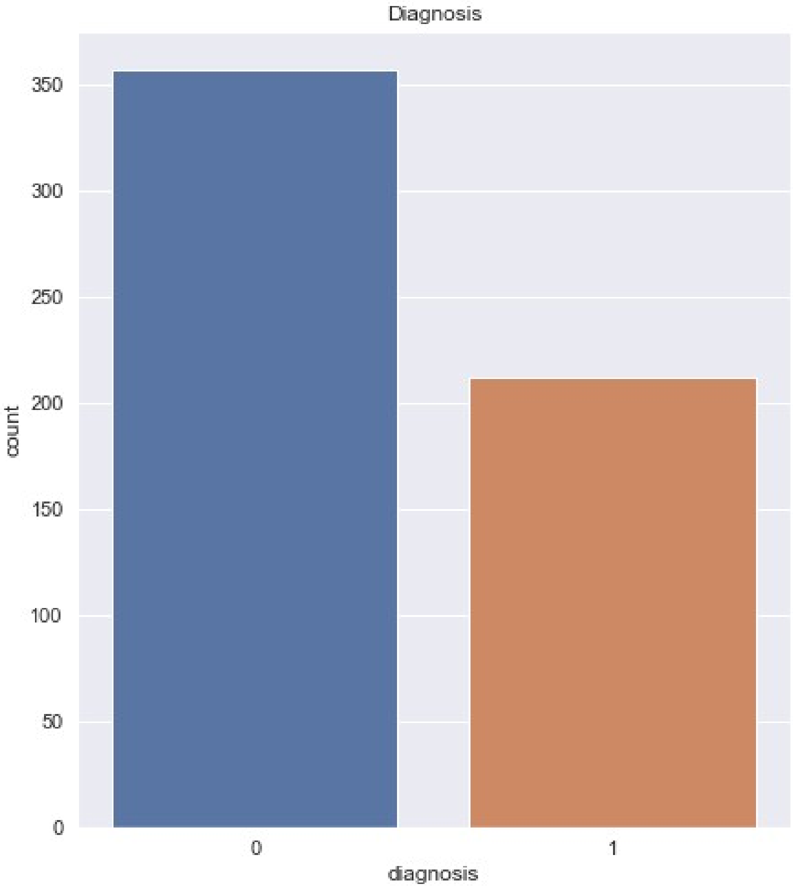
Dev Joshi

Final Report : Benign or Malignant Cancer Prediction

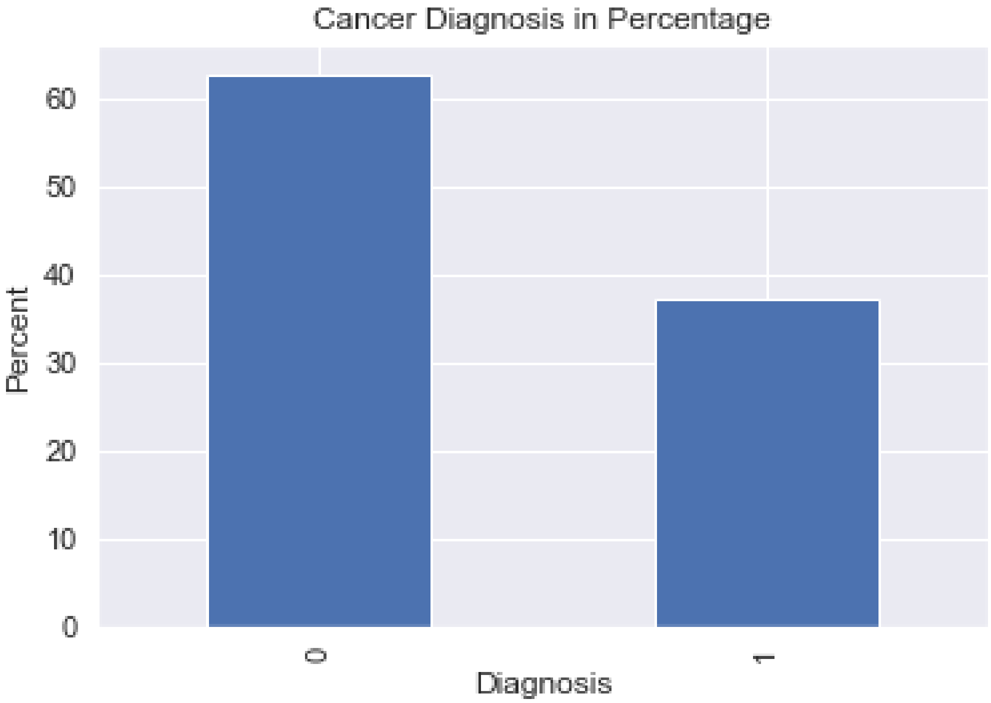
Problem Statement

The dataset consisting of digitized imaging of fine needle aspirate (FNA) of a breast tumor cell mass can be used to distinguish between a benign and malignant tumor to aid in clinical diagnosis. Each cell nucleus has ten real-valued features which describe characteristics of the cell nuclei present in the image. The project goal is to deploy machine learning algorithms to accurately distinguish between a benign and malignant tumor to aid in clinical diagnosis. The original dataset is available at the https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+Diagnostic .

Data Wrangling: We began by downloading the dataset from the above link. The dataset consisted of 569 rows and 33 columns. We removed two columns – 'id' and 'Unnamed: 32' as they weren't essential for the analysis. We explored the data and looked at various columns of the dataframe. We made the target variable (column) 'diagnosis ' as numeric by assigning '1' as Malignant and '0' as benign. We plotted this column to observe that the number of the benign cases are 357 and the number of malignant cases are 212.



The same plot in terms of percentage yielded


Cancer Diagnosis in Percentage
Percent
Diagnosis