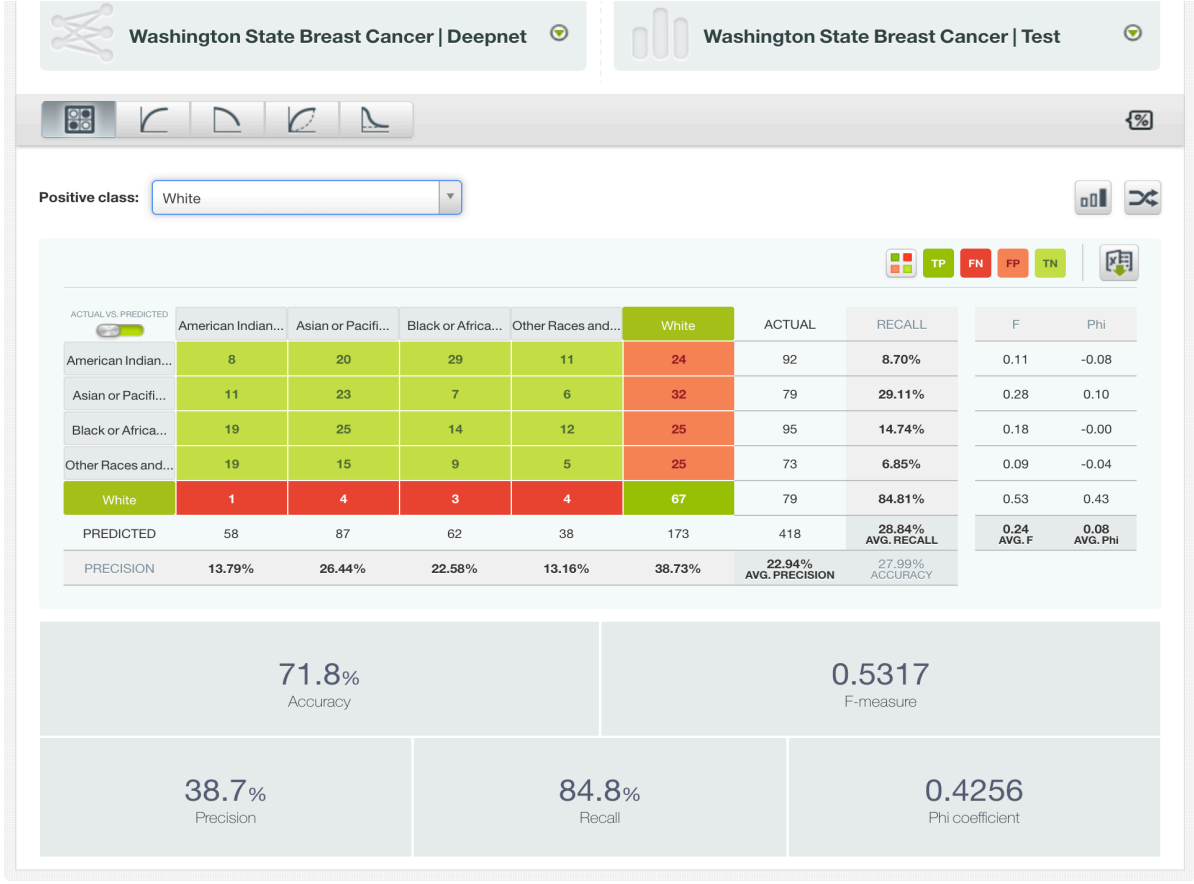
TP FN FP TN

| ACTUAL VS. PREDICTED | American Indian... | Asian or Pacifi... | Black or Africa... | Other Races and... | White | ACTUAL | RECALL | F | Phi |
|---|---|---|---|---|---|---|---|---|---|
| American Indian... | 0 | 12 | 15 | 57 | 8 | 92 | 0.00% | 0.00 | -0.20 |
| Asian or Pacifi... | 13 | 12 | 11 | 34 | 9 | 79 | 15.19% | 0.17 | 0.00 |
| Black or Africa... | 14 | 17 | 6 | 51 | 7 | 95 | 6.32% | 0.08 | -0.08 |
| Other Races and... | 20 | 16 | 10 | 21 | 6 | 73 | 28.77% | 0.16 | -0.17 |
| White | 7 | 5 | 5 | 33 | 29 | 79 | 36.71% | 0.42 | 0.31 |
| PREDICTED | 54 | 62 | 47 | 196 | 59 | 418 | 17.40% AVG. RECALL | 0.17 AVG. F | -0.03 AVG. Phi |
| PRECISION | 0.00% | 19.36% | 12.77% | 10.71% | 49.15% | 18.40% AVG. PRECISION | 16.27% ACCURACY | | |

