# Amazon SageMaker Autopilot Data Exploration Report

This report contains insights about the dataset you provided as input to the AutoML job. This data report was generated by **Assignment2-2** AutoML job. To check for any issues with your data and possible improvements that can be made to it, consult the sections below for guidance. You can use information about the predictive power of each feature in the **Data Sample** section and from the correlation matrix in the **Cross Column Statistics** section to help select a subset of the data that is most significant for making predictions.

**Note**: SageMaker Autopilot data reports are subject to change and updates. It is not recommended to parse the report using automated tools, as they may be impacted by such changes.

## Dataset Summary

### Dataset Properties

| Rows | Columns | Duplicate rows | Target column | Missing target values | Invalid target values | Detected problem type |
|---|---|---|---|---|---|---|
| 537 | 9 | 0.00% | Outcome | 0.00% | 0.00% | BinaryClassification |

### Detected Column Types

| | Numeric | Categorical | Text | Datetime | Sequence |
|---|---|---|---|---|---|
| **Column Count** | 8 | 0 | 0 | 0 | 0 |
| **Percentage** | 100.00% | 0.00% | 0.00% | 0.00% | 0.00% |

## Report Contents

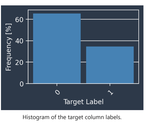


## Target Analysis

The column **Outcome** is used as the target column. See the distribution of values (labels) in the target column below:

| Number of Classes | Invalid Percentage | Missing Percentage |
|---|---|---|
| 2 | 0.00% | 0.00% |

| Target Label | Frequency Percentage | Label Count |
|---|---|---|
| 0 | 65.36% | 351 |
| 1 | 34.64% | 186 |




Histogram of the target column labels.

## Data Sample

The following table contains a random sample of **10** rows from the dataset. The top two rows provide the type and prediction power of each column. Verify the input headers correctly align with the columns of the dataset sample. If they are incorrect, update the header names of your input dataset in Amazon Simple Storage Service (Amazon S3).

| | Outc | Pregnc | BloodPr | SkinThi | Ag | Gluc | Insu | BM | DiabetesPedig |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | |
| | | | | | | | | | |

## Descriptive Stats

For each of the input features that has at least one numeric value, several descriptive statistics are computed from the data sample.

SageMaker Autopilot may treat numerical features as `Categorical` if the number of unique entries is sufficiently low. For `Numerical` features, we may apply numerical transformations such as normalization, log and quantile transforms, and binning to manage outlier values and difference in feature scales.

We found **9 of the 9** columns contained at least one numerical value. The table below shows the **9** columns which have the largest percentage of numerical values. Percentage of outliers is calculated only for columns which Autopilot detected to be of numeric type. Percentage of outliers is not calculated for the target column.

> **Suggested Action Items** - Investigate the origin of the data field. Are some values non-finite (e.g. infinity, nan)? Are they missing or is it an error in data input? - Missing and extreme values may indicate a bug in the data collection process. Verify the numerical descriptions align with expectations. For example, use domain knowledge to check that the range of values for a feature meets with expectations.

| | % of Numerical Values | Mean | Median | Min | Max |
|---|---|---|---|---|---|
| **Outcome** | 100.0% | 0.346369 | 0.0 | 0.0 | 1.0 |
| **Pregnancies** | 100.0% | 3.92551 | 3.0 | 0.0 | 17.0 |
| **Glucose** | 100.0% | 121.587 | 117.0 | 0.0 | 199.0 |
| **BloodPressure** | 100.0% | 68.9888 | 72.0 | 0.0 | 114.0 |
| **SkinThickness** | 100.0% | 20.4115 | 22.0 | 0.0 | 99.0 |
| **Insulin** | 100.0% | 77.7523 | 22.0 | 0.0 | 846.0 |
| **BMI** | 100.0% | 31.8501 | 32.0 | 0.0 | 67.1 |
| **DiabetesPedigreeFunction** | 100.0% | 0.469642 | 0.364 | 0.084 | 2.42 |
| **Age** | 100.0% | 33.2775 | 29.0 | 21.0 | 81.0 |

## Definitions

### Feature types

**Numeric:** Numeric values, either floats or integers. For example: age, income. When training a machine learning model, it is assumed that numeric values are ordered and a distance is defined between them. For example, 3 is closer to 4 than to 10 and 3 < 4 < 10.

**Categorical:** The column entries belong to a set of unique values that is usually much smaller than number of rows in the dataset. For example, a column from datasets with 100 rows with the unique values "Dog", "Cat" and "Mouse". The values could be numeric, textual, or combination of both. For example, "Horse", "House", 8, "Love" and 3.1 are all valid values and can be found in the same categorical column. When manipulating column of categorical values, a machine learning model does not assume that they are ordered or that distance function is defined on them, even if all of the values are numbers.

**Binary:** A special case of categorical column for which the cardinality of the set of unique values is 2.

**Text:** A text column that contains many non-numeric unique values, often a human readable text. In extreme cases, all the elements of the column are unique, so no two entries are the same.

**Datetime:** This column contains date and/or time information.

### Feature statistics

**Prediction power:** Prediction power of a column (feature) is a measure of how useful it is for predicting the target variable. It is measured using a stratified split into 80%/20% training and validation folds. We fit a model for each feature separately on the training fold after applying minimal feature pre-processing and measure prediction performance on the validation data. The scores are normalized to the range [0,1]. A higher prediction power score near 1 indicate that a column is more useful for predicting the target on its own. A lower score near 0 indicate that a column contains little useful information for predicting the target on their own. Although it is possible that a column that is uninformative on its own can be useful in predicting the target when used in tandem with other features, a low score usually indicates the feature is redundant. A score of 1 implies perfect predictive abilities, which often indicates an error called target leakage. The cause is typically a column present in dataset that is hard or impossible to obtain at prediction time, such as a duplicate of the target.

**Outliers:** Outliers are detected using two statistics that are robust to outliers: median and robust standard deviation (RSTD). RSTD is derived by clipping the feature values to the range [5 percentile, 95 percentile] and calculating the standard deviation of the clipped vector. All values larger than median + 5 * *RSTD or smaller than median* - 5 RSTD are considered to be outliers.

**Skew:** Skew measures the symmetry of the distribution and is defined as the third moment of the distribution divided by the third power of the standard deviation. The skewness of the normal distribution or any other symmetric distribution is zero. Positive values imply that the right tail of the distribution is longer than the left tail. Negative values imply that the left tail of the distribution is longer than the right tail. As a thumb rule, a distribution is considered skewed when the absolute value of the skew is larger than 3.

**Kurtosis:** Pearson's kurtosis measures the heaviness of the tail of the distribution and is defined as the fourth moment of the distribution divided by the fourth power of the standard deviation. The kurtosis of the normal distribution is 3. Thus, kurtosis values lower than 3 imply that the distribution is more concentrated around the mean and the tails are lighter than the tails of the normal distribution. Kurtosis values higher than 3 imply heavier tails than the normal distribution or that the data contains outliers.

**Missing Values:** Empty strings and strings composed of only white spaces are considered missing.

**Valid values:**

- **Numeric features / regression target:** All values that could be casted to finite floats are valid. Missing values are not valid.
- **Categorical / binary / text features / classification target:** All values that are not missing are valid.
- **Datetime features:** All values that could be casted to datetime object are valid. Missing values are not valid.

**Invalid values:** values that are either missing or that could not be casted to the desired type. See the definition of valid values for more information