# ANALISIS PENERAPAN *KNOWLEDGE DISCOVERY IN DATABASES*: DATASETS *BREAST CANCER*

# TUGAS MANAJEMEN TEKNOLOGI INFORMASI

# 

# Disusun Oleh:

# Fadilah Nur Imani

# 09031281621039

# PROGRAM STUDI SISTEM INFORMASI

# FAKULTAS ILMU KOMPUTER

# UNIVERSITAS SRIWIJAYA

# INDRALAYA

# 2018

# DAFTAR ISI

[DAFTAR ISI 1](#_Toc530369023)

[BAB 1: Definisi dan Informasi Umum 2](#_Toc530369024)

[1.1. KDD 2](#_Toc530369025)

[1.2. *Data Mining* 2](#_Toc530369026)

[1.3. *Dataset* 3](#_Toc530369027)

[1.4. *Breast Cancer* 3](#_Toc530369028)

[BAB 2: *Datasets* 3](#_Toc530369029)

[2.1. *Breast Cancer* 3](#_Toc530369030)

[2.2. *Breast Cancer Wisconsin (Original)* 3](#_Toc530369031)

[2.3. *Breast Cancer Wisconsin (Prognostic)* 4](#_Toc530369032)

[2.4. *Breast Cancer Wisconsin (Diagnostic)* 5](#_Toc530369033)

[BAB 3: Jurnal dan Analisis 6](#_Toc530369034)

[3.1. *Analysis of feature selection with classification: Breast cancer datasets.* 6](#_Toc530369035)

[3.2. Peningkatan Metode Naive Bayes Classification untuk Penentuan Tingkat Keganasan Kanker Payudara Menggunakan Particle Swarm Optimization 7](#_Toc530369036)

[3.3. Diagnosis Penyakit Kanker Payudara Menggunakan Metode Naive Bayes Berbasis Desktop 8](#_Toc530369037)

[3.4. Evaluasi Tiga Jenis Algoritme Berbasis Pembelajaran Mesin untuk Klasifikasi Jenis Tumor Payudara 9](#_Toc530369038)

[3.5. *Breast Cancer Diagnosis on Three Different Datasets Using Multi-Classifiers* 10](#_Toc530369039)

[BAB 4: Kesimpulan 11](#_Toc530369040)

# BAB 1: Definisi dan Informasi Umum

## KDD

KDD adalah singkatan dari *Knowledge Discovery in Databases* yang merupakan sebuah proses penemuan pengetahuan yang bermanfaat dari kumpulan data-data. Dalam KDD terdapat beberapa proses, yaitu:

* pembersihan data *(data cleaning)*
* integrasi data *(data integration)*
* pemilihan data *(data selection)*
* transformasi data *(data transformation)*
* penambangan data *(data mining)*
* evaluasi pola *(pattern evaluation)*
* presentasi pengetahuan *(knowledge presentation)*

## *Data Mining*

*Data Mining* adalah proses penemuan informasi atau pengetahuan yang berguna dari data yang jumlahnya besar secara otomatis. *Data Mining* adalah salah satu proses dari KDD.

## *Dataset*

*Dataset* adalah sebuah koleksi dari data – data.

## *Breast Cancer*

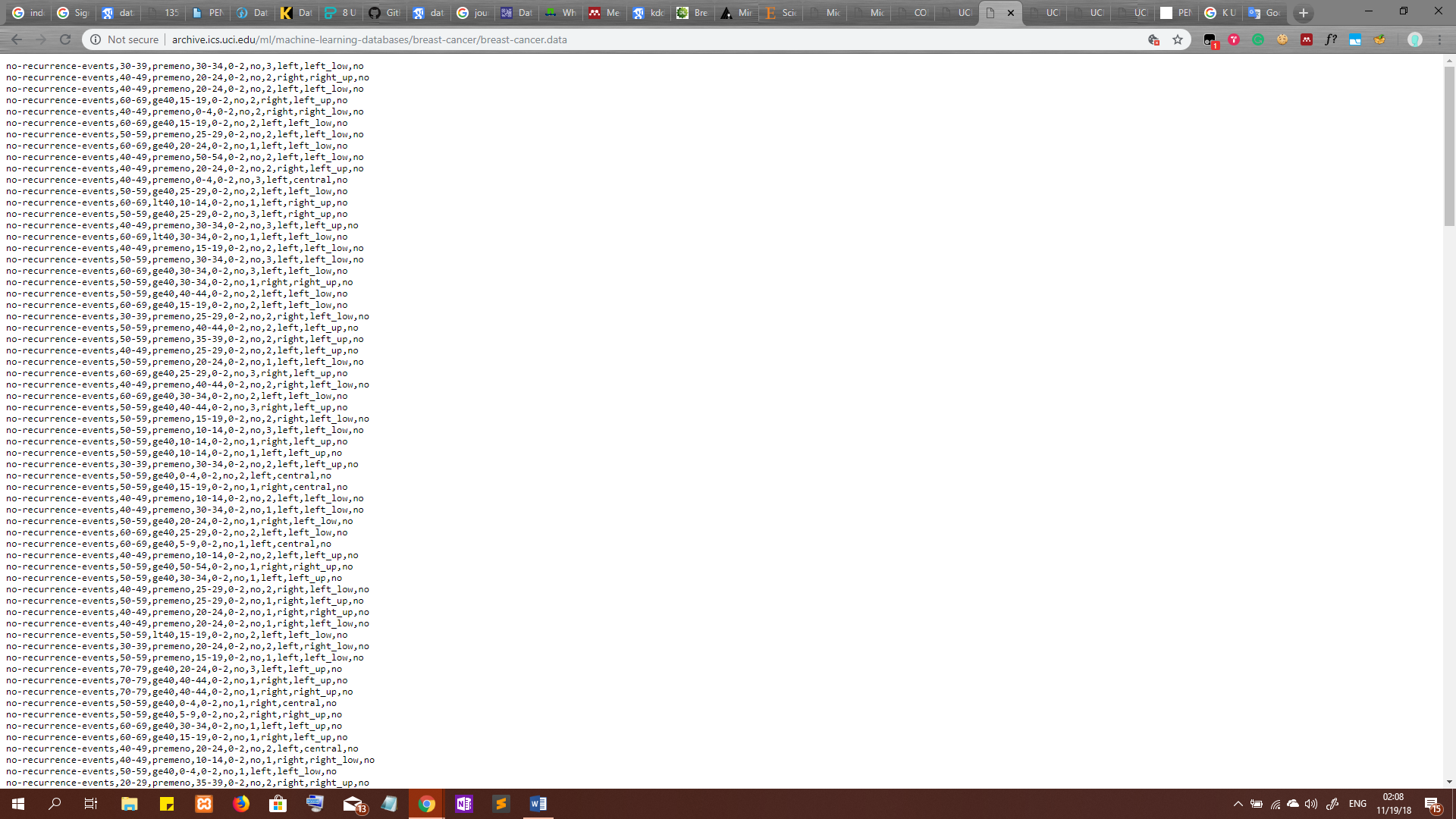
*Breast Cancer* atau selanjutnya disebut kanker payudara, adalah kanker pertumbuhan abnormal pada sel – sel payudara, yang terkadang Nampak seperti benjolan atau massa yang disebut tumor. Ini adalah jenis kanker paling umum yang diderita kaum wanita. Kaum pria juga dapat terserang kanker payudara, walaupun kemungkinannya lebih kecil dari 1 di antara 1000.

# BAB 2: *Datasets*

Telah dikumpulkan dataset dari situs UCI – Machine Learning Repository (<http://archive.ics.uci.edu/ml/datasets.html>) yang terdiri dari 3 dataset dengan nama serupa, yaitu:

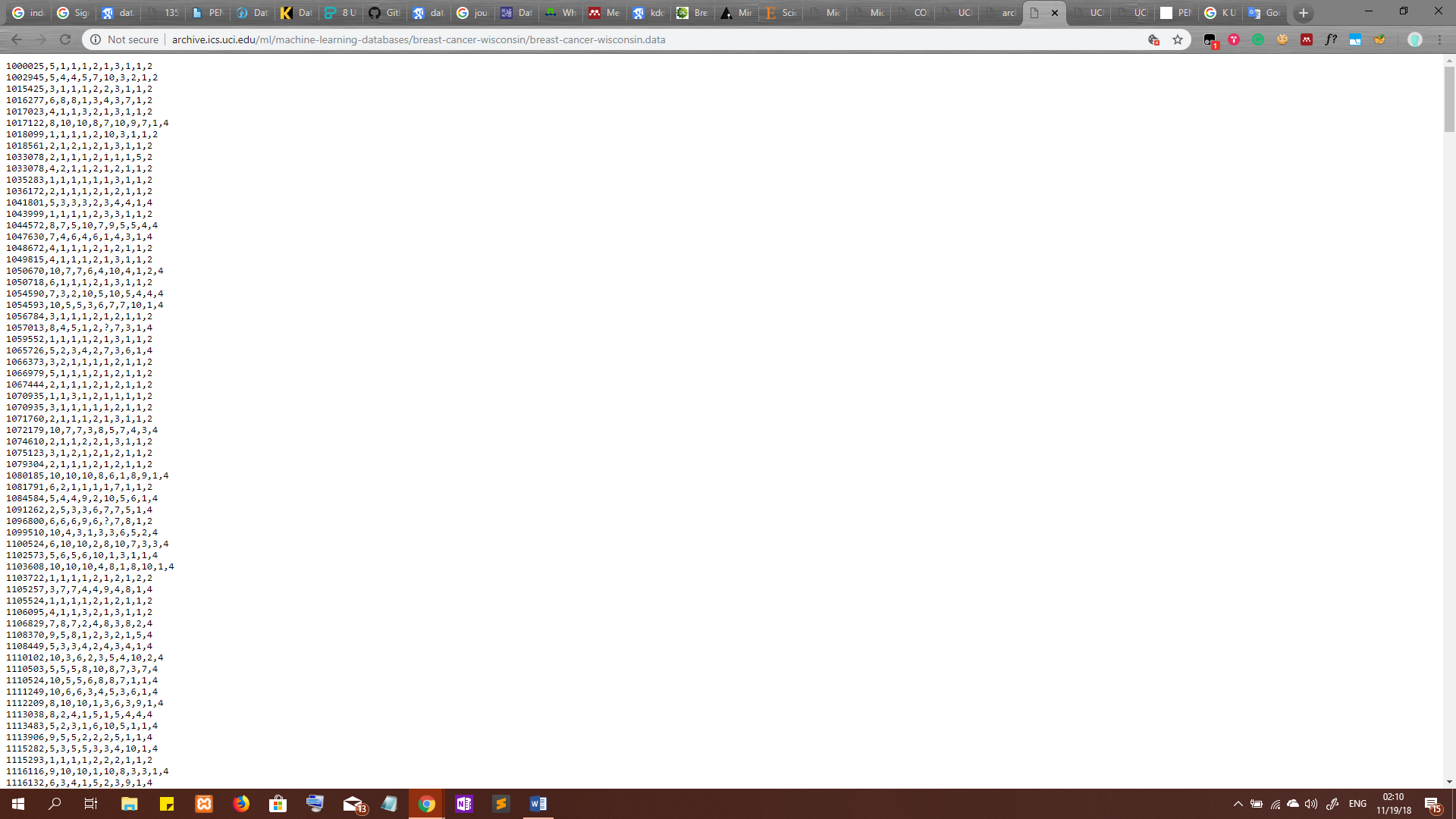
## *Breast Cancer*

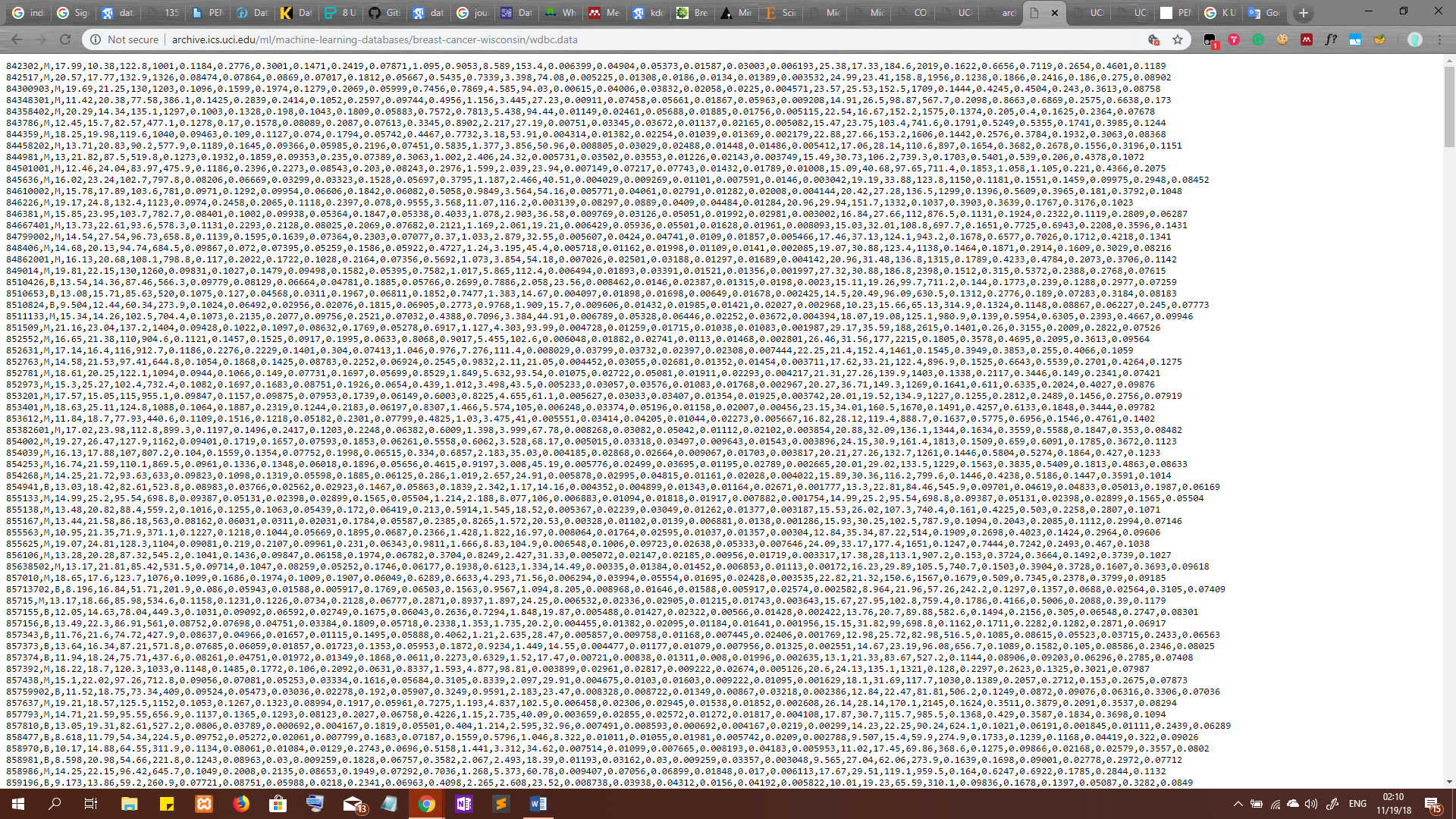
* *Attribute Type: Categorical*
* *Attributes: 9*
* *Instances: 286 records*
* *Year: 1988.*

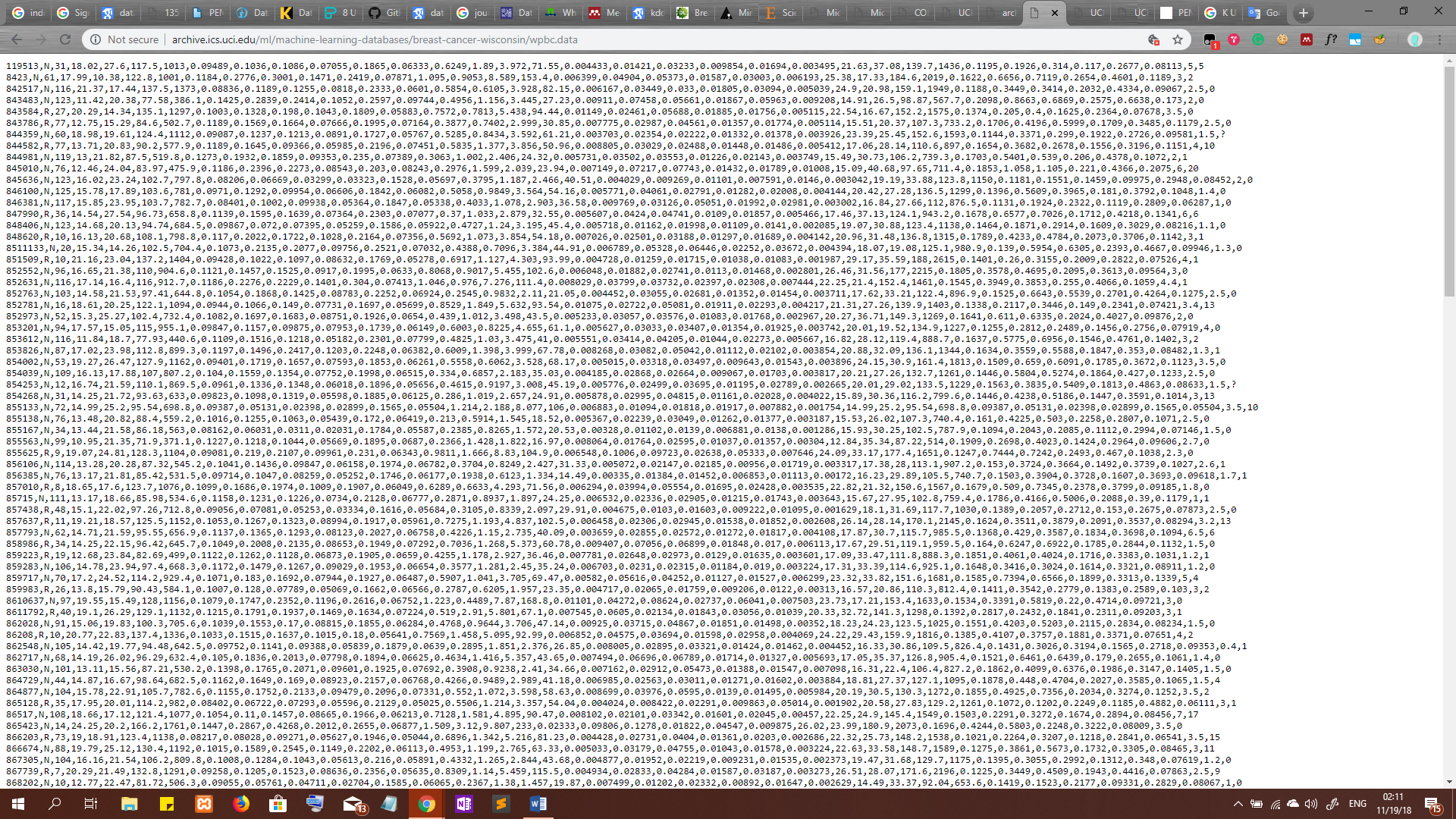


## *Breast Cancer Wisconsin (Original)*

* *Attribute Type: Integer*
* *Attributes: 10*
* *Instances: 699 records*
* *Year: 1992.*

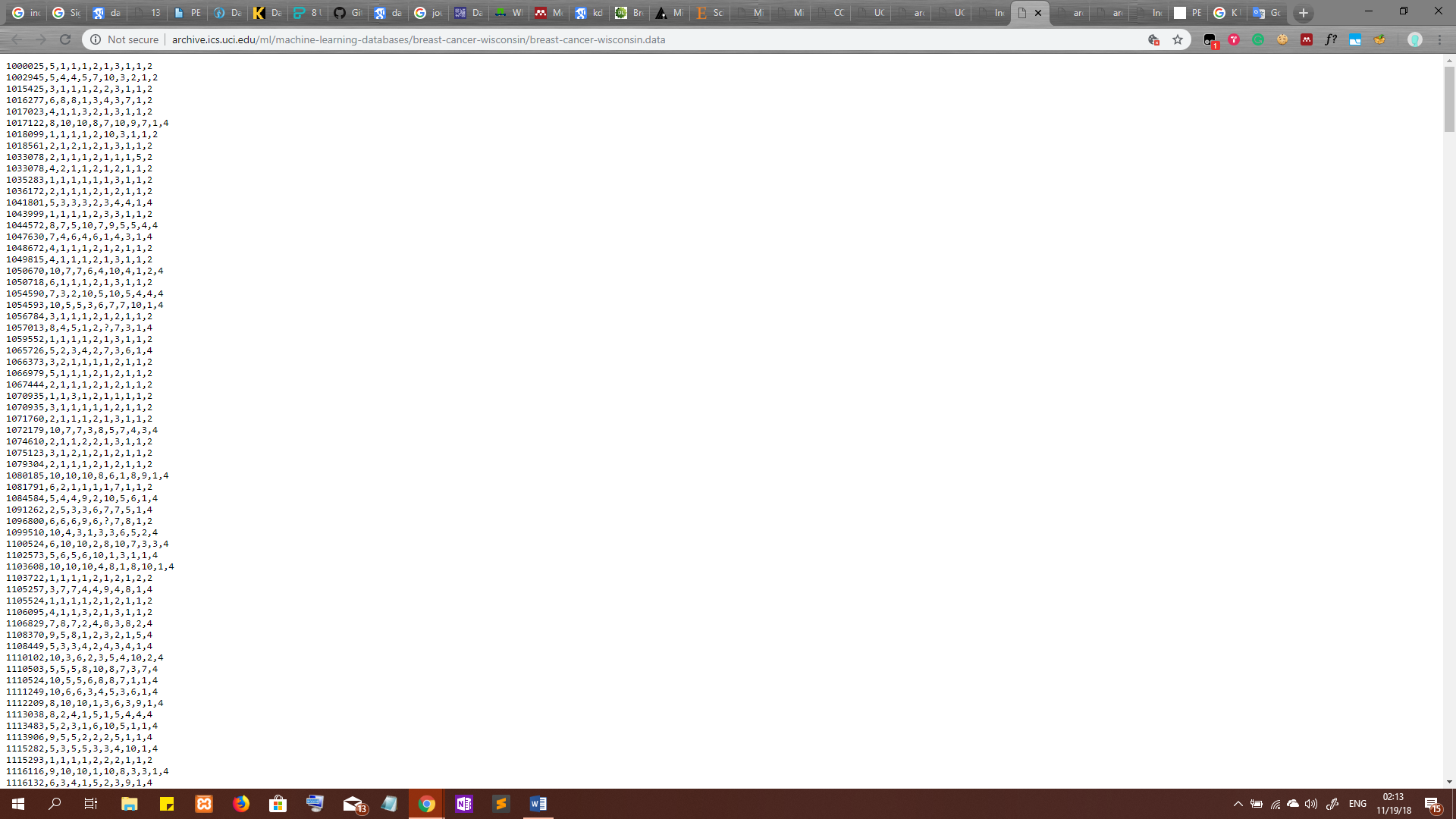


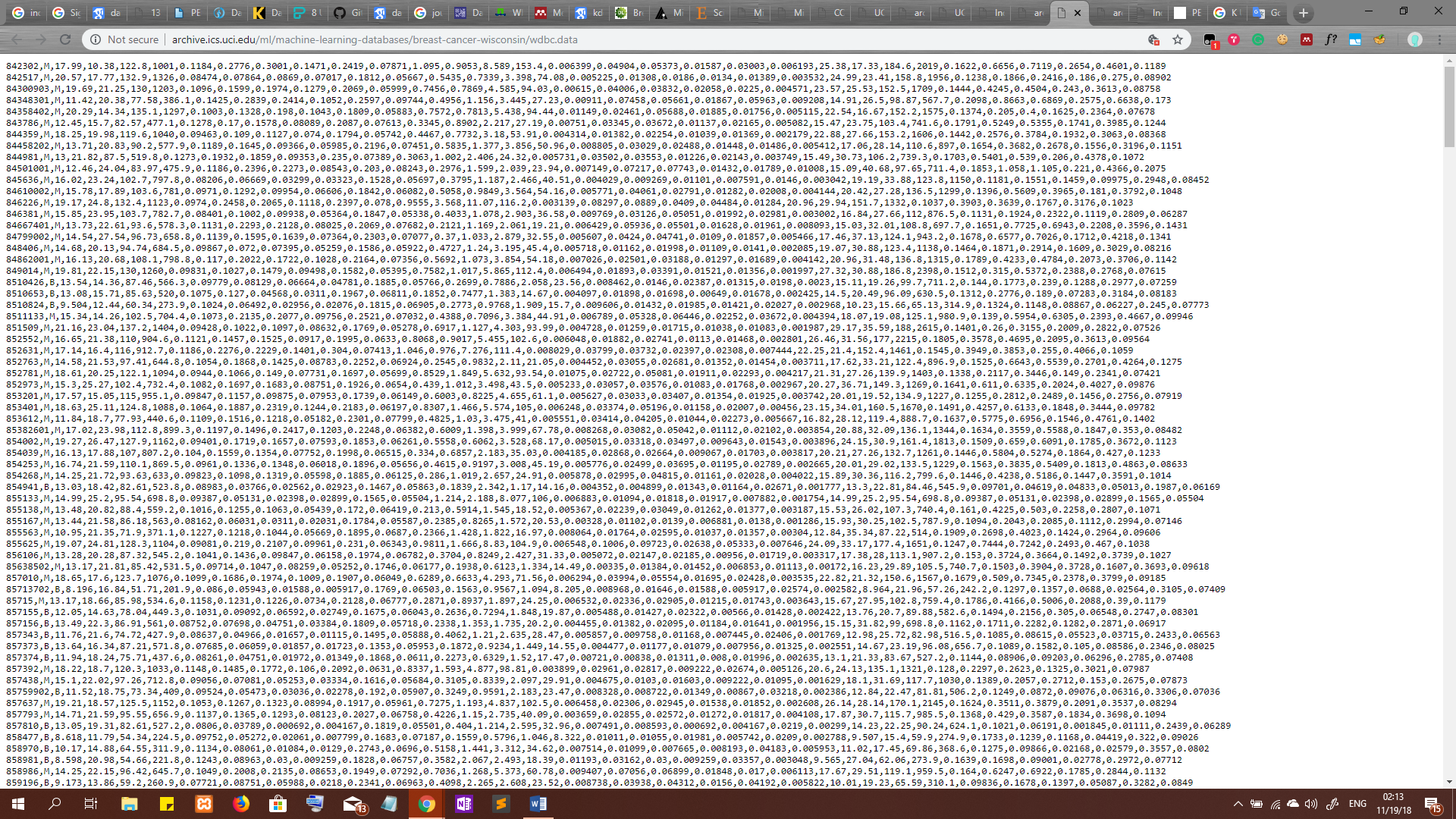


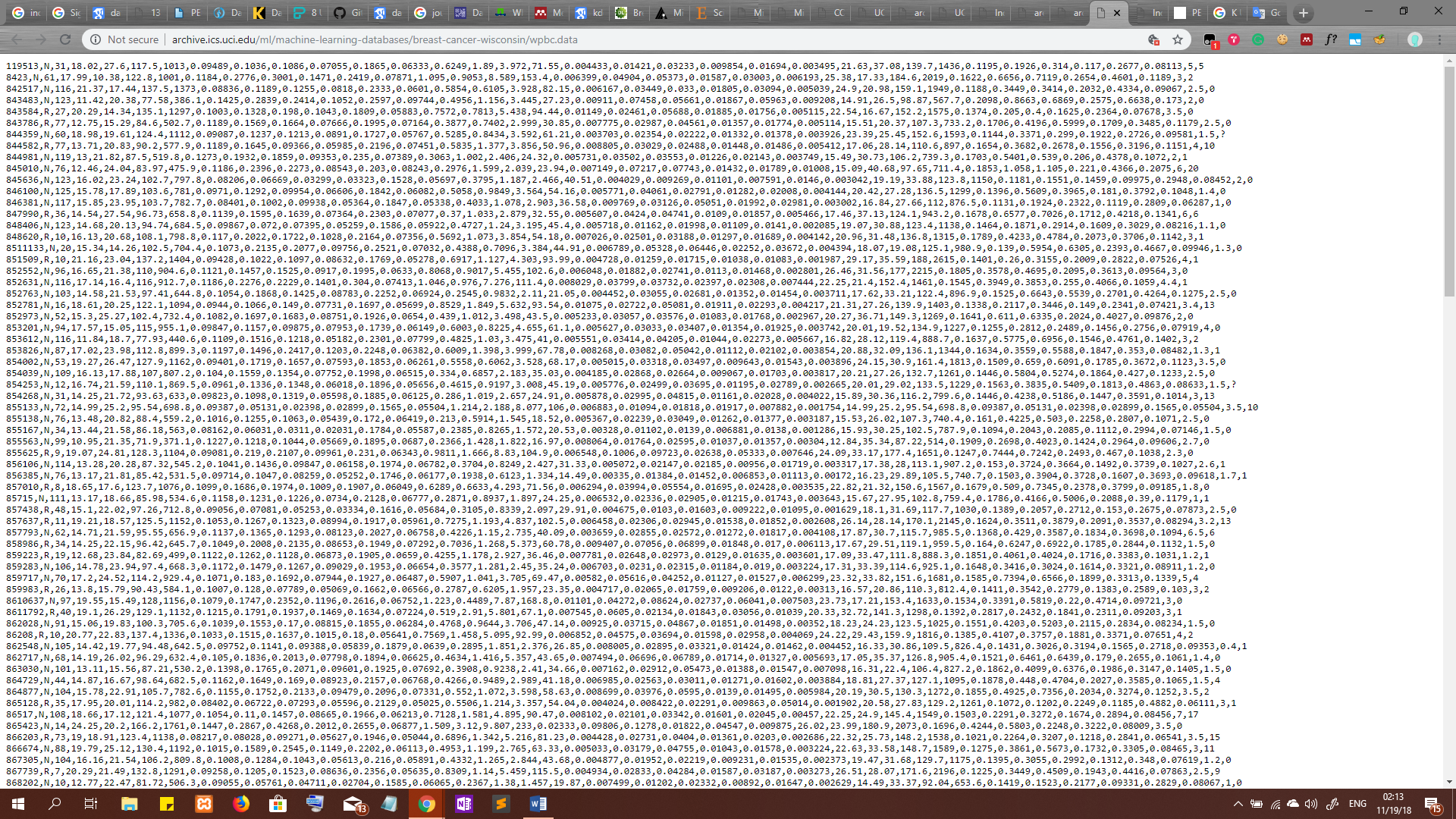


## *Breast Cancer Wisconsin (Prognostic)*

* *Attribute Type: Real*
* *Attributes: 32*
* *Instances: 569 records*
* *Year: 1995.*

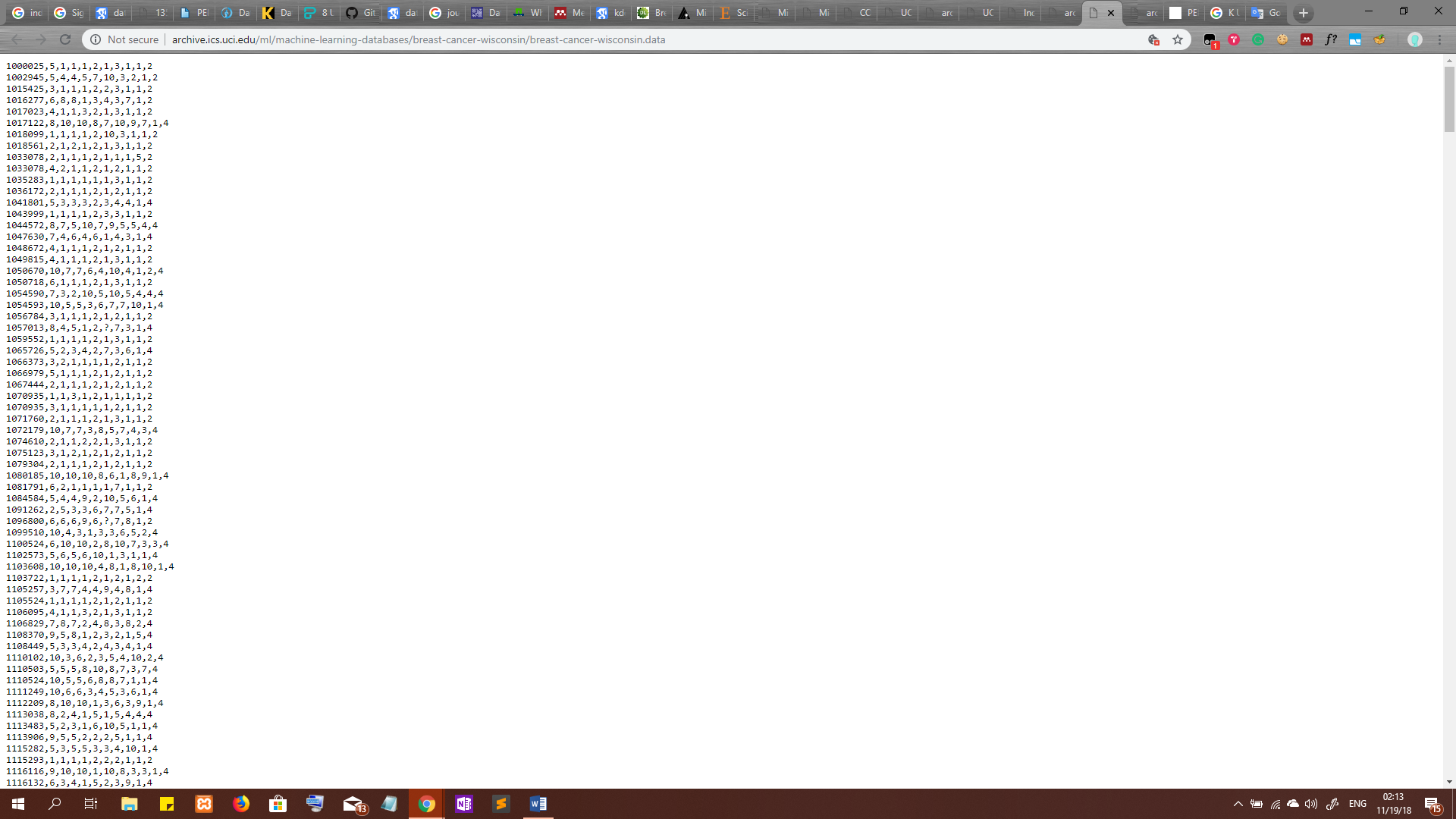
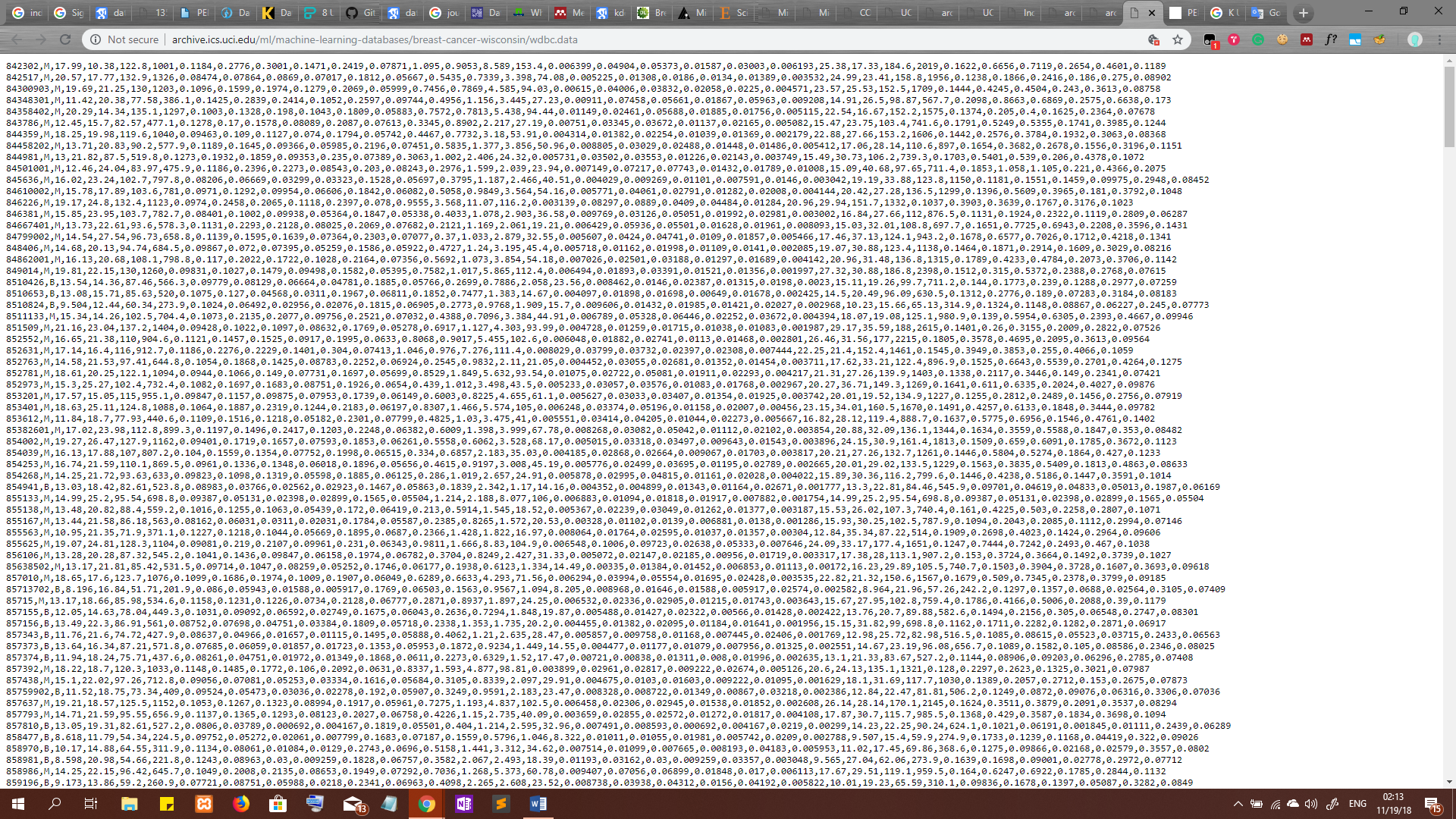


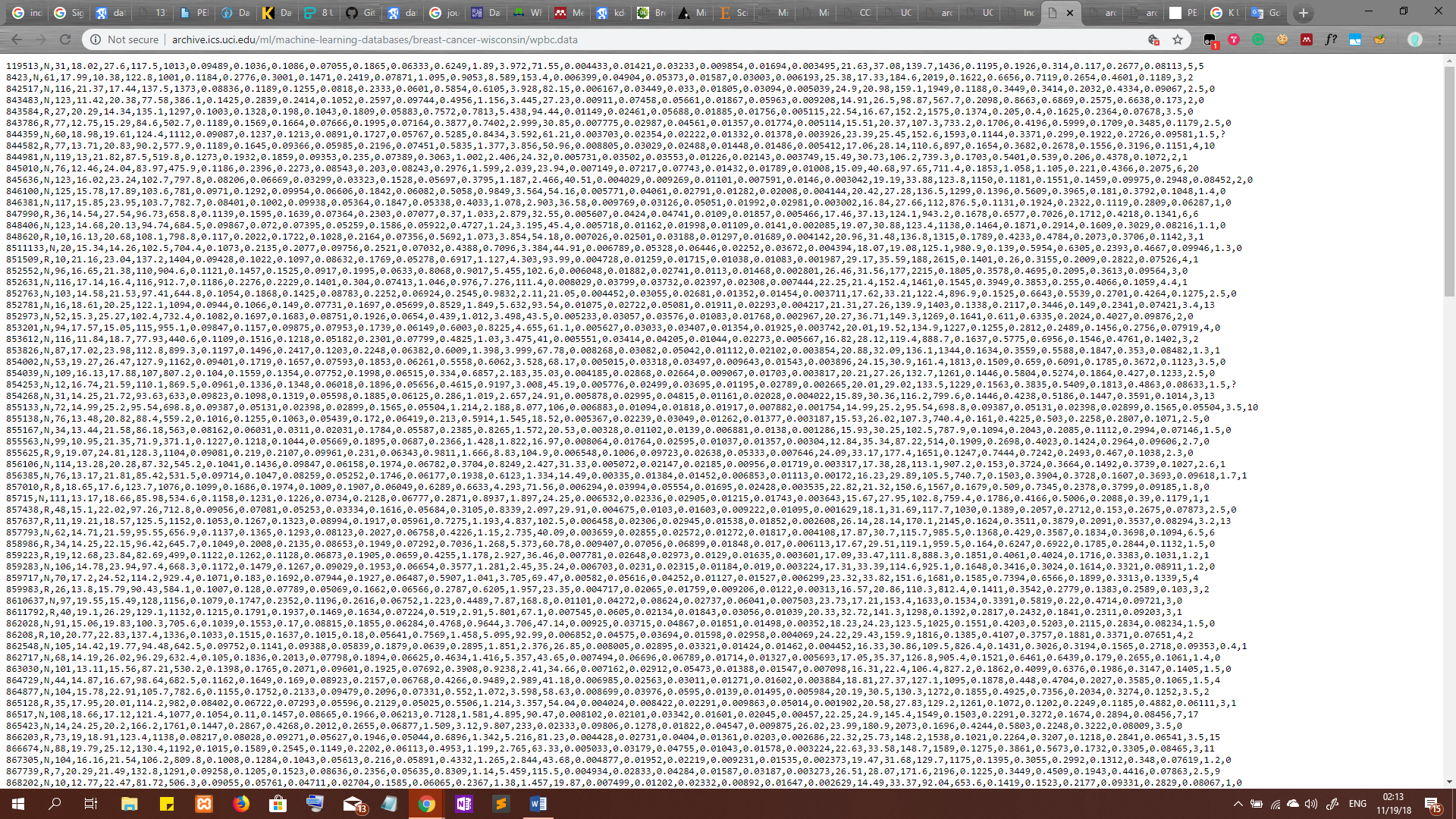




## *Breast Cancer Wisconsin (Diagnostic)*

* *Attribute Type: Real*
* *Attributes: 34*
* *Instances: 198 records*
* *Year: 1995.*



# BAB 3: Jurnal dan Analisis

## *Analysis of feature selection with classification: Breast cancer datasets.*

Lavanya, D. and Rani, D.K.U., 2011. Analysis of feature selection with classification: Breast cancer datasets. *Indian Journal of Computer Science and Engineering (IJCSE)*, 2(5), pp.756-763.

Jurnal ini menganalisis performa dari decision tree classifier-CART dengan dan tanpa seleksi fitur dalam hal akurasi, waktu untuk membangun model dan ukuran pohon pada berbagai Kanker Payudara. Hasilnya menunjukkan bahwa pilihan fitur tertentu menggunakan CART telah meningkatkan klasifikasi akurasi dataset tertentu.

* **Analisis**

Dalam studi ini algoritma CART dipilih untuk mengklasifikasikan data kanker payudara karena memberikan akurasi yang lebih baik untuk dataset medis dari algoritma ID3, C4.5. CART menangani baik secara kategoris dan kontinyu atribut untuk membangun pohon keputusan. Ini juga menangani nilai-nilai yang hilang. Dengan tujuan untuk mencari tahu apakah metode pemilihan fitur yang sama dapat menghasilkan akurasi terbaik untuk berbagai kumpulan data domain yang sama, berbagai percobaan dilakukan pada tiga dataset kanker payudara yang berbeda: Breast Cancer, Breast Cancer Wisconsin (Original) and Breast Cancer Wisconsin (Diagnostic) dengan tipe atribut yang berbeda.

Hasil percobaan menunjukkan bahwa Seleksi Fitur, teknik *Preprocessing* sangat meningkatkan akurasi klasifikasi. Juga disimpulkan bahwa keakuratan *classifier* meningkat dengan penggunaan salah satu dari pemilihan Fitur metode daripada akurasi *classifier* tanpa pemilihan fitur. Dengan niat untuk mencari tahu apakah hal yang sama metode pemilihan fitur dapat mengarah pada akurasi terbaik untuk berbagai kumpulan data dari domain yang sama, berbagai eksperimen dilakukan pada tiga dataset kanker payudara yang berbeda. Kinerja Decision tree classifier-CART dengan dan tanpa pemilihan fitur dalam hal akurasi, waktu untuk membangun model dan ukuran pohon pada berbagai payudara Kumpulan Data Kanker diamati. Dari hasil, terlihat bahwa meskipun peneliti hanya mempertimbangkan kanker payudara dataset, pilihan fitur tertentu mungkin tidak mengarah pada akurasi terbaik untuk semua dataset kanker payudara. Metode pemilihan fitur terbaik untuk dataset tertentu tergantung pada jumlah atribut, tipe atribut dan contoh.

## Peningkatan Metode Naive Bayes Classification untuk Penentuan Tingkat Keganasan Kanker Payudara Menggunakan Particle Swarm Optimization

Fitriani, I.R., 2014. Peningkatan Metode Naive Bayes Classification untuk Penentuan Tingkat Keganasan Kanker Payudara Menggunakan Particle Swarm Optimiza-tion.(Graduate Thesis). *Universitas Dian Nuswantoro. Jawa Tengah*.

Penelitian ini menganalisis tentang pengelompokan data kanker payudara untuk mengetahui kanker tersebut termasuk kanker jinak atau kanker ganas. Untuk mengklasifikasi tingkat keganasan dapat dilakukan dengan pemanfaatan bioinformatics menggunakan teknik data mining salah satuya dengan algoritma klasifikasi *Naive Bayes Classifier* (NBC). NBC dapat bekerja lebih efektif jika dikombinasikan dengan beberapa prosedur pemilihan atribut seperti *Particle Swarm Optimization* (PSO) untuk membobot atribut. Desain penelitian menggunakan model proses CRISP-DM karena penyelesaian masalah dalam penelitian ini mengarah pada masalah strategi bisnis. Penelitian ini menggunakan dataset public Breast Cancer Wisconsin (WBC). Dari hasil pengujian dengan tenfold cross validation dan confusion matrix diketahui bahwa Naive Bayes Classifier (NBC) dalam PSO terbukti memiliki akurasi 96,86%, sedangkan algoritma NBC memiliki akurasi 95,85%. Hasil penelitian ini terbukti bahwa PSO dapat meningkatkan akurasi algoritma NBC.

* **Analisis**

Data yang digunakan pada penelitian ini menggunakan public dataset berasal dari University of California, Irvine (UCI) Machine Learning dengan judul Wisconsin Breast Cancer (Original). Data ini berjumlah 699 record dan terdiri dari 11 atribut, dengan 10 atribut bertipe numerik dan 1 bertipe kategorikal. Elemininasi dilakukan sehingga hanya menggunakan 9 atribut sebagai predictor dan 1 atribut sebagai target. Model NBC menghasilkan nilai akurasi confusion matrix sebesar 95,85%. Lalu, percobaan NBC-PSO dilakukan dengan memberikan 2 nilai population size, yaitu secara default 5, 10-600 dengan maximum number of generation 100 bernilai konstran. Dan kedua, population size bernilai tetap 10 dan maximum number of generation bernilai 100-1500. Akurasi tertinggi dan waktu eksekusi terendah terjadi pada saat NBC-PSO population size bernilai 10 dan maximum number of generation bernilai 100 dengan nilai akurasi sebesar 96,86%.

Algoritma naive bayes classifier (NBC) dengan PSO dan algoritma NBC tanpa PSO, dapat diterapkan untuk penentuan tingkat keganasan kanker payudara. Namun, hasil penelitian menunjukan bahwa model NBC-PSO memiliki akurasi yang lebih baik.

## Diagnosis Penyakit Kanker Payudara Menggunakan Metode Naive Bayes Berbasis Desktop

Safutra, A.R. and Prabowo, D.W., 2016. Diagnosis Penyakit Kanker Payudara Menggunakan Metode Naive Bayes Berbasis Desktop. *Jurnal Penelitian Dosen FIKOM (UNDA),* 6(1).

Ada beberapa tahapan pemeriksaan lebih lanjut dalam mendeteksi penyakit kanker payudara apakah termasuk dalam kategori jinak atau ganas. Pemeriksaan Patologi Anatomi dilakukan di laboratorium dengan memeriksa contoh jaringan tumor yang diambil melalui biopsi. Dalam pemberian keputusan, dokter patalogi anatomi harus menghitung apa yang telah dihasilkan oleh laboratorium dalam pemeriksaan seorang pasien. Perhitungan tersebut memerlukan waktu yang tidak singkat karena menyangkut vonis seorang pasien kanker payudara.

Tujuan dari penelitian ini adalah melakukan diagnosis penyakit kanker payudara berbasis komputer (aplikasi) yang dapat menghasilkan vonis apakah seorang pasien masuk dalam kategori jinak atau ganas. Aplikasi diagnosis penyakit kanker payudara ini menggunakan dataset Wisconsin Diagnostic sebagai data pembelajarannya. Metode Naïve Bayes diterapkan dalam aplikasi diagnosis penyakit kanker payudara karena teknik klasifikasi dari Naïve Bayes tersebut dapat menghasilkan hasil diagnosis yang akurat sebesar 98,6726%.

* **Analisis**

Dalam mendiagnosis penyakit kanker payudara, sebelum ditemukan metode modern, ada metode konvensional, yaitu biopsy dan patologi anatomi. Dalam hal pemberian keputusan, dokter patologi anatomi harus menghitung apa yang sudah dihasilkan oleh laboratorium patologi anatominya. Perhitungan dokter memerlukan waktu yang tidak singkat, oleh karena itu, hal ini memotivasi pengembangan suatu metode komputer (aplikasi) yang dapat mendiagnosis keberadaan penyakit kanker payudara. Aplikasi ini dapat menyediakan prosedur diagnosis penyakit kanker payudara terhadap pasien dengan waktu yang singkat dan memudahkan pekerjaan dokter. Dalam mengembangkan sistem ini, peneliti menggunakan metode Naïve Bayes yang diawali dengan pemilihan dataset Wisconsin Diagnostic sebagai data training untuk mendiagnosis yang sebelumnya disaring sebelum dapat digunakan. Dalam perancangan sistem, output akhir didapat dari perhitungan kemungkinan kanker jinak dan ganas dalam tiap diagnosis, dimana nilai kemungkinan tertinggi antara keduanya, adalah kesimpulan diagnosis.

## Evaluasi Tiga Jenis Algoritme Berbasis Pembelajaran Mesin untuk Klasifikasi Jenis Tumor Payudara

Handayani, A., Jamal, A. and Septiandri, A.A., 2017. Evaluasi Tiga Jenis Algoritme Berbasis Pembelajaran Mesin untuk Klasifikasi Jenis Tumor Payudara. *Jurnal Nasional Teknik Elektro dan Teknologi Informasi (JNTETI)*, 6(4).

Salah satu cara untuk membedakan tumor ganas payudara dan tumor jinak payudara adalah dengan melakukan tes Fine Needle Aspiration (FNA). Metode ini disukai karena mudah dilakukan, aman, sederhana, murah, serta dapat dilakukan pada pasien rawat jalan maupun rawat inap. Meskipun metode ini banyak disukai, tetapi FNA memiliki tingkat akurasi yang bervariasi. Oleh karena itu, makalah ini disusun untuk menemukan model terbaik dari data tes FNA untuk mengklasifikasikan tumor ganas dan tumor jinak payudara dengan melakukan evaluasi pada nilai Area Under the Curve (AUC) dari Algoritme Extreme Gradient Boosting (XGBoost), Support Vector Machine Kernel Radial Basic Function (SVM-RBF), dan Multilayer Perceptron (MLP). Data yang digunakan merupakan data Wisconsin Breast Cancer (WBC) yang tersedia di UCI Machine Learning Repository. Hasil terbaik berdasarkan nilai AUC dan score cost terendah diraih oleh algoritme SVM-RBF pada data set yang - dilakukan eliminate missing value dengan nilai AUC 99,23 dan score cost $2.740,20.

* **Analisis**

Penelitian ini meneliti antara 3 jenis algoritma pembelajaran mesin, yaitu (1) Support Vector Machine (SVM), (2) Neural Networks, (3) Extreme Gradient Boosting. Penelitian membuktikan bahwa model terbaik dihasilkan oleh SVM. Berdasarkan kurva RoC secara keseluruhan, model terbaik dihasilkan oleh algoritma SVM pada Data set B dengan nilai AUC 99,23. Sedangkan berdasarkan confusion matrix secara keseluruhan, score cost terendah dihasilkan oleh algoritme SVM pada Data set B, yaitu $2.740,20, sehingga dari ketiga algoritma ini, berdasarkan nilai AUC dan score cost confusion matrix, algoritme SVM merupakan algoritme yang paling tepat digunakan untuk kasus klasifikasi jenis tumor payudara dengan jumlah data yang tersedia. Selain itu, kesimpulan lain yang dapat disampaikan adalah perbedaan perlakuan yang dilakukan terhadap data set A dan B tidak menghasilkan perubahan yang signifikan dari jumlah FP dan FN, walaupun menghasilkan score cost yang cukup jauh perbedaannya

## *Breast Cancer Diagnosis on Three Different Datasets Using Multi-Classifiers*

Salama, G.I., Abdelhalim, M. and Zeid, M.A.E., 2012. Breast cancer diagnosis on three different datasets using multi-classifiers. *Breast Cancer (WDBC)*, 32(569), p.2.

Makalah ini menyajikan perbandingan antara pohon hasil pengklasifikasi berbeda (J48), Multi-Layer Persepsi (MLP), Naive Bayes (NB), Pengoptimalan Minimal Berurutan (SMO), dan Instance Berbasis untuk K-Nearest tetangga (IBK) di tiga database kanker payudara yang berbeda (Wisconsin Payudara Kanker (WBC), Wisconsin Diagnosis Kanker Payudara (WDBC) dan Wisconsin Prognosis Kanker Payudara (WPBC)) dengan menggunakan akurasi klasifikasi dan confusion matrix berdasarkan 10 kali lipat metode validasi silang. Juga, peneliti memperkenalkan fusi di tingkat klasifikasi antara pengklasifikasi ini untuk mendapatkan hasil maksimal pendekatan multi-penggolongan yang cocok untuk setiap dataset. Hasil percobaan menunjukkan bahwa dalam klasifikasi menggunakan fusi MLP dan J48 dengan PCA lebih unggul dari pengklasifikasi lainnya menggunakan data kumpulan WBC. PCAA dalam dataset WBC sebagai mengatur fitur yang berkorelasi. Atribut yang dipilih adalah: Keseragaman Ukuran Sel, Mitosis, Ketebalan Gumpalan, Bare Nuclei, Single Ukuran sel epitel, adhesi Marginal, Bland Chromatin dan Kelas. Eksperimen dilakukan di alat penambangan data WEKA.

* **Analisis**

Penelitian ini menggunakan tiga dataset berbeda, yaitu Wisconsin Payudara Kanker (WBC), Wisconsin Diagnosis Kanker Payudara (WDBC) dan Wisconsin Prognosis Kanker Payudara (WPBC). Data dalam WDBC menunjukkan bahwa klasifikasi menggunakan SMO saja atau menggunakan fusi SMO dan MLP atau SMO dan IBK lebih unggul dari pengklasifikasi lainnya. Data dalam WPBC menunjukkan bahwa klasifikasi menggunakan fusi MLP, J48, SMO dan IBK lebih unggul dari pengklasifikasi lainnya.

# BAB 4: Kesimpulan

Dari 4 dataset yang berbeda yang digunakan dalam penelitian 5 jurnal berkaitan, dapat disimpulkan bahwa penggunaan metode algoritma yang berbeda dalam data mining dapat menghasilkan hasil yang berbeda satu sama lain, terutama selisih dalam tingkat akurasi. 5 penelitian yang dituangkan dalam jurnal tersebut menunjukkan bahwa, pengelolaan data dengan metode yang berbeda, dapat mempengaruhi tingkat akurasi hasil akhirnya. Dari beragam penelitian tersebut, terlihat bahwa data mining memiliki beragam metode algoritma, diantaranya:

* C 4.5
* K-Means
* Support Vector Machines
* Neural Networks
* Extreme Gradient Boosting
* Naive Bayes
* Classiﬁcation and Regression Trees