If this test dataset is a good representation of your production data, this ensemble may give poor performance when no threshold is configured; predicting the **mode** of the data for every instance gives better predictions than the ensemble.

| | | |
|---|---|---|
| 80.9% Accuracy | 0.4203 F-measure | |
| 49.2% Precision | 36.7% Recall | 0.3133 Phi coefficient |

[Click Here to View in BigML]



Washington State Breast Cancer | Deepnet

Washington State Breast Cancer | Test

Positive class: White

TP FN FP TN

| ACTUAL VS. PREDICTED | American Indian... | Asian or Pacifi... | Black or Africa... | Other Races and... | White | ACTUAL | RECALL | F | Phi |
|---|---|---|---|---|---|---|---|---|---|
| American Indian... | 8 | 20 | 29 | 11 | 24 | 92 | 8.70% | 0.11 | -0.08 |
| Asian or Pacifi... | 11 | 23 | 7 | 6 | 32 | 79 | 29.11% | 0.28 | 0.10 |
| Black or Africa... | 19 | 25 | 14 | 12 | 25 | 95 | 14.74% | 0.18 | -0.00 |
| Other Races and... | 19 | 15 | 9 | 5 | 25 | 73 | 6.85% | 0.09 | -0.04 |
| White | 1 | 4 | 3 | 4 | 67 | 79 | 84.81% | 0.53 | 0.43 |
| PREDICTED | 58 | 87 | 62 | 38 | 173 | 418 | 28.84% AVG. RECALL | 0.24 AVG. F | 0.08 AVG. Phi |
| PRECISION | 13.79% | 26.44% | 22.58% | 13.16% | 38.73% | 22.94% AVG. PRECISION | 27.99% ACCURACY | | |

| | | |
|---|---|---|
| 71.8% Accuracy | 0.5317 F-measure | |
| 38.7% Precision | 84.8% Recall | 0.4256 Phi coefficient |

[Click Here to View in BigML]

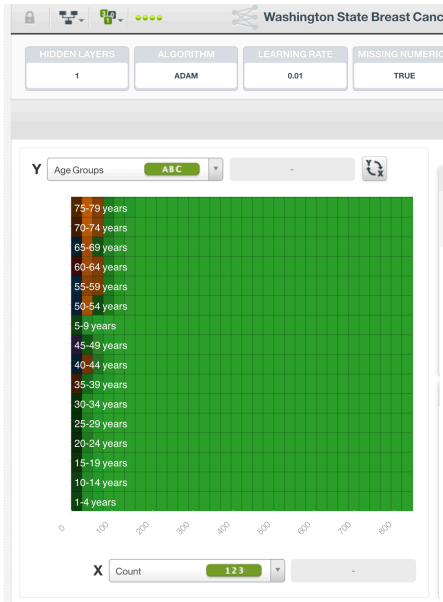
# Should I use a Deepnet or Ensemble to make my predictions?

A **Deepnet** is better to make my prediction because:

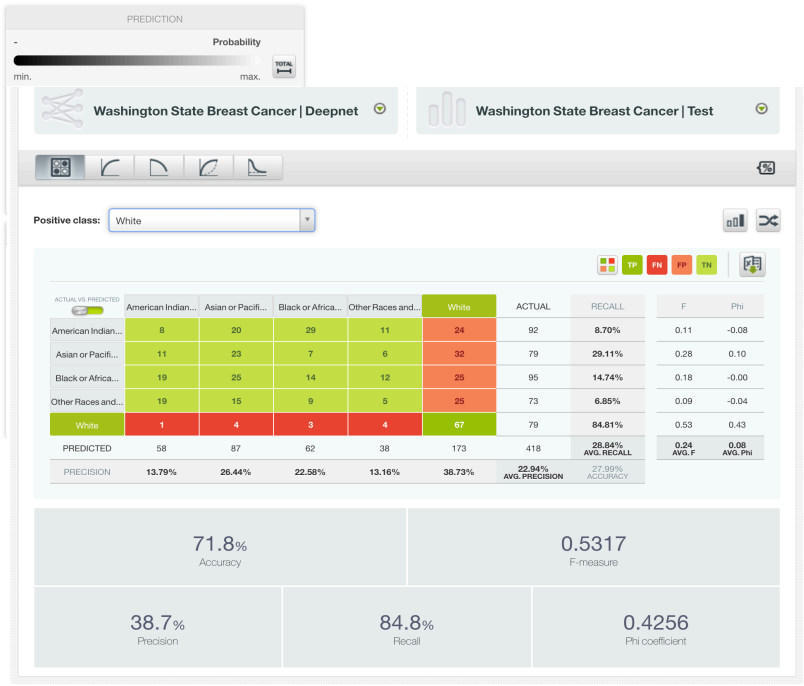
1. A Deepnet thinks more in terms of a human – which is crucial when predicting health care information
2. There was not a warning sign when the Deepnet evaluation appeared
3. There are not as many columns and the Deepnet appears to have handled less columns better than an Ensemble, due to the warning sign given on the Ensemble and not on the Deepnet Model
4. The Performance level was higher for the Deepnet than for the Ensemble

## Caveats:

1. An ensemble may help determine predictions better the variables did not have many interpretations/scales of the variable - because there are many variables within the data of health care
2. “0” counts and “Suppressed Counts” – which means that many people were not counted in the data due to privacy reasons which would skew the data
3. There still could be bias in the data from the number of systems the data had to “go through” before it reached the CDC – which could skew the data



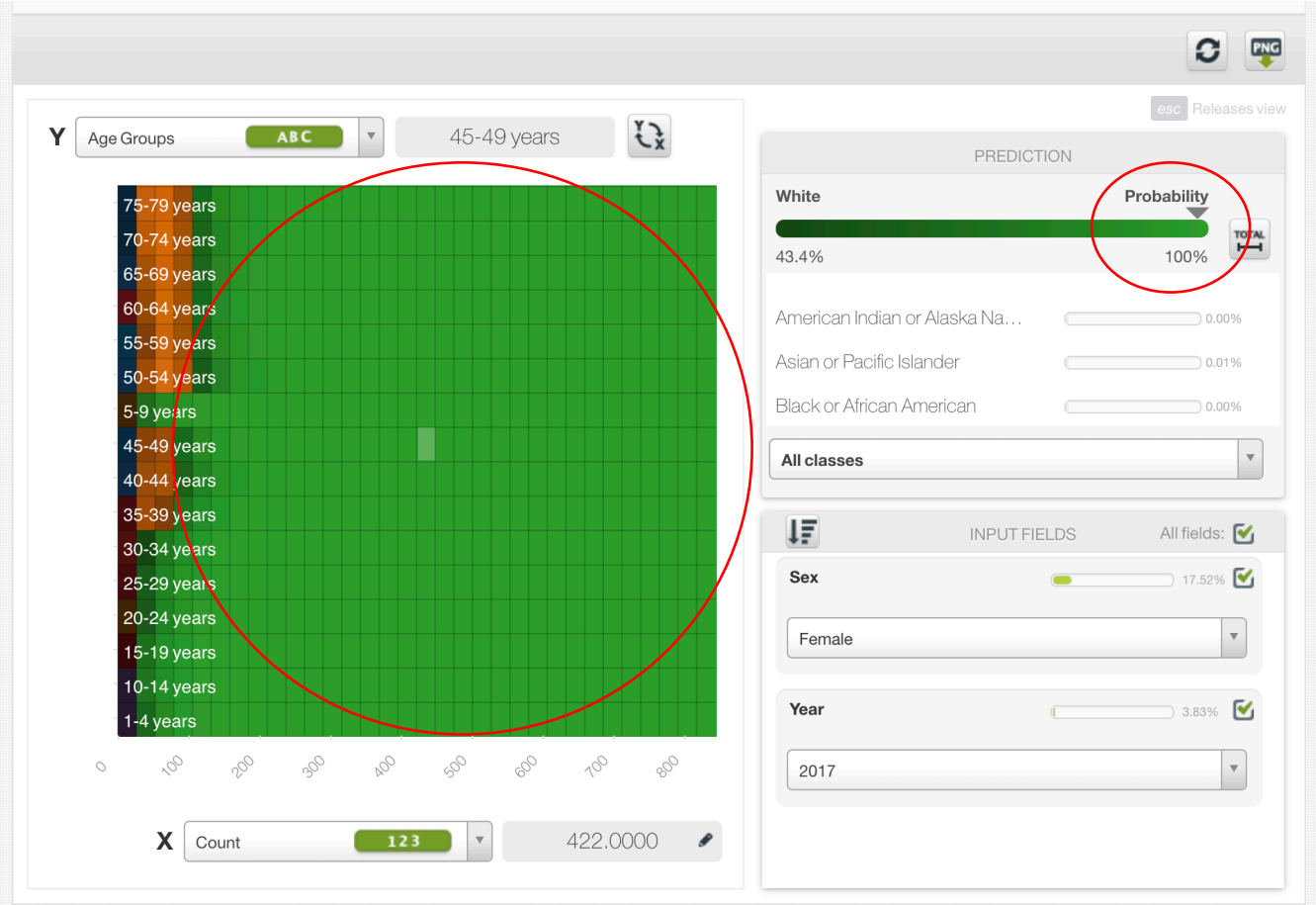

Click Here to View in BigML




Click Here to View in BigML

How does my chosen model perform?

# How does my chosen model perform?




[Click Here to View in BigML]

**Findings:**

There are many “highest points of probability” for white people in comparison to other races (White people are indicated on the Deepnet as green)

**Caveats:**

This shows the lack of representation marginalized groups have within the data set which disproportionately gives Cancer Patients of Color less care – because they are not as likely to be represented

# How does my chosen model perform? (As a whole)

**Findings:** Predictions based on as a **whole**

**Accuracy** – 28.0% came out right as a whole

**Precision** – Measures False Positives rate. About 77% of the time the model will return a false positive. About 23% of the time the model can get a true positive as we would expect.

- 77% of the time the model will state that someone does have Breast Cancer when they do not

**Recall** – Measures False Negative rate. About 73% of the time the model returns a false negative when it shouldn't have. About 28% of the time, the model was able to get a true positive as we would expect.

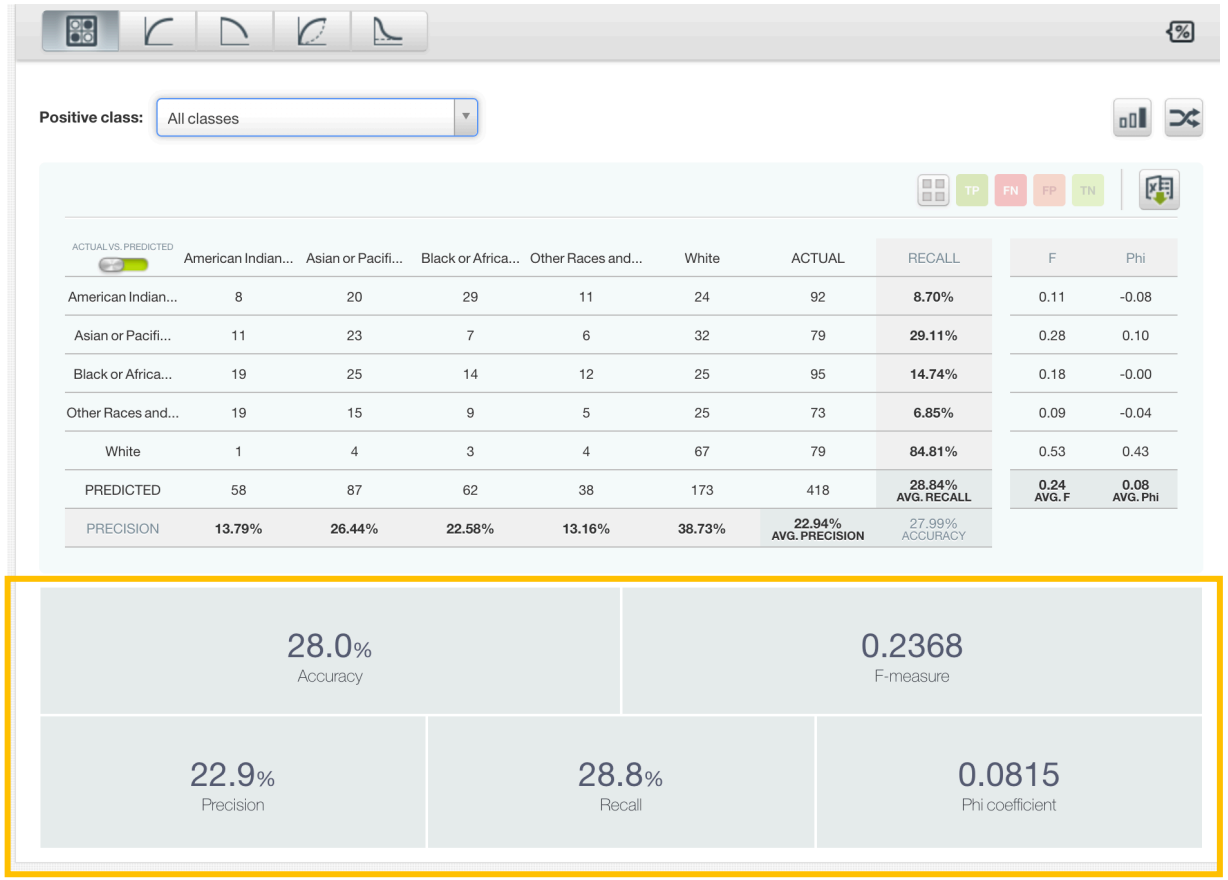
- 73% of the time the model will state that someone does not have Breast Cancer when they do

**F-measure** - .2368 is the balance point between Precision and Recall

- This model does poorly at predicting the correct number of Breast Cancer Patients across all demographics (race). This is concerning, because this is private/sensitive information that would make life-changing decisions for the patient

**Caveats:**

1. There were many "suppressed" values that could skew the data
2. "Other races and Unknown Combined" is concerning, because people were counted as "unknown"
3. Breast Cancer can be a life-or-death situation, which shows that there needs to be multiple trials of testing to prove if someone truly does have cancer, and its not only a false positive
4. May be not enough columns/features/descriptions that could help this model determine if someone has Breast Cancer or not




Click Here to View in BigML

# How does my chosen model perform? (Specified)

**Findings:** **Patients who were white had the most correct predictions**

**Accuracy** – 71.8% of the predictions came out right for white people

**Precision** – Measures False Positives. About 61% of the time the model will return a false positive. About 39% of the time the model was able to get a true positive as we would expect.

- 61% of the time the model will state that someone does have Breast Cancer when they do not

**Recall** – Measures False Negative rate. About 16% of the time the model returns a false negative when it shouldn't have. About 84.4% of the time, the model was able to get a true positive as we would expect.

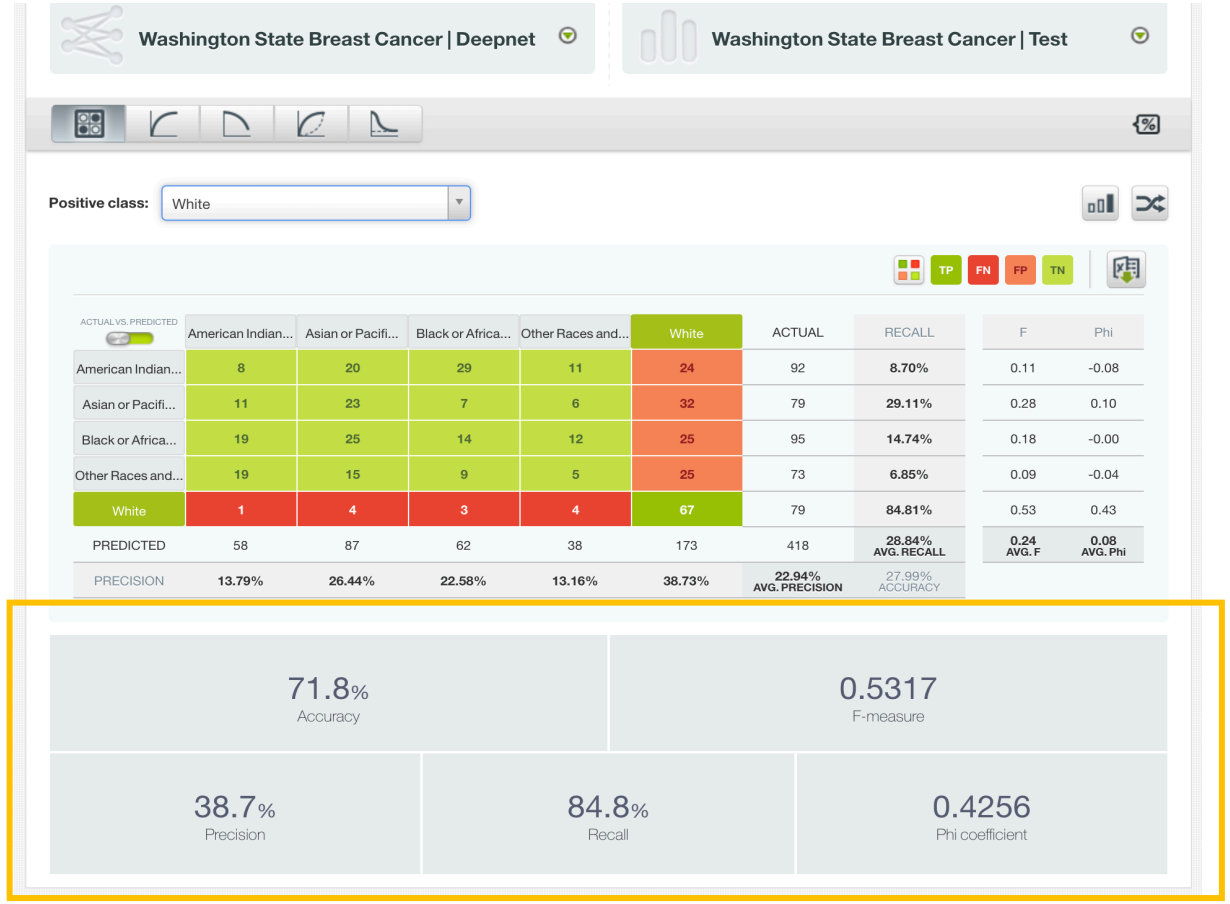
- 16% of the time the model will state that someone does not have Breast Cancer when they do

**F-measure** - .5317 is the balance between Precision and Recall

- Historically, White people had the most representation with the best quality in the health care system – which could explain why white people could be predicted better
- In this scenario, it is better to emphasis Precision, because it is better for someone to have been diagnosed with Breast Cancer when they don't – than someone who does have Breast cancer and don't receive the treatment they need

**Caveats:**

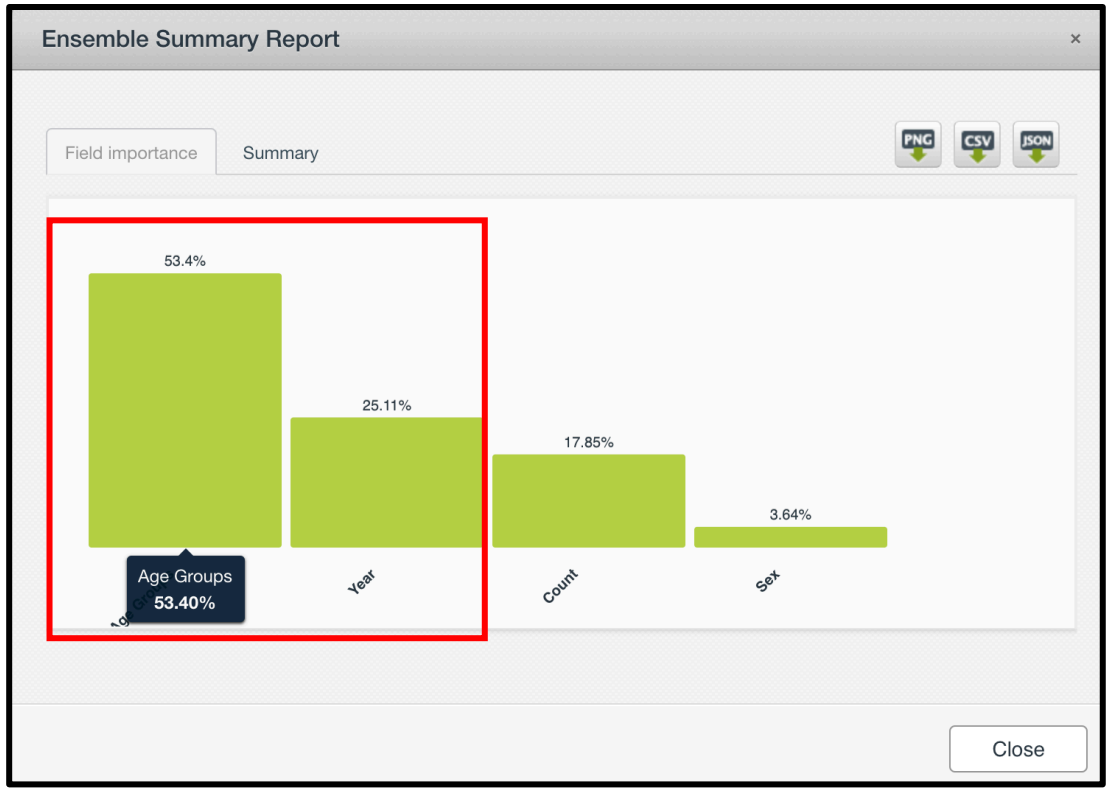
1. "Suppressed" values could have skewed the data, making this model harder to predict people of color because they are not specifically represented by a number (even if it is a low number) within the CDC data set
2. I don't have many other description columns, such as the County in Washington State or the Income levels in Washington state based on Census tracts, which could also give more accurate predictions than the limited amount of columns




Click Here to View in BigML

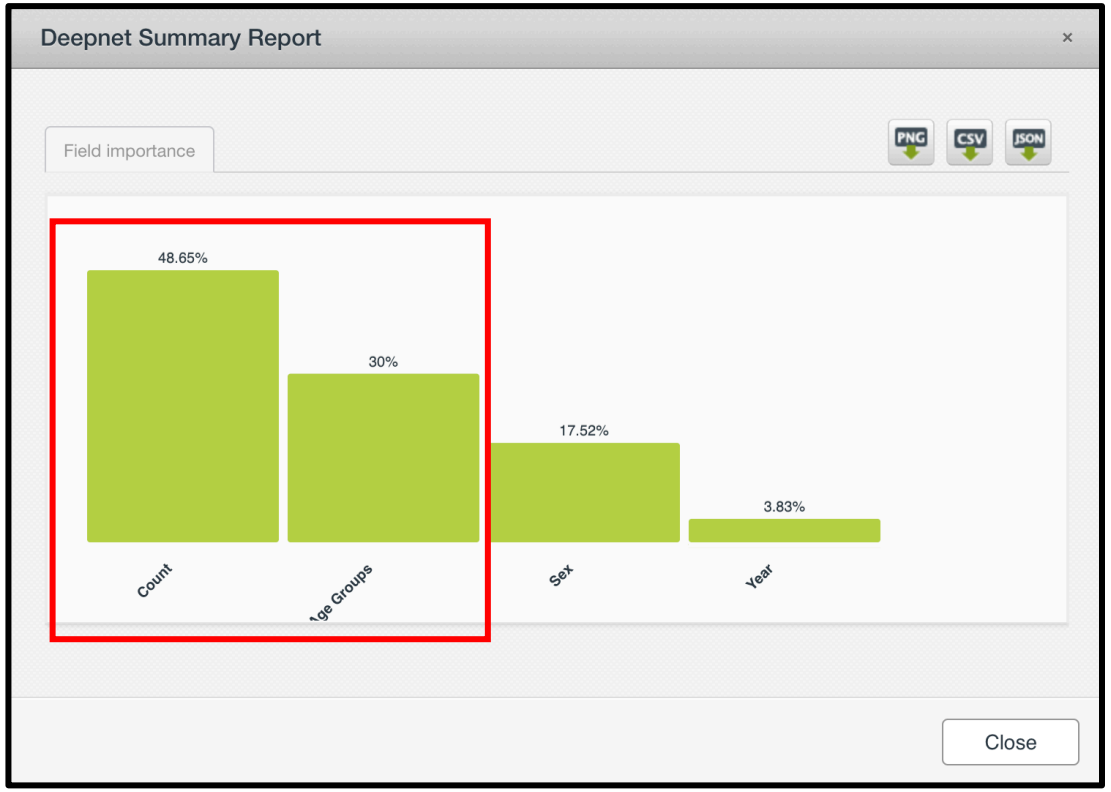
What factors are most influential in predicting the categorical variable?

# What factors are most influential in predicting the categorical variable?




[Click Here to View in BigML]

**Top Two Influential:**
1. Age groups
2. Year




[Click Here to View in BigML]

**Top Two Influential:**
1. Count
2. Age Groups

According to my scenario, what are my predictions and what should I recommend?

# According to my scenario, what are my predictions and what should I recommend?

**Before Analysis**

**Scenario Question(s):**

1. Based on age group, race and sex - what race will most likely have the most amount of reported Breast Cancer incidences?
2. Can this prediction be targeted for future years of Breast Cancer incidences?
3. Should Washington State increase funding?

**Predicted Variable (objective Variable):** Race

**Predicted Outcome:** White

**Predicted Outcome:** Out of age group, race, sex and year – White people of any age will have the most reported Breast Cancer incidences. This outcome will be able to be targeted for future years.

Scenario – After Analysis

**After Analysis**

**Scenario Answer(s):**

1. From the data set and analysis models, White people are most likely to have the most amount of reported Breast Cancer incidences
2. No because this data set only provides data up until 2017 and is not current data.
   - At the best, these prediction may be targeted for future years of Breast Cancer incidences, for white people. I would not use this data set to Predict other races other than white people because of the lack of representation (“suppressed” values).
3. Yes, Washington State should Increase Funding because there are still hundreds of people being diagnosed with Breast Cancer every year.

**Analyzed Variable (objective Variable):** Race

**Best Analysis Outcome:** White

**Analysis Outcome:** Out of age group, race, sex and year – White people of any age will have the most reported Breast Cancer incidences. This outcome should not be used to target future years of Breast Cancer Incidences. If someone still decided to use this data set to predict future incidences, this set may be able to predict future years for White people, but not accurately for people of Color.

# According to my scenario, what are my predictions and what should I recommend? Cont.

**What I learned from this Analysis:**

1. I did not expect that the BigML models would be able to analyze the details of the columns as specifically as they did, and it shows how powerful machine learning can be
   - (e.g., the variance of the interpretations of the columns can affect the performance of the wanted model)
2. A common theme throughout this entire data set was representation and since this data set was lacking in equal race representation – it can also show how the health care industry lacks equity across different demographic representations – from both patients and in the work field

**Recommendations:**

1. More (quality) columns/the better – in general health records/information is private which means the general public wouldn't be able to download csv files of people's information
2. Specific Columns/Variables – the data set had the variable "suppressed" if there was less than 16 cases which disproportionality suppressed marginalized groups from being equally represented in the data set. If I were to "help" this data set or redo this analysis, I would pick a data set that would give "all the numbers" for all rows/observations

**Overall Caveats:**

1. This data went through many systems and was (most likely) viewed by many people – which means that biases can happen at many points when this data traveled to the CDC – whether it was a bias by humans or algorithmic biases
2. Handling health records is tough – from experience this data set made me question both the integrity of health records themselves and the number of stakeholders that both contribute to data recording to the ones who face repercussions from not being represented

# Results Recap

1) Can I predict a categorical value from a data set?

1) Yes! - through both and Ensemble and a Deepnet

2) Should I use a deepnet or ensemble to make my predictions?

2) A Deepnet is better based on its performance rate shown by BigML and my specific data set

3) How does my chosen model perform?

3) My chosen model performs the best for White People when determining which demographic would most likely have Breast Cancer in Washington State. This data set would not be applicable for future predictions and/or predictions based on time, because of the lack of equal representations and the production date (2017 was the latest data)

4) What factors are most influential in predicting the categorical variable?

4) For Ensembles – Age Groups and Year
For Deepnet – Count and Age Groups