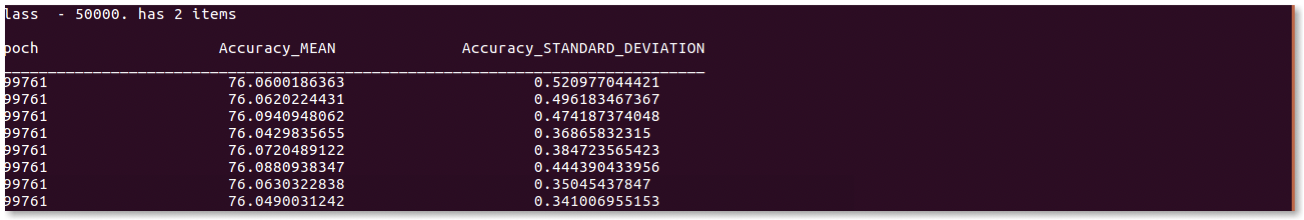
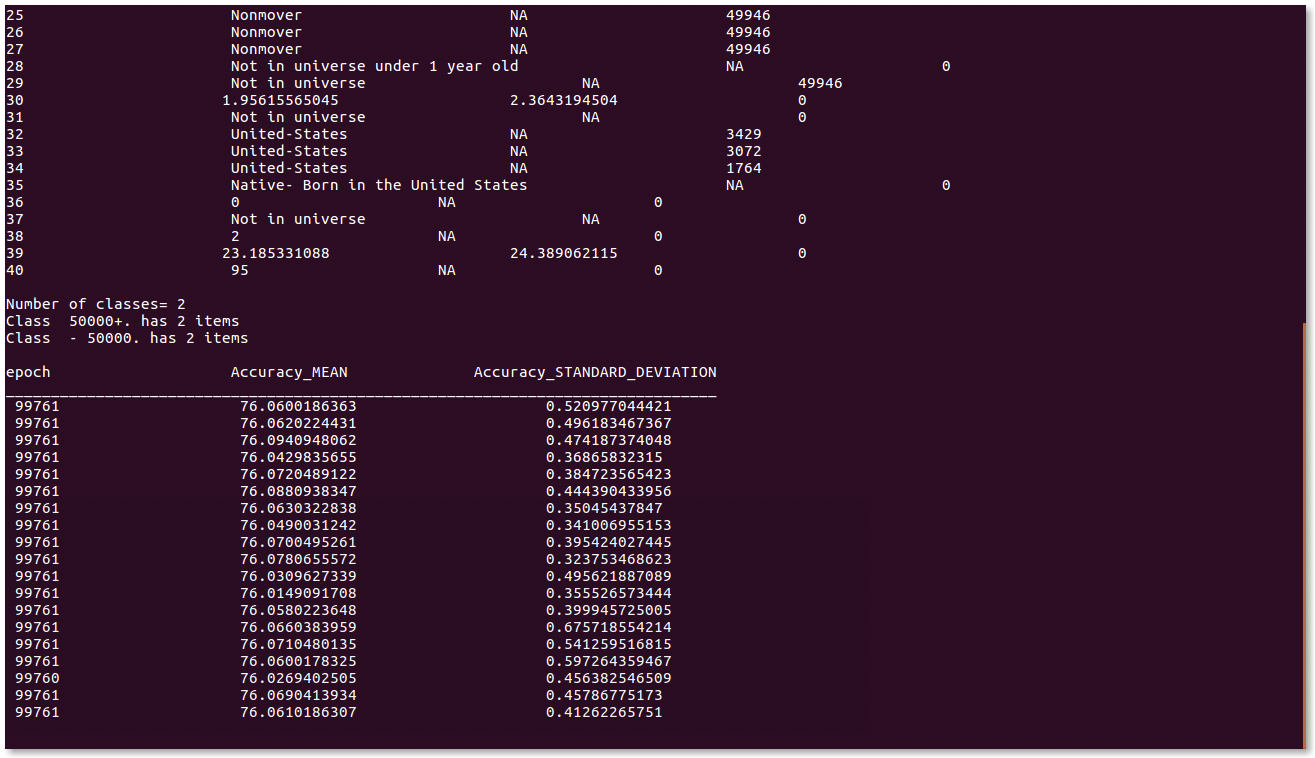
Implemented a Naive Bayes Classifier for UCI Census-Income (KDD) Data Set using only the Discrete and Categorical attributes/features.

OutPut:  
With considering continuous data and calculation gaussion probabiility for that.



Without considering continuous data and calculation gaussion probabiility for that.



Part 1:  
Summary of the data.

Number of classes= 2  
Class 50000+. has 6186 items

Class - 50000. has 93576 items Database - 50000. has 99762 items

1  
{' Private': 36070, ' Self-employed-incorporated': 1647, ' State government': 2166, ' Self-employed-not incorporated': 4279, ' Not in universe': 50078, ' Without pay': 74, ' Federal government': 1404, ' Never worked': 203, ' Local government': 3832}  
2  
{' 10': 0, ' 11': 902, ' 38': 845, ' 39': 1455, ' 14': 141, ' 15': 221, ' 16': 252, ' 17': 68, ' 32': 1821, ' 33': 8711, ' 30': 594, ' 31': 569, ' 36': 494, ' 37': 2138, ' 34': 1353, ' 35': 1664, ' 13': 466, ' 28': 62, ' 49': 261, ' 48': 290, ' 50': 820, ' 18': 206, ' 19': 720, ' 12': 679, ' 25': 482, ' 24': 730, ' 27': 294, ' 26': 56, ' 21': 280, ' 20': 19, ' 23': 273, ' 22': 468, ' 47': 803, ' 43': 4226, ' 45': 2203, ' 44': 1221, ' 29': 2210, ' 42': 2303, ' 41': 1850, ' 40': 820, ' 2': 1076, ' 3': 324, ' 0': 50282, ' 1': 380, ' 6': 241, ' 7': 246, ' 4': 3066, ' 5': 268, ' 8': 252, ' 9': 489, ' 46': 101, ' 51': 15}  
3

{' 10': 1805, ' 11': 296, ' 12': 1692, ' 13': 614, ' 14': 417, ' 15': 438, ' 16': 1741, ' 17': 902, ' 32': 1188, ' 33': 1707, ' 30': 940, ' 31': 1375, ' 36': 2092, ' 37': 1092, ' 34': 2021, ' 35': 1623, ' 39': 480, ' 9': 370, ' 18': 573, ' 19': 2770, ' 25': 415, ' 24': 893, ' 27': 425, ' 26': 3813, ' 21': 272, ' 20': 30, ' 23': 1634, ' 22': 219, ' 43': 660, ' 45': 79, ' 44': 778, ' 29': 2670, ' 28': 786, ' 41': 803, ' 42': 929, ' 2': 4355, ' 3': 1655, ' 0': 50282, ' 1': 275, ' 6': 199, ' 7': 363, ' 4': 680, ' 5': 438, ' 8': 1017, ' 40': 336, ' 46': 15, ' 38': 1558}

4  
{' 1st 2nd 3rd or 4th grade': 905, ' 12th grade no diploma': 1136, ' Less than 1st grade': 415, ' Some college but no degree': 13953, ' Masters degree(MA MS MEng MEd MSW MBA)': 3305, ' 10th grade': 3812, ' 11th grade': 3522, ' Bachelors degree(BA AB BS)': 9884, ' Prof school degree (MD DDS DVM LLB JD)': 875, ' Associates degree-academic program': 2078, ' 5th or 6th grade': 1713, ' Doctorate degree(PhD EdD)': 619, ' High school graduate': 24146, ' 7th and 8th grade': 4148, ' Children': 23441, ' Associates degree-occup /vocational': 2689, ' 9th grade': 3104}  
6  
{' Not in universe': 93423, ' College or university': 2831, ' High school': 3505}  
7  
{' Separated': 1695, ' Divorced': 6449, ' Widowed': 5324, ' Married-spouse absent': 715, ' Married- civilian spouse present': 42092, ' Married-A F spouse present': 338, ' Never married': 43142}  
8  
{' Mining': 324, ' Transportation': 2210, ' Medical except hospital': 2303, ' Retail trade': 8711, ' Finance insurance and real estate': 3018, ' Construction': 3066, ' Wholesale trade': 1821, ' Manufacturing- nondurable goods': 3393, ' Private household services': 494, ' Communications': 594, ' Utilities and sanitary services': 569, ' Not in universe or children': 50282, ' Forestry and fisheries': 101, ' Other professional services': 2203, ' Education': 4226, ' Personal services except private HH': 1455, ' Agriculture': 1457, ' Entertainment': 820, ' Public administration': 2177, ' Hospital services': 1850, ' Manufacturing-durable goods': 4444, ' Social services': 1221, ' Business and repair services': 2984, ' Armed Forces': 15}  
9  
{' Other service': 6176, ' Technicians and related support': 1471, ' Not in universe': 50282, ' Private household services': 425, ' Professional specialty': 6868, ' Protective services': 786, ' Farming forestry and fishing': 1519, ' Sales': 6020, ' Machine operators assmblrs & inspctrs': 3185, ' Transportation and material moving': 2039, ' Handlers equip cleaners etc ': 2070, ' Precision production craft & repair': 5353, ' Adm support including clerical': 7251, ' Armed Forces': 15, ' Executive admin and managerial': 6287}  
10  
{' Other': 1902, ' White': 83612, ' Asian or Pacific Islander': 2912, ' Amer Indian Aleut or Eskimo': 1208, ' Black': 10123}  
11  
{' Cuban': 615, ' Chicano': 169, ' Other Spanish': 1242, ' Mexican-American': 4000, ' NA': 404, ' Do not know': 139, ' Mexican (Mexicano)': 3694, ' Central or South American': 1986, ' Puerto Rican': 1581, ' All other': 85922}  
12  
{' Male': 47970, ' Female': 51790}  
13  
{' Not in universe': 90282, ' Yes': 1444, ' No': 8033}  
14  
{' Job loser - on layoff': 507, ' Other job loser': 1107, ' Job leaver': 285, ' Not in universe': 96631, ' New entrant': 203, ' Re-entrant': 1023}  
15

{' PT for econ reasons usually FT': 257, ' Children or Armed Forces': 61689, ' Not in labor force': 13412, ' Full-time schedules': 20466, ' PT for econ reasons usually PT': 582, ' PT for non-econ reasons usually FT': 1734, ' Unemployed part- time': 410, ' Unemployed full-time': 1204}  
19  
{' Nonfiler': 37367, ' Joint one under 65 & one 65+': 1976, ' Joint both 65+': 4193, ' Single': 18893, ' Joint both under 65': 33571, ' Head of household': