|  |  |
| --- | --- |
| **<http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29>**  **Breast Cancer Wisconsin (Diagnostic) Data Set**  *Download*: [Data Folder](http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/), [Data Set Description](http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wdbc.names)  **Abstract**: Diagnostic Wisconsin Breast Cancer Database | http://archive.ics.uci.edu/ml/assets/MLimages/Large14.jpg |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Data Set Characteristics:** | Multivariate | **Number of Instances:** | 569 | **Area:** | Life |
| **Attribute Characteristics:** | Real | **Number of Attributes:** | 32 | **Date Donated** | 1995-11-01 |
| **Associated Tasks:** | Classification | **Missing Values?** | No | **Number of Web Hits:** | 272977 |

**Source:**

Creators:   
  
1. Dr. William H. Wolberg, General Surgery Dept.   
University of Wisconsin, Clinical Sciences Center   
Madison, WI 53792   
wolberg **'@'** eagle.surgery.wisc.edu   
  
2. W. Nick Street, Computer Sciences Dept.   
University of Wisconsin, 1210 West Dayton St., Madison, WI 53706   
street **'@'** cs.wisc.edu 608-262-6619   
  
3. Olvi L. Mangasarian, Computer Sciences Dept.   
University of Wisconsin, 1210 West Dayton St., Madison, WI 53706   
olvi **'@'** cs.wisc.edu   
  
Donor:   
  
Nick Street

**Data Set Information:**

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. A few of the images can be found at [[Web Link]](http://www.cs.wisc.edu/~street/images/)   
  
Separating plane described above was obtained using Multisurface Method-Tree (MSM-T) [K. P. Bennett, "Decision Tree Construction Via Linear Programming." Proceedings of the 4th Midwest Artificial Intelligence and Cognitive Science Society, pp. 97-101, 1992], a classification method which uses linear programming to construct a decision tree. Relevant features were selected using an exhaustive search in the space of 1-4 features and 1-3 separating planes.   
  
The actual linear program used to obtain the separating plane in the 3-dimensional space is that described in: [K. P. Bennett and O. L. Mangasarian: "Robust Linear Programming Discrimination of Two Linearly Inseparable Sets", Optimization Methods and Software 1, 1992, 23-34].   
  
This database is also available through the UW CS ftp server:   
ftp ftp.cs.wisc.edu   
cd math-prog/cpo-dataset/machine-learn/WDBC/

**Attribute Information:**

1. ID number   
   2) Diagnosis (M 1= malignant, B 0= benign)    
   3-32)   
     
   Ten real-valued features are computed for each cell nucleus:   
     
   a) radius (mean of distances from center to points on the perimeter)   
   b) texture (standard deviation of gray-scale values)   
   c) perimeter   
   d) area   
   e) smoothness (local variation in radius lengths)   
   f) compactness (perimeter^2 / area - 1.0)   
   g) concavity (severity of concave portions of the contour)   
   h) concave points (number of concave portions of the contour)   
   i) symmetry   
   j) fractal dimension ("coastline approximation" - 1)

|  |  |
| --- | --- |
| **<http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Prognostic%29>**  **Breast Cancer Wisconsin (Prognostic) Data Set**  *Download*: [Data Folder](http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/), [Data Set Description](http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wpbc.names)  **Abstract**: Prognostic Wisconsin Breast Cancer Database | http://archive.ics.uci.edu/ml/assets/MLimages/Large14.jpg |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Data Set Characteristics:** | Multivariate | **Number of Instances:** | 198 | **Area:** | Life |
| **Attribute Characteristics:** | Real | **Number of Attributes:** | 34 | **Date Donated** | 1995-12-01 |
| **Associated Tasks:** | Classification, Regression | **Missing Values?** | Yes | **Number of Web Hits:** | 54475 |

**Source:**

Creators:   
  
1. Dr. William H. Wolberg, General Surgery Dept.   
University of Wisconsin, Clinical Sciences Center   
Madison, WI 53792   
wolberg **'@'** eagle.surgery.wisc.edu   
  
2. W. Nick Street, Computer Sciences Dept.   
University of Wisconsin   
1210 West Dayton St., Madison, WI 53706   
street **'@'** cs.wisc.edu 608-262-6619   
  
3. Olvi L. Mangasarian, Computer Sciences Dept.,   
University of Wisconsin   
1210 West Dayton St., Madison, WI 53706   
olvi **'@'** cs.wisc.edu   
  
Donor:   
  
Nick Street

**Data Set Information:**

Each record represents follow-up data for one breast cancer case. These are consecutive patients seen by Dr. Wolberg since 1984, and include only those cases exhibiting invasive breast cancer and no evidence of distant metastases at the time of diagnosis.   
  
The first 30 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. A few of the images can be found at [[Web Link]](http://www.cs.wisc.edu/~street/images/)   
  
The separation described above was obtained using Multisurface Method-Tree (MSM-T) [K. P. Bennett, "Decision Tree Construction Via Linear Programming." Proceedings of the 4th Midwest Artificial Intelligence and Cognitive Science Society, pp. 97-101, 1992], a classification method which uses linear programming to construct a decision tree. Relevant features were selected using an exhaustive search in the space of 1-4 features and 1-3 separating planes.   
  
The actual linear program used to obtain the separating plane in the 3-dimensional space is that described in:   
[K. P. Bennett and O. L. Mangasarian: "Robust Linear Programming Discrimination of Two Linearly Inseparable Sets", Optimization Methods and Software 1, 1992, 23-34].   
  
The Recurrence Surface Approximation (RSA) method is a linear programming model which predicts Time To Recur using both recurrent and nonrecurrent cases. See references (i) and (ii) above for details of the RSA method.   
  
This database is also available through the UW CS ftp server:   
  
ftp ftp.cs.wisc.edu   
cd math-prog/cpo-dataset/machine-learn/WPBC/

**Attribute Information:**

1) ID number   
2) Outcome (R = recur, N 0= nonrecur)   
3) Time (recurrence time if field 2 = R, disease-free time if field 2 = N)   
4-33) Ten real-valued features are computed for each cell nucleus:   
  
a) radius (mean of distances from center to points on the perimeter)   
b) texture (standard deviation of gray-scale values)   
c) perimeter   
d) area   
e) smoothness (local variation in radius lengths)   
f) compactness (perimeter^2 / area - 1.0)   
g) concavity (severity of concave portions of the contour)   
h) concave points (number of concave portions of the contour)   
i) symmetry   
j) fractal dimension ("coastline approximation" - 1)