Breast Cancer Prediction Analysis

Joseph Gyorda, Kevin Rouse, Anton Hung, Digvijay Yadav, Bofan Chen
(Team Dunphy 101)

Overview

1. Introduction and Motivations
2. Exploratory Data Analysis (EDA)
   a. Principal Component Analysis (PCA)
3. Feature Selection with LASSO
4. Prediction of Malignant Tumors
   a. K-Means Clustering
   b. Linear Discriminant Analysis (LDA)
   c. Decision Tree/Random Forest
   d. Neural Network
5. Discussion of Results/Conclusion

Introduction and Motivations

Motivations

- Breast cancer is a serious and prevalent disease
  - 1 out of 8 women develop breast cancer in the US[1]
  - 1 out of 3 new cancer cases in women are breast cancer[1]
- Most biopsied tumors are benign (non-cancerous)—only 20% are malignant (cancerous)[2]
- Accordingly, identifying characteristics of malignant tumors is important to avoid mistreatment of cancer (e.g., unnecessary chemotherapy)
- Highly accurate classification models may support healthcare decision making and tumor identification

[1] https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html

[2] https://cancer.stonybrookmedicine.edu/breast-cancer-team/patients/bse/breastlumps

Introduction to Dataset

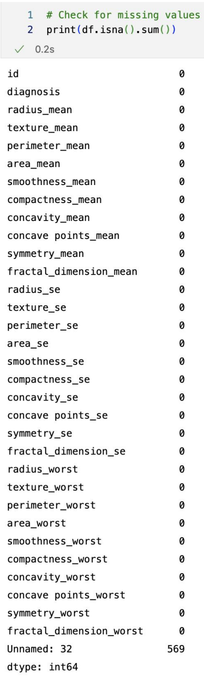
- Wisconsin Tumor Dataset with tumor information collected from **N=569** individuals[3]
- "Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image."
- Outcome variable was tumor diagnosis: benign (coded as 0, **N=357**, ~63%) malignant (coded as 1, **N=212,** ~37%)
- **Variables:**
  1. **radius** (mean of distances from center to points on the perimeter)
  2. **texture** (standard deviation of gray-scale values)
  3. **perimeter**
  4. **area**
  5. **smoothness** (local variation in radius lengths)
  6. **compactness** (perimeter^2 / area - 1.0)
  7. **concavity** (severity of concave portions of the contour)
  8. **concave points** (number of concave portions of the contour)
  9. **symmetry**
  10. **fractal dimension** ("coastline approximation" - 1)
- Mean, standard error, and "worst" value for each feature calculated, totalling 10*3=30 variables

[3] https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data

Exploratory Data Analysis

Exploratory Data Analysis

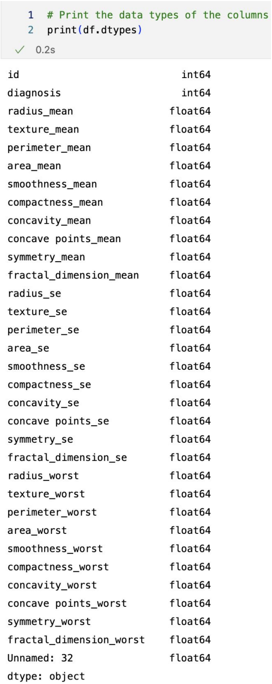
- Looked at missing values, data types, number of unique values in each column
- Checked the distribution of the variables
- Encode target variable (M = malignant, B = benign) to 1, 0



```
# Check for missing values
print(df.isna().sum())
```

✓ 0.2s

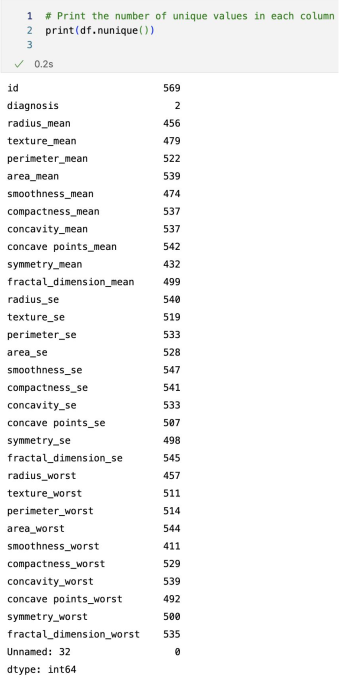
```
id                          0
diagnosis                   0
radius_mean                 0
texture_mean                0
perimeter_mean              0
area_mean                   0
smoothness_mean             0
compactness_mean            0
concavity_mean              0
concave points_mean         0
symmetry_mean               0
fractal_dimension_mean      0
radius_se                   0
texture_se                  0
perimeter_se                0
area_se                     0
smoothness_se               0
compactness_se              0
concavity_se                0
concave points_se           0
symmetry_se                 0
fractal_dimension_se        0
radius_worst                0
texture_worst               0
perimeter_worst             0
area_worst                  0
smoothness_worst            0
compactness_worst           0
concavity_worst             0
concave points_worst        0
symmetry_worst              0
fractal_dimension_worst     0
Unnamed: 32               569
dtype: int64
```



```
# Print the data types of the columns
print(df.dtypes)
```

✓ 0.2s

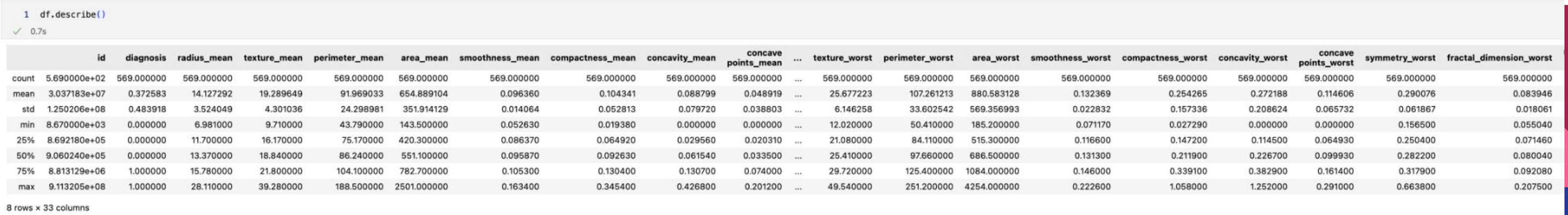
```
id                           int64
diagnosis                    int64
radius_mean                float64
texture_mean               float64
perimeter_mean             float64
area_mean                  float64
smoothness_mean            float64
compactness_mean           float64
concavity_mean             float64
concave points_mean        float64
symmetry_mean              float64
fractal_dimension_mean     float64
radius_se                  float64
texture_se                 float64
perimeter_se               float64
area_se                    float64
smoothness_se              float64
compactness_se             float64
concavity_se               float64
concave points_se          float64
symmetry_se                float64
fractal_dimension_se       float64
radius_worst               float64
texture_worst              float64
perimeter_worst            float64
area_worst                 float64
smoothness_worst           float64
compactness_worst          float64
concavity_worst            float64
concave points_worst       float64
symmetry_worst             float64
fractal_dimension_worst    float64
Unnamed: 32                float64
dtype: object
```



```
# Print the number of unique values in each column
print(df.nunique())

```

✓ 0.2s

```
id                         569
diagnosis                    2
radius_mean                456
texture_mean               479
perimeter_mean             522
area_mean                  539
smoothness_mean            474
compactness_mean           537
concavity_mean             537
concave points_mean        542
symmetry_mean              432
fractal_dimension_mean     499
radius_se                  540
texture_se                 519
perimeter_se               533
area_se                    528
smoothness_se              547
compactness_se             541
concavity_se               533
concave points_se          507
symmetry_se                498
fractal_dimension_se       545
radius_worst               457
texture_worst              511
perimeter_worst            514
area_worst                 544
smoothness_worst           411
compactness_worst          529
concavity_worst            539
concave points_worst       492
symmetry_worst             500
fractal_dimension_worst    535
Unnamed: 32                  0
dtype: int64
```



```
df.describe()
```

✓ 0.7s

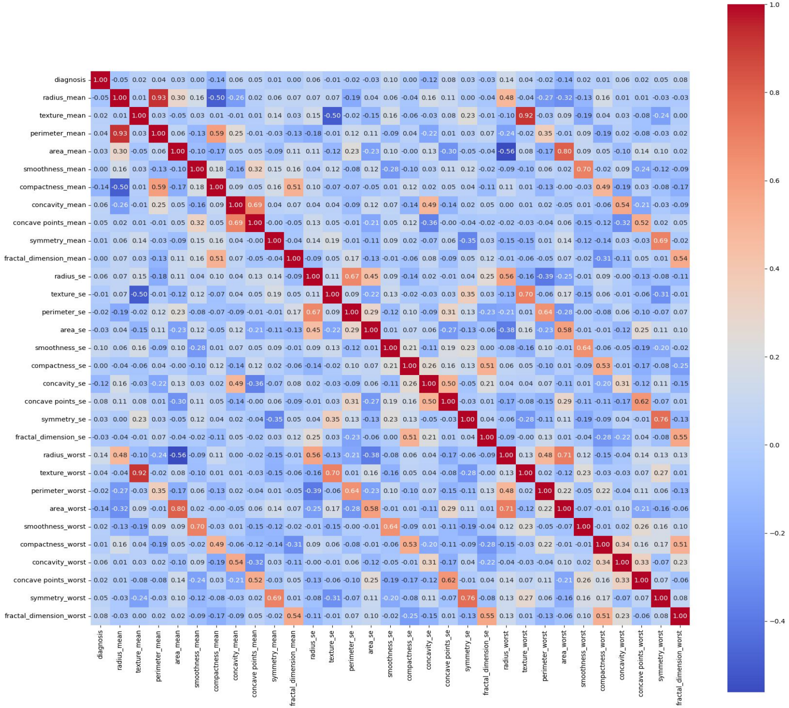
	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...	texture_worst	perimeter_worst	area_worst	smoothness_worst	compactness_worst	concavity_worst	concave points_worst	symmetry_worst	fractal_dimension_worst
count	5.690000e+02	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	...	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000
mean	3.037183e+07	0.372583	14.127292	19.289649	91.969033	654.889104	0.096360	0.104341	0.088799	0.048919	...	25.677223	107.261213	880.583128	0.132369	0.254265	0.272188	0.114606	0.290076	0.083946
std	1.250206e+08	0.483918	3.524049	4.301036	24.298981	351.914129	0.014064	0.052813	0.079720	0.038803	...	6.146258	33.602542	569.356993	0.022832	0.157336	0.208624	0.065732	0.061867	0.018061
min	8.670000e+03	0.000000	6.981000	9.710000	43.790000	143.500000	0.052630	0.019380	0.000000	0.000000	...	12.020000	50.410000	185.200000	0.071170	0.027290	0.000000	0.000000	0.156500	0.055040
25%	8.692180e+05	0.000000	11.700000	16.170000	75.170000	420.300000	0.086370	0.064920	0.029560	0.020310	...	21.080000	84.110000	515.300000	0.116600	0.147200	0.114500	0.064930	0.250400	0.071460
50%	9.060240e+05	0.000000	13.370000	18.840000	86.240000	551.100000	0.095870	0.092630	0.061540	0.033500	...	25.410000	97.660000	686.500000	0.131300	0.211900	0.226700	0.099930	0.282200	0.080040
75%	8.813129e+06	1.000000	15.780000	21.800000	104.100000	782.700000	0.105300	0.130400	0.130700	0.074000	...	29.720000	125.400000	1084.000000	0.146000	0.339100	0.382900	0.161400	0.317900	0.092080
max	9.113205e+08	1.000000	28.110000	39.280000	188.500000	2501.000000	0.163400	0.345400	0.426800	0.201200	...	49.540000	251.200000	4254.000000	0.222600	1.058000	1.252000	0.291000	0.663800	0.207500

8 rows × 33 columns

Exploratory Data Analysis

- Partial correlation plot: measures the strength of the linear relationship between two variables while controlling for the effects of one or more additional variables
- Help remove the influence of confounding factors and produce a more accurate prediction model
- The variables that are highly correlated with the diagnosis may have low correlation values in the partial correlation plot due to the effects of other variables

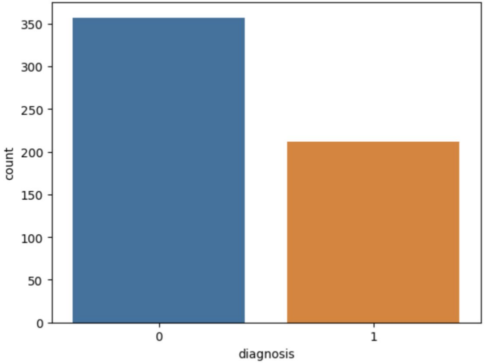
```
Top 21 Variables Correlated with Diagnosis:
 concavity_worst            0.95
symmetry_se                0.94
fractal_dimension_worst    0.94
compactness_worst          0.93
compactness_se             0.93
concavity_se               0.93
concavity_mean             0.92
smoothness_worst           0.92
symmetry_worst             0.90
fractal_dimension_se       0.90
perimeter_se               0.88
concave points_worst       0.87
compactness_mean           0.86
perimeter_worst            0.85
radius_se                  0.85
radius_worst               0.85
concave points_mean        0.84
perimeter_mean             0.82
symmetry_mean              0.82
concave points_se          0.81
fractal_dimension_mean     0.80
Name: diagnosis, dtype: object
```



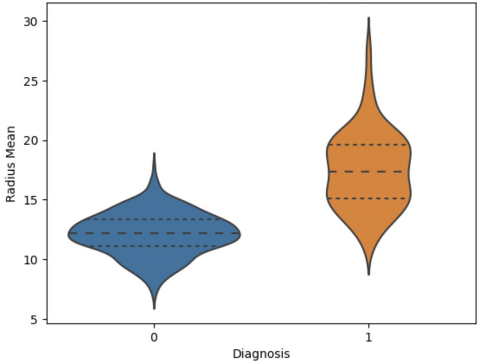
	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean	fractal_dimension_mean	radius_se	texture_se	perimeter_se	area_se	smoothness_se	compactness_se	concavity_se	concave points_se	symmetry_se	fractal_dimension_se	radius_worst	texture_worst	perimeter_worst	area_worst	smoothness_worst	compactness_worst	concavity_worst	concave points_worst	symmetry_worst	fractal_dimension_worst
diagnosis	1.00	-0.05	0.02	0.04	0.03	0.00	-0.14	0.06	0.05	0.01	0.00	0.06	-0.01	-0.02	-0.03	0.10	0.00	-0.12	0.08	0.03	-0.03	0.14	0.04	-0.02	-0.14	0.02	0.01	0.06	0.02	0.05	0.08
radius_mean	-0.05	1.00	0.01	0.93	0.30	0.16	-0.50	-0.26	0.02	0.06	0.07	0.07	0.07	-0.19	0.04	0.06	-0.04	0.16	0.11	0.00	-0.04	0.48	-0.04	-0.27	-0.32	-0.13	0.16	0.01	0.01	-0.03	-0.03
texture_mean	0.02	0.01	1.00	0.03	-0.05	0.03	0.01	-0.01	0.01	0.14	0.03	0.15	-0.50	-0.02	-0.15	0.16	-0.06	-0.03	0.08	0.23	-0.01	-0.10	0.92	-0.03	0.09	-0.19	0.04	0.03	-0.08	-0.24	0.00
perimeter_mean	0.04	0.93	0.03	1.00	0.06	-0.13	0.59	0.25	-0.01	-0.03	-0.13	-0.18	-0.01	0.12	0.11	-0.09	0.04	-0.22	0.01	0.03	0.07	-0.24	-0.02	0.35	-0.01	0.09	-0.19	0.02	-0.08	-0.03	0.02
area_mean	0.03	0.30	-0.05	0.06	1.00	-0.10	-0.17	0.05	0.05	-0.09	0.11	0.11	-0.12	0.23	-0.23	0.10	-0.00	0.13	-0.30	-0.05	-0.04	-0.56	0.08	-0.17	0.80	0.09	0.05	-0.10	0.14	0.10	0.02
smoothness_mean	0.00	0.16	0.03	-0.13	-0.10	1.00	0.18	-0.16	0.32	0.15	0.16	0.04	0.12	-0.08	0.12	-0.28	-0.10	0.03	0.11	0.12	-0.02	-0.09	-0.10	0.06	0.02	0.70	-0.02	0.09	-0.24	-0.12	-0.09
compactness_mean	-0.14	-0.50	0.01	0.59	-0.17	0.18	1.00	0.09	0.05	0.16	0.51	0.10	-0.07	-0.07	-0.05	0.01	0.12	0.02	0.05	0.04	-0.11	0.11	0.01	-0.13	-0.00	-0.03	0.49	-0.19	0.03	-0.08	-0.17
concavity_mean	0.06	-0.26	-0.01	0.25	0.05	-0.16	0.09	1.00	0.69	0.04	0.07	0.04	0.04	-0.09	0.12	0.07	-0.14	0.49	-0.14	0.02	0.05	0.00	0.01	0.02	-0.05	0.01	-0.06	0.54	-0.21	-0.03	-0.09
concave points_mean	0.05	0.02	0.01	-0.01	0.05	0.32	0.05	0.69	1.00	-0.00	-0.05	0.13	0.05	-0.01	-0.21	0.05	0.12	-0.36	-0.00	-0.04	-0.02	-0.02	-0.03	-0.04	0.06	-0.15	-0.12	-0.32	0.52	0.02	0.05
symmetry_mean	0.01	0.06	0.14	-0.03	-0.09	0.15	0.16	0.04	-0.00	1.00	-0.04	0.14	0.19	-0.01	-0.11	0.09	0.02	-0.07	0.06	-0.35	0.03	-0.15	-0.15	0.01	0.14	-0.12	-0.14	0.03	-0.03	0.69	-0.02
fractal_dimension_mean	0.00	0.07	0.03	-0.13	0.11	0.16	0.51	0.07	-0.05	-0.04	1.00	-0.09	0.05	0.17	-0.13	-0.01	-0.06	0.08	-0.09	0.05	0.12	-0.01	-0.06	-0.05	0.07	-0.02	-0.31	-0.11	0.05	0.01	0.54
radius_se	0.06	0.07	0.15	-0.18	0.11	0.04	0.10	0.04	0.13	0.14	-0.09	1.00	0.11	0.67	0.45	0.09	-0.14	0.02	-0.01	0.04	0.25	0.56	-0.16	-0.39	-0.25	-0.01	0.09	-0.00	-0.13	-0.08	-0.11
texture_se	-0.01	0.07	-0.50	-0.01	-0.12	0.12	-0.07	0.04	0.05	0.19	0.05	0.11	1.00	0.09	-0.22	0.13	-0.02	-0.03	0.03	0.35	0.03	-0.13	0.70	-0.06	0.17	-0.15	0.06	-0.01	-0.06	-0.31	-0.01
perimeter_se	-0.02	-0.19	-0.02	0.12	0.23	-0.08	-0.07	-0.09	-0.01	-0.01	0.17	0.67	0.09	1.00	0.29	-0.12	0.10	-0.09	0.31	0.13	-0.23	-0.21	0.01	0.64	-0.28	-0.00	-0.08	0.06	-0.10	-0.07	0.07
area_se	-0.03	0.04	-0.15	0.11	-0.23	0.12	-0.05	0.12	-0.21	-0.11	-0.13	0.45	-0.22	0.29	1.00	0.01	0.07	0.06	-0.27	-0.13	-0.06	-0.38	0.16	-0.23	0.58	-0.01	-0.01	-0.12	0.25	0.11	0.10
smoothness_se	0.10	0.06	0.16	-0.09	0.10	-0.28	0.01	0.07	0.05	0.09	-0.01	0.09	0.13	-0.12	0.01	1.00	0.21	-0.11	0.19	0.23	0.00	-0.08	-0.16	0.10	-0.01	0.64	-0.06	-0.05	-0.19	-0.20	-0.02
compactness_se	0.00	-0.04	-0.06	0.04	-0.00	-0.10	0.12	-0.14	0.12	0.02	-0.06	-0.14	-0.02	0.10	0.07	0.21	1.00	0.26	0.16	0.13	0.51	0.06	0.05	-0.10	0.01	-0.09	0.53	-0.01	-0.17	-0.08	-0.25
concavity_se	-0.12	0.16	-0.03	-0.22	0.13	0.03	0.02	0.49	-0.36	-0.07	0.08	0.02	-0.03	-0.09	0.06	-0.11	0.26	1.00	0.50	-0.05	0.21	0.04	0.04	0.07	-0.11	0.01	-0.20	0.31	-0.12	0.11	-0.15
concave points_se	0.08	0.11	0.08	0.01	-0.30	0.11	0.05	-0.14	-0.00	0.06	-0.09	-0.01	0.03	0.31	-0.27	0.19	0.16	0.50	1.00	-0.03	0.01	-0.17	-0.08	-0.15	0.29	-0.11	-0.11	-0.17	0.62	-0.07	0.01
symmetry_se	0.03	0.00	0.23	0.03	-0.05	0.12	0.04	0.02	-0.04	-0.35	0.05	0.04	0.35	0.13	-0.13	0.23	0.13	-0.05	-0.03	1.00	0.04	-0.06	-0.28	-0.11	0.11	-0.19	-0.09	0.04	-0.01	0.76	-0.13
fractal_dimension_se	-0.03	-0.04	-0.01	0.07	-0.04	-0.02	-0.11	0.05	-0.02	0.03	0.12	0.25	0.03	-0.23	-0.06	0.00	0.51	0.21	0.01	0.04	1.00	-0.09	-0.00	0.13	0.01	-0.04	-0.28	-0.22	0.04	-0.08	0.55
radius_worst	0.14	0.48	-0.10	-0.24	-0.56	-0.09	0.11	0.00	-0.02	-0.15	-0.01	0.56	-0.13	-0.21	-0.38	-0.08	0.06	0.04	-0.17	-0.06	-0.09	1.00	0.13	0.48	0.71	0.12	-0.15	-0.04	0.14	0.13	0.13
texture_worst	0.04	-0.04	0.92	-0.02	0.08	-0.10	0.01	0.01	-0.03	-0.15	-0.06	-0.16	0.70	0.01	0.16	-0.16	0.05	0.04	-0.08	-0.28	-0.00	0.13	1.00	0.02	-0.12	0.23	-0.03	-0.03	0.07	0.27	0.01
perimeter_worst	-0.02	-0.27	-0.03	0.35	-0.17	0.06	-0.13	0.02	-0.04	0.01	-0.05	-0.39	-0.06	0.64	-0.23	0.10	-0.10	0.07	-0.15	-0.11	0.13	0.48	0.02	1.00	0.22	-0.05	0.22	-0.04	0.11	0.06	-0.13
area_worst	-0.14	-0.32	0.09	-0.01	0.80	0.02	-0.00	-0.05	0.06	0.14	0.07	-0.25	0.17	-0.28	0.58	-0.01	0.01	-0.11	0.29	0.11	0.01	0.71	-0.12	0.22	1.00	-0.07	-0.01	0.10	-0.21	-0.16	-0.06
smoothness_worst	0.02	-0.13	-0.19	0.09	0.09	0.70	-0.03	0.01	-0.15	-0.12	-0.02	-0.01	-0.15	-0.00	-0.01	0.64	-0.09	0.01	-0.11	-0.19	-0.04	0.12	0.23	-0.05	-0.07	1.00	-0.01	0.02	0.26	0.16	0.10
compactness_worst	0.01	0.16	0.04	-0.19	0.05	-0.02	0.49	-0.06	-0.12	-0.14	-0.31	0.09	0.06	-0.08	-0.01	-0.06	0.53	-0.20	-0.11	-0.09	-0.28	-0.15	-0.03	0.22	-0.01	-0.01	1.00	0.34	0.16	0.17	0.51
concavity_worst	0.06	0.01	0.03	0.02	-0.10	0.09	-0.19	0.54	-0.32	0.03	-0.11	-0.00	-0.01	0.06	-0.12	-0.05	-0.01	0.31	-0.17	0.04	-0.22	-0.04	-0.03	-0.04	0.10	0.02	0.34	1.00	0.33	-0.07	0.23
concave points_worst	0.02	0.01	-0.08	-0.08	0.14	-0.24	0.03	-0.21	0.52	-0.03	0.05	-0.13	-0.06	-0.10	0.25	-0.19	-0.17	-0.12	0.62	-0.01	0.04	0.14	0.07	0.11	-0.21	0.26	0.16	0.33	1.00	0.07	-0.06
symmetry_worst	0.05	-0.03	-0.24	-0.03	0.10	-0.12	-0.08	-0.03	0.02	0.69	0.01	-0.08	-0.31	-0.07	0.11	-0.20	-0.08	0.11	-0.07	0.76	-0.08	0.13	0.27	0.06	-0.16	0.16	0.17	-0.07	0.07	1.00	0.08
fractal_dimension_worst	0.08	-0.03	0.00	0.02	0.02	-0.09	-0.17	-0.09	0.05	-0.02	0.54	-0.11	-0.01	0.07	0.10	-0.02	-0.25	-0.15	0.01	-0.13	0.55	0.13	0.01	-0.13	-0.06	0.10	0.51	0.23	-0.06	0.08	1.00

Exploratory Data Analysis

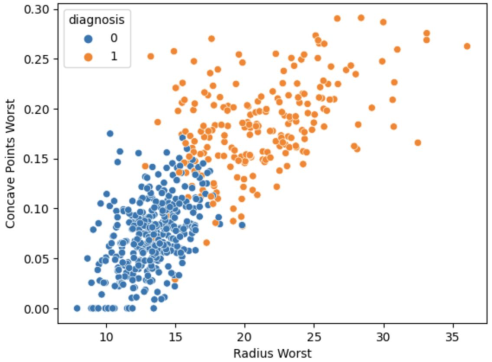
- Balanced dataset
- Clear separation between the diagnosis groups in the scatter plot of concave points_worst vs radius_worst
- The mean radius for malignant tumors tends to be larger than that for benign tumors in the violin plot of radius_mean by diagnosis



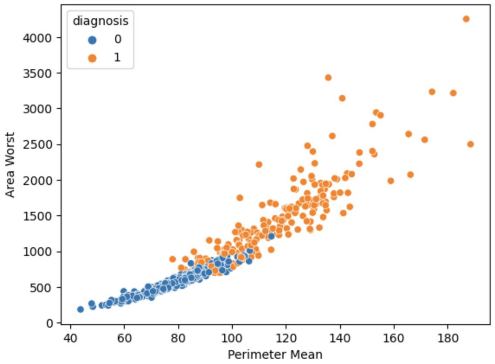
Violin Plot of Radius Mean by Diagnosis



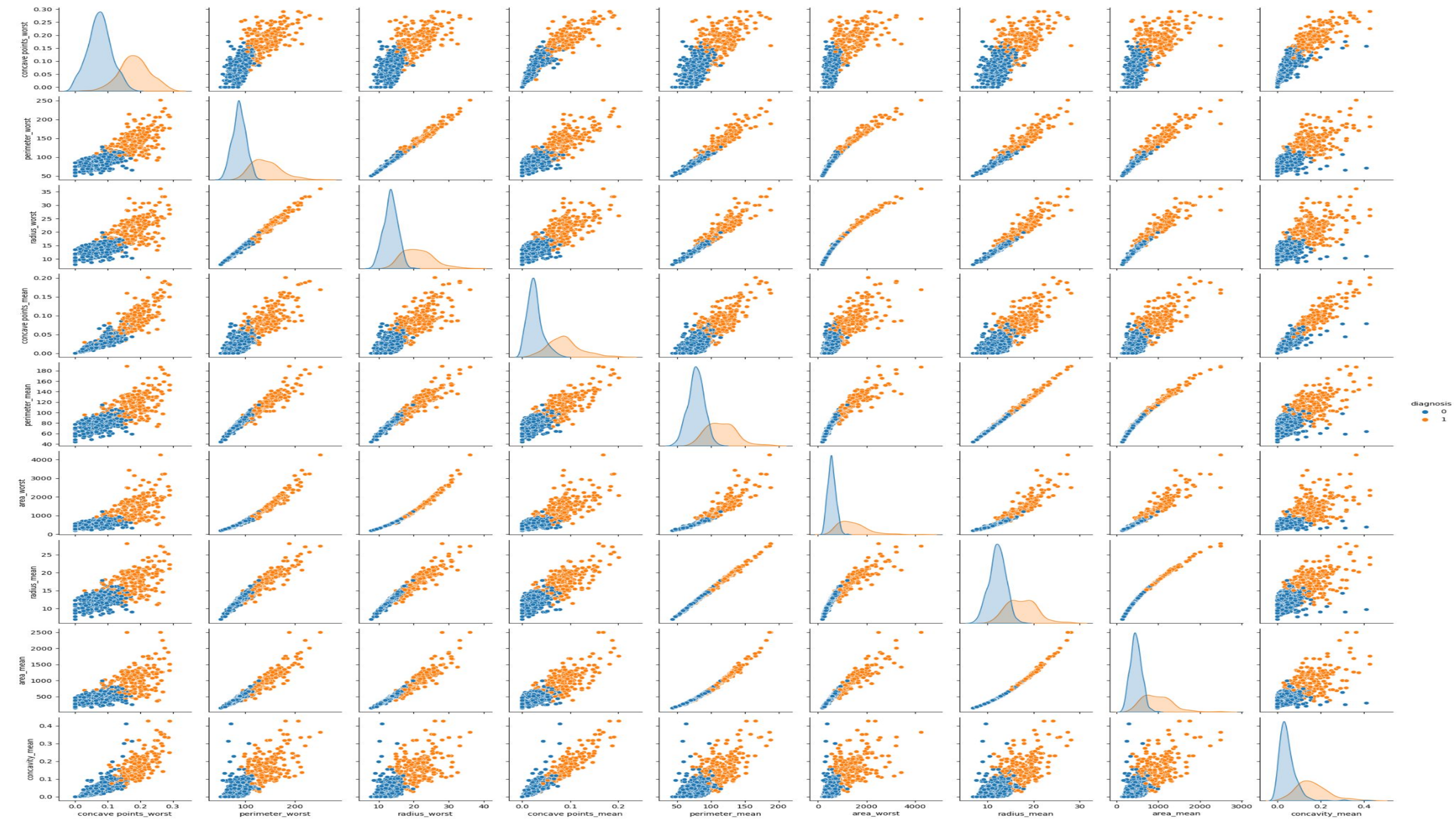
Scatter Plot of Concave Points Worst vs Radius Worst



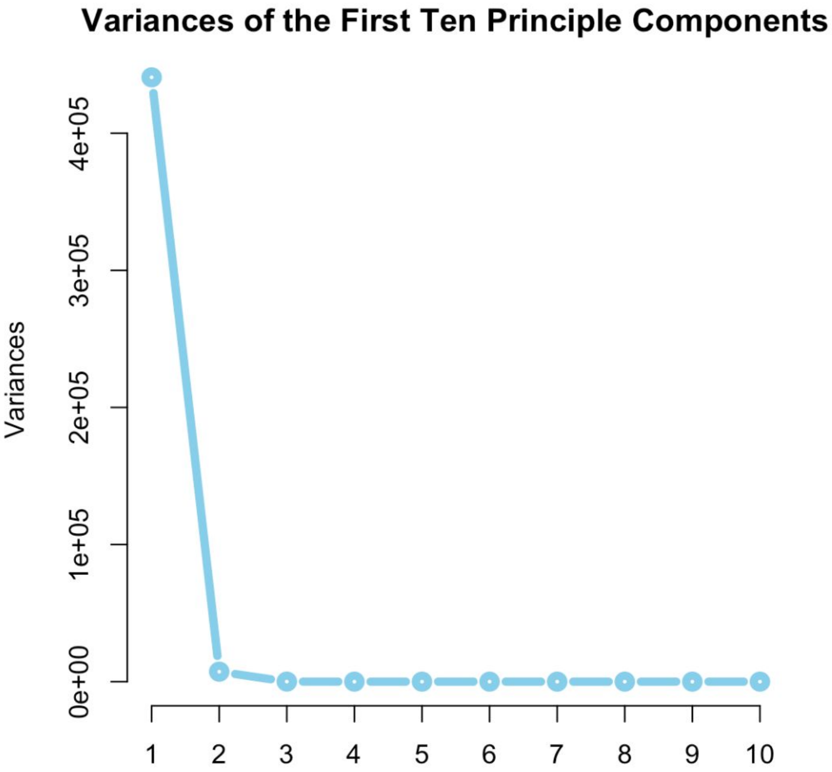
Scatter Plot of Area Worst vs Perimeter Mean

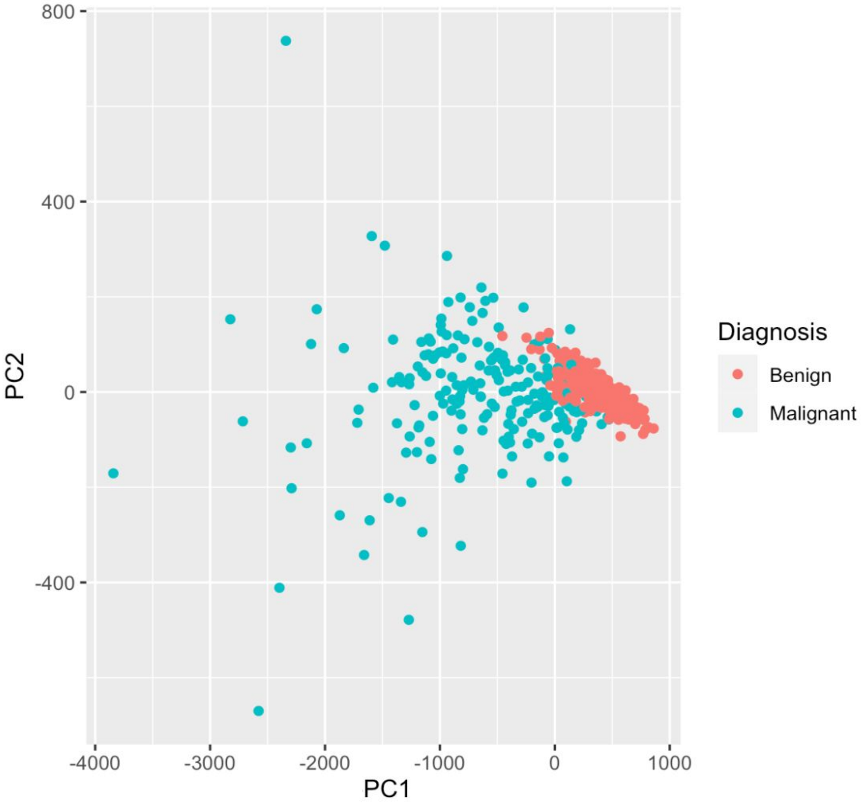


Exploratory Data Analysis



Principal Component Analysis (PCA)

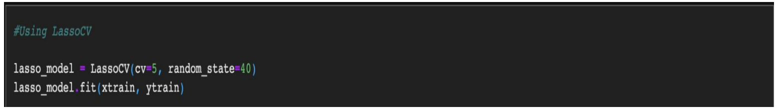




Feature Selection: LASSO

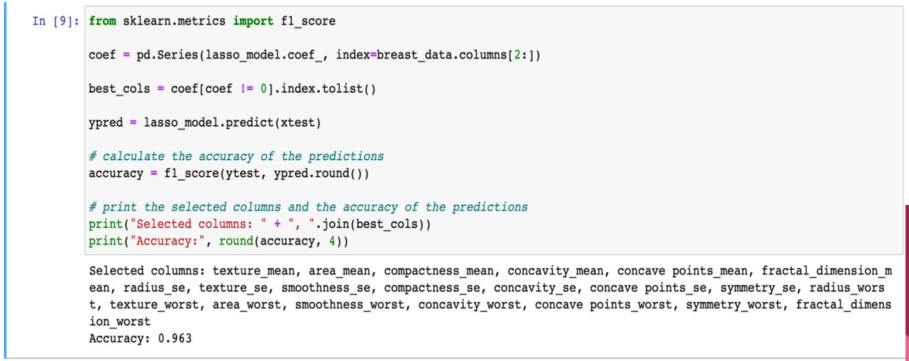
LASSO Regression with Cross Validation

- We scaled the dataset using MinMax Scaling algorithm. Ensuring that all the features are on similar scale, thereby preventing any bias in the model. Since some machine learning algorithms tend to be impacted by numerical instability, using this transformation we eliminate those issues.
- Scaled data was passed into Lasso regression model with 5 fold cross validation to avoid data leakage, and overfitting.
- From 30 variables, 21 variables were selected by the regression model, with an Accuracy of 0.963.



```
#Using LassoCV

lasso_model = LassoCV(cv=5, random_state=40)
lasso_model.fit(xtrain, ytrain)
```

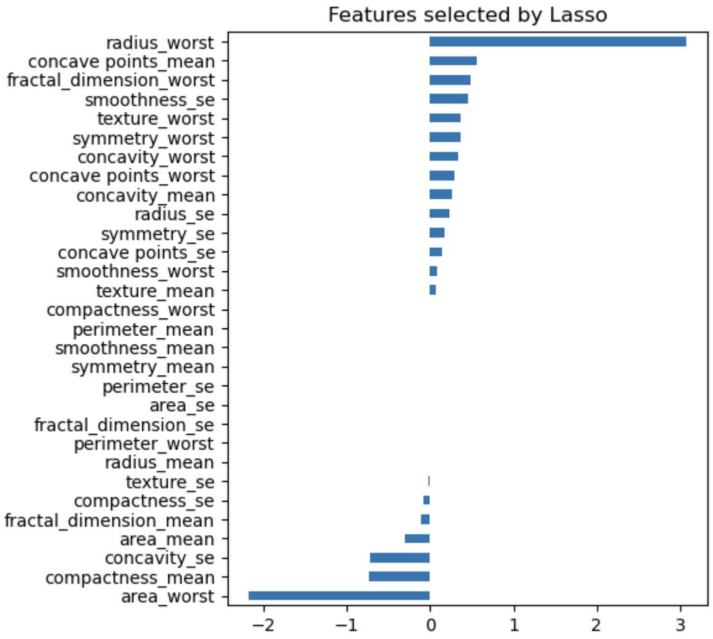
$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$



```
In [9]: from sklearn.metrics import f1_score

        coef = pd.Series(lasso_model.coef_, index=breast_data.columns[2:])

        best_cols = coef[coef != 0].index.tolist()

        ypred = lasso_model.predict(xtest)

        # calculate the accuracy of the predictions
        accuracy = f1_score(ytest, ypred.round())

        # print the selected columns and the accuracy of the predictions
        print("Selected columns: " + ", ".join(best_cols))
        print("Accuracy:", round(accuracy, 4))

        Selected columns: texture_mean, area_mean, compactness_mean, concavity_mean, concave points_mean, fractal_dimension_m
        ean, radius_se, texture_se, smoothness_se, compactness_se, concavity_se, concave points_se, symmetry_se, radius_wors
        t, texture_worst, area_worst, smoothness_worst, concavity_worst, concave points_worst, symmetry_worst, fractal_dimens
        ion_worst
        Accuracy: 0.963
```

Features selected by Lasso

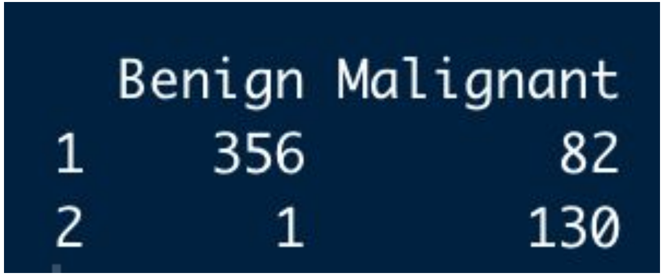
- After analyzing the data on 5 fold cross-validation set, we found that 21 features were important for further analysis, since their estimates weren't reduced to zero.
- The model also explains relationship between different columns and the outcome variable, ex. Columns like concavity_se, area_worst negatively influence the outcome of cancer, whereas columns like smoothness, radius_worst, positively influence the outcome.



Prediction Models

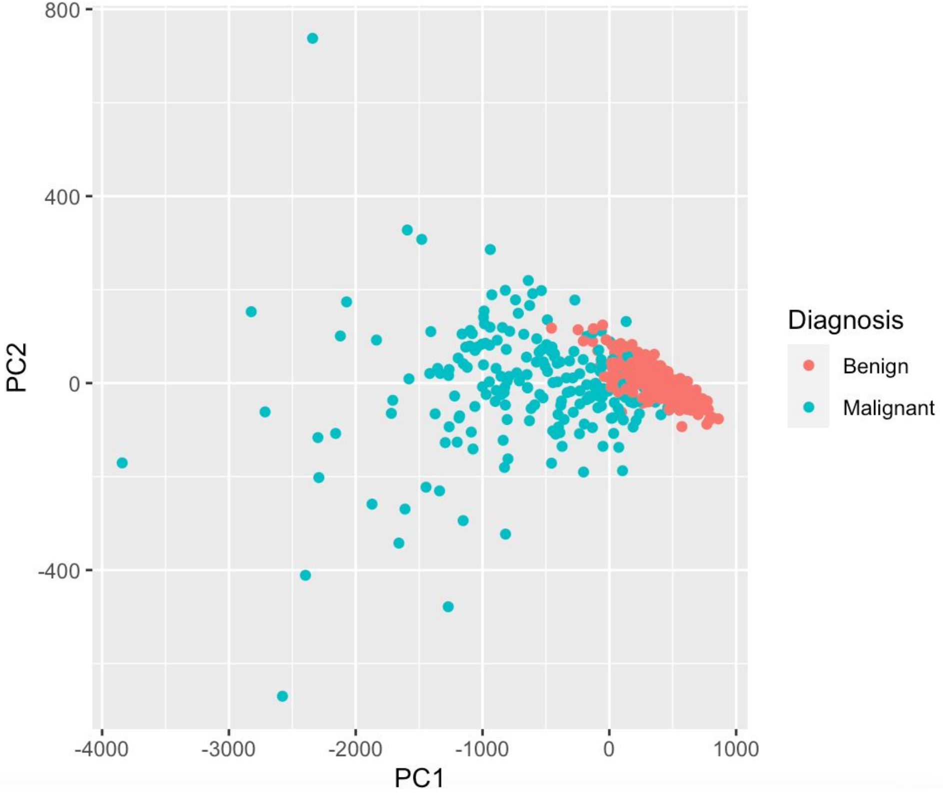
K-Means Clustering

- Used k-means clustering and compared to given labels (malignant vs. benign)
- Used k = 2 since the actual outcome is binary
- In the table, 1 = benign, 2 = malignant for the clusters
- K-means algorithm over classifies individuals into the benign group
- Accuracy of 85%, silhouette score 0.70

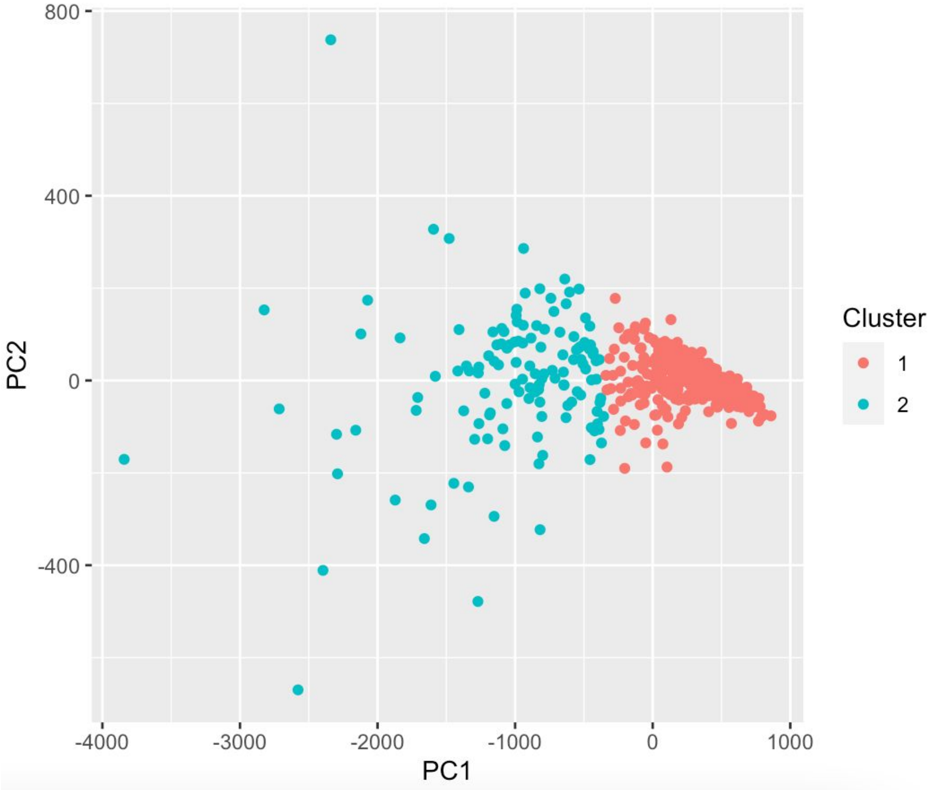


```
  Benign Malignant
1    356        82
2      1       130
```

Labels From Dataset



K-Means clustering



Linear Discriminant Analysis (LDA)

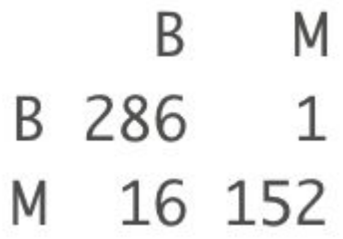
- Supervised classification method
- LDA projects the data onto a subspace
  - Minimize variance within classes, maximize distance between classes

Approach:

- Train vs test split
- Fit to training data, predict on testing data
- Compared the fitted labels to the actual labels
- Accuracy of 96%
- Higher accuracy than k-means

Training Accuracy:

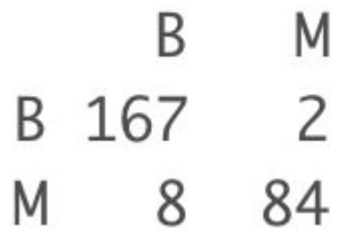
96%



	B	M
B	286	1
M	16	152

Testing Accuracy:

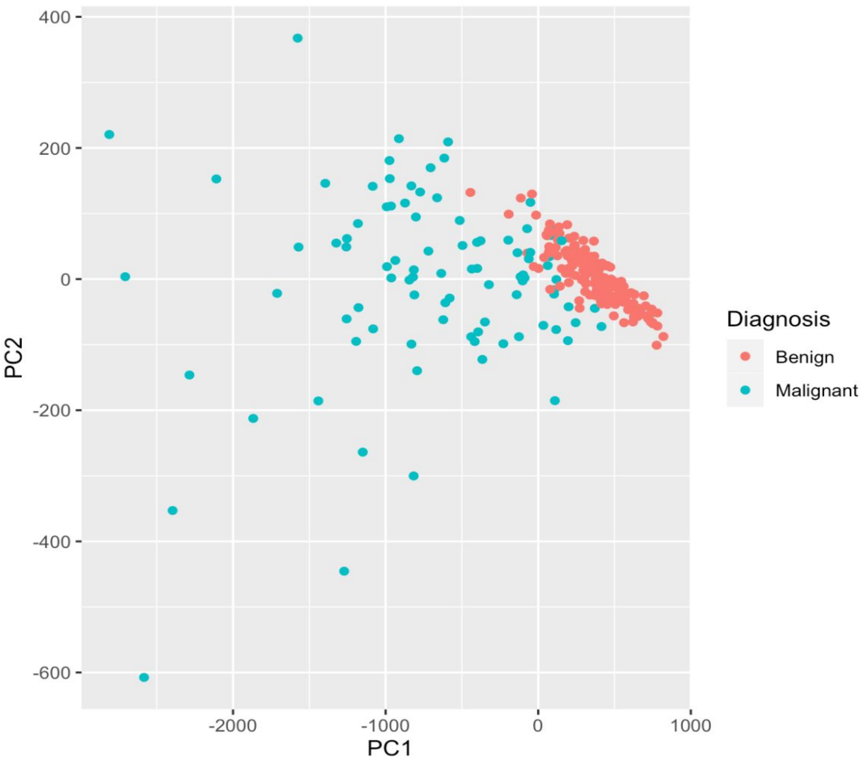
96%



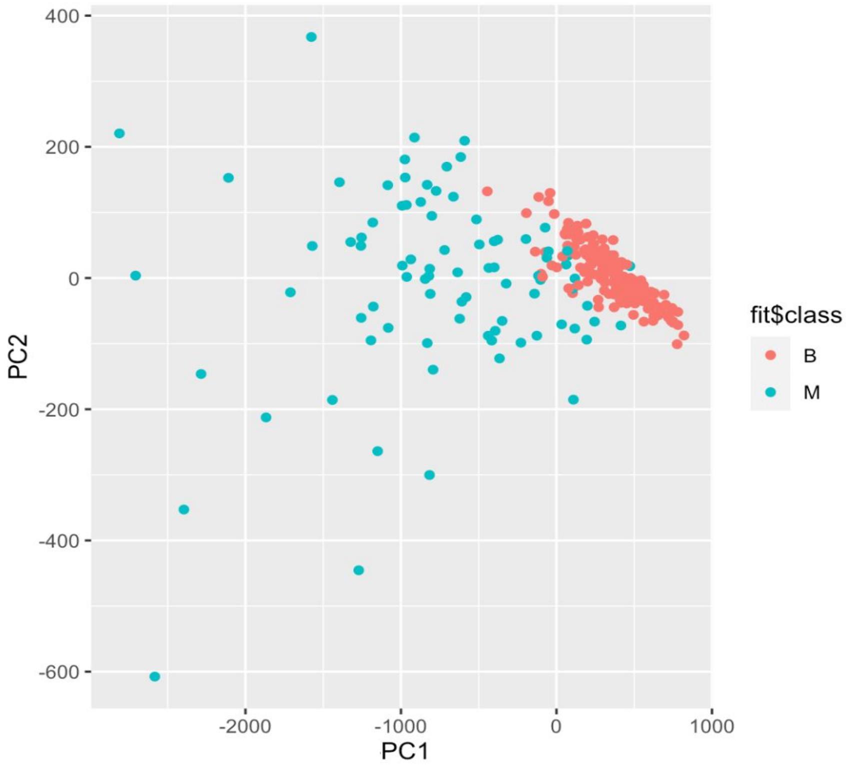
	B	M
B	167	2
M	8	84

LDA

Original Labels

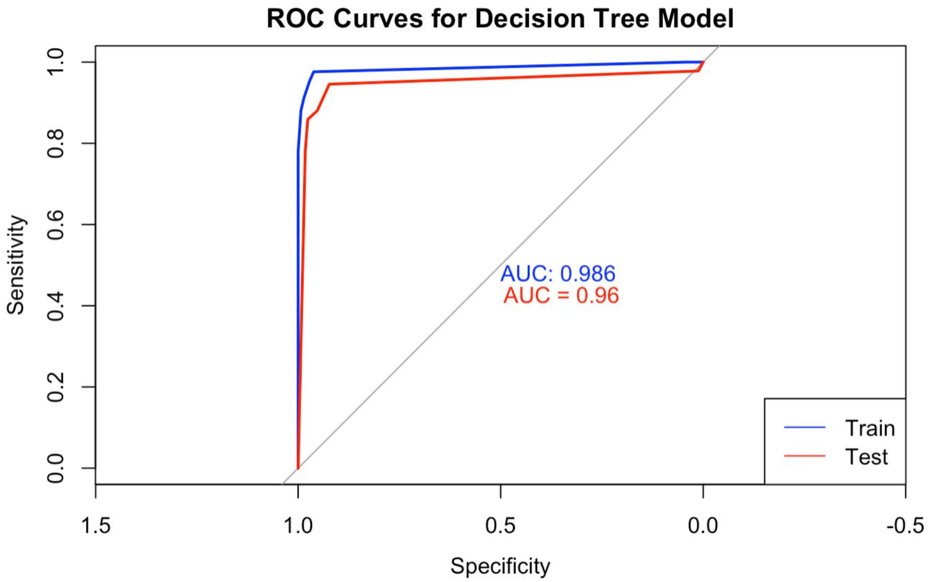


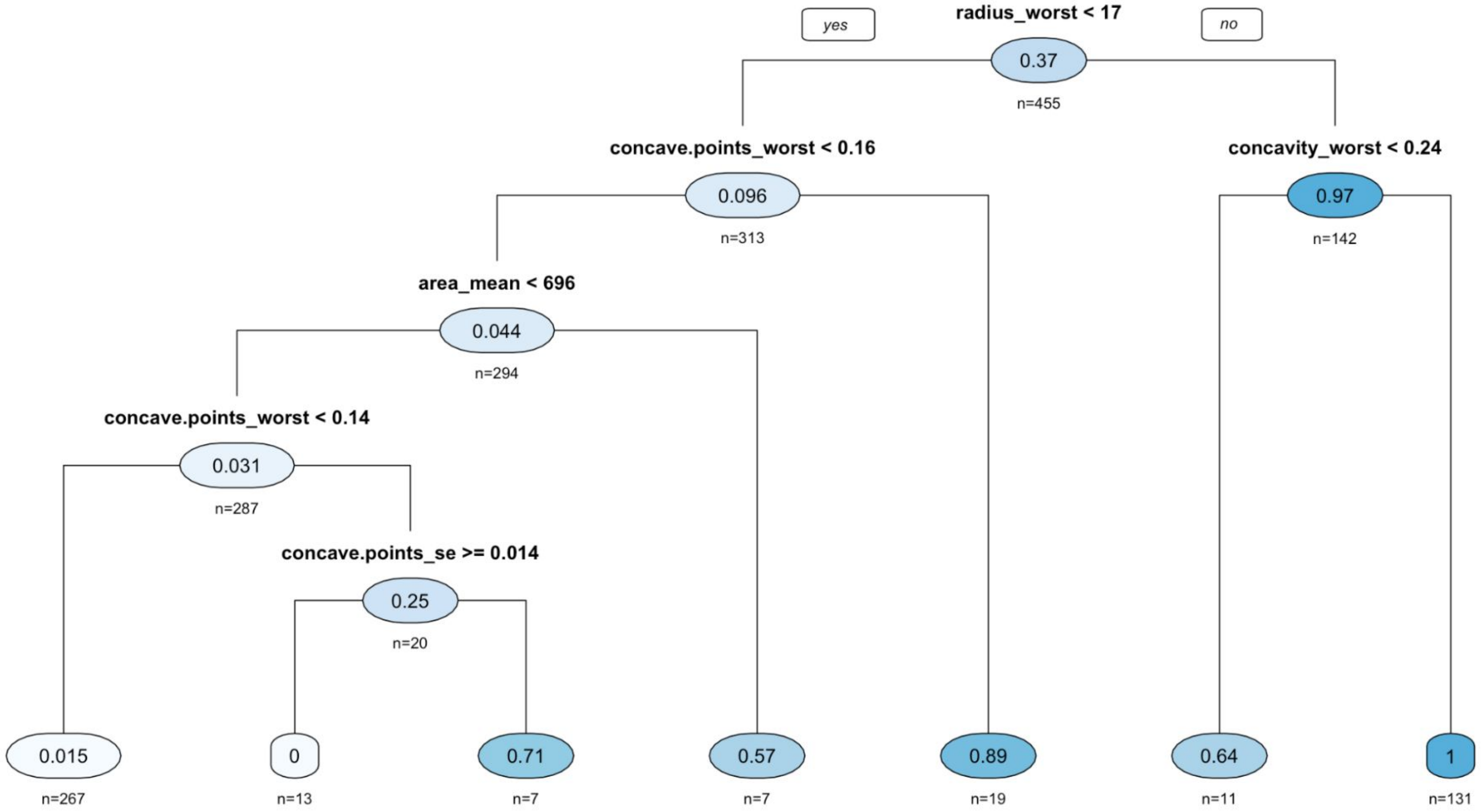
LDA fit on test subset



Decision Tree

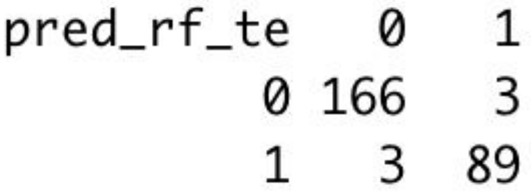
- Used 80%/20% train/test split to subset data
- rpart() function in R used with default parameters to fit model on train set, then predict on test



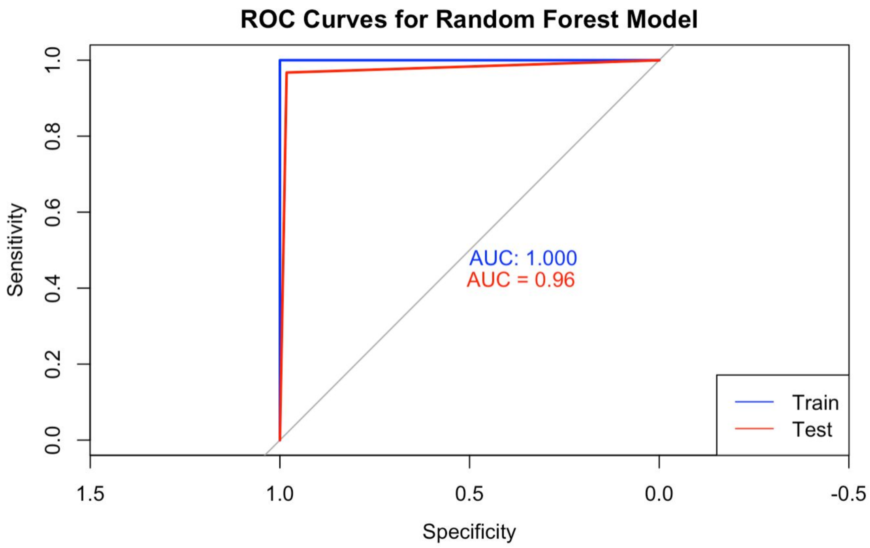

radius_worst < 17
yes
no
0.37
n=455
concave.points_worst < 0.16
0.096
n=313
concavity_worst < 0.24
0.97
n=142
area_mean < 696
0.044
n=294
concave.points_worst < 0.14
0.031
n=287
concave.points_se >= 0.014
0.25
n=20
0.015
n=267
0
n=13
0.71
n=7
0.57
n=7
0.89
n=19
0.64
n=11
1
n=131

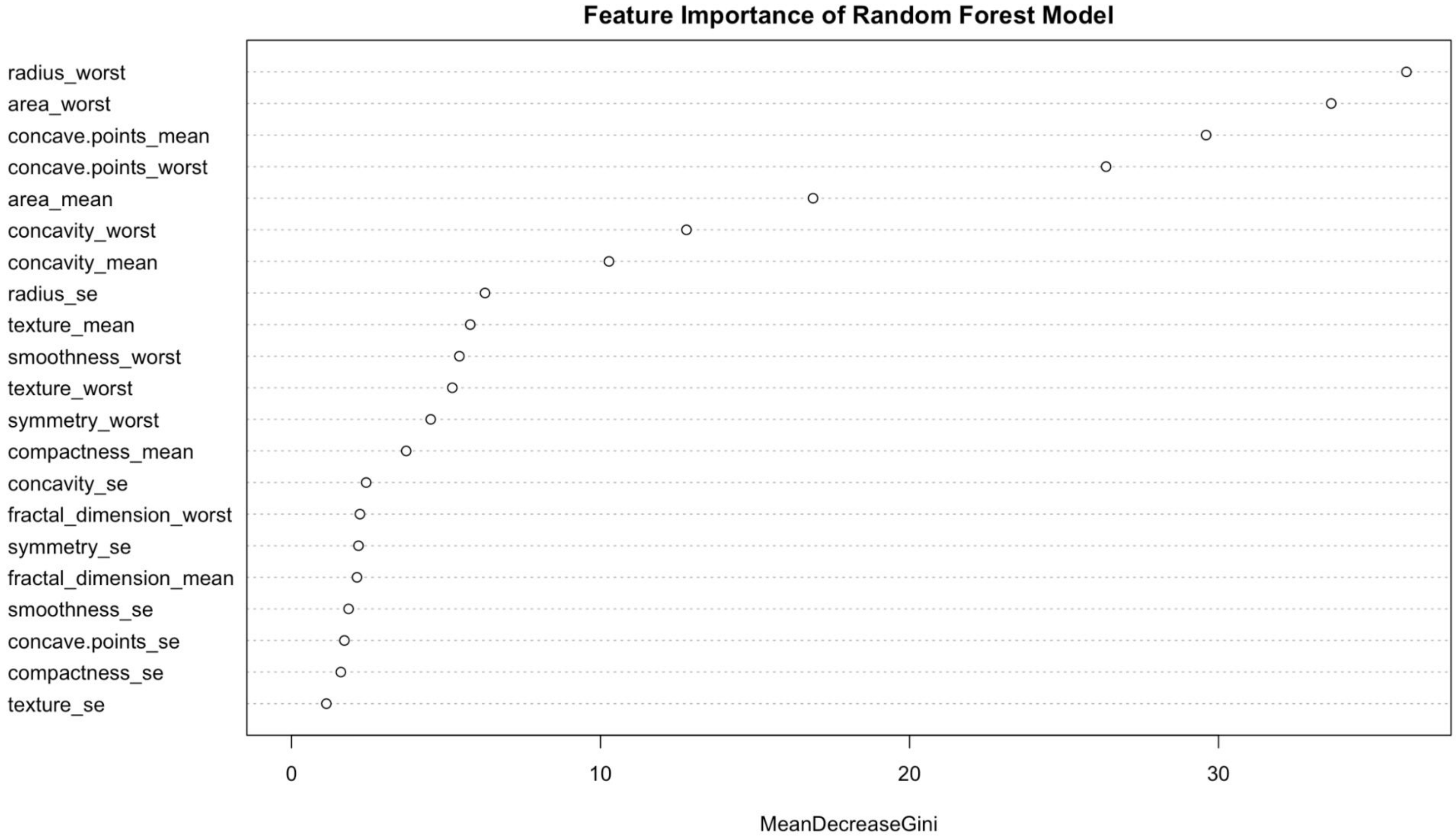
Random Forest

- Used RandomForest() in R to fit model on 80% train set
- Number of trees = 500, features per split = sqrt(p) (p=21)



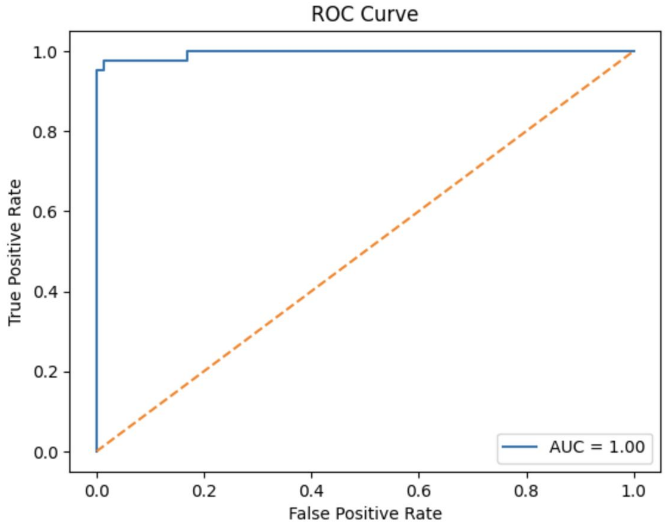
```
pred_rf_te   0   1
         0 166   3
         1   3  89
```




Feature Importance of Random Forest Model
radius_worst
area_worst
concave.points_mean
concave.points_worst
area_mean
concavity_worst
concavity_mean
radius_se
texture_mean
smoothness_worst
texture_worst
symmetry_worst
compactness_mean
concavity_se
fractal_dimension_worst
symmetry_se
fractal_dimension_mean
smoothness_se
concave.points_se
compactness_se
texture_se
0
10
20
30
MeanDecreaseGini

Neural Network

- 1 input layer, 1 hidden layer, and 1 output layer of 1 neuron, with ReLU activation functions for the input and hidden layers (simple, computationally efficient), and sigmoid activation function for the output layer (used when output layer of binary classification problems)
- 50 epochs (number of times the entire dataset is passed through the neural network during the training process)
- batch size of 32 (the model updates its weights and biases based on 32 training examples at a time.)



```
Test loss: 0.069, Test accuracy: 0.982
```

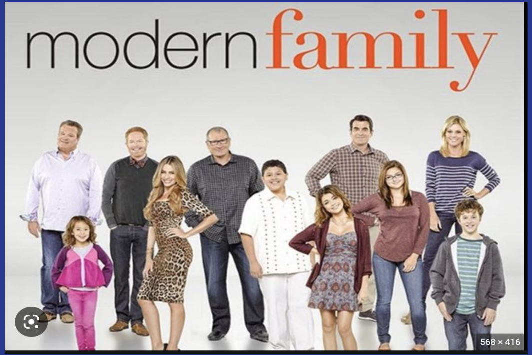
```
ReLu: f(x) = max(0, x)
```

```
Sigmoid: f(x) = 1 / (1 + exp(-x))
```

Discussion

Discussion of Model Results

- PCA able to explain majority of variance in data & visualize diagnosis clusters
- Out of 30 variables, Lasso Regression model identifies 21 as most informative; these variables were included in all classification models
- All classifications obtained high performance (AUC > 0.95)
  - Neural network had the best performance
  - Decision tree likely the best model for clinical interpretability
- Variables for tumor radius, concavity points, and area were among most informative predictors



Thank you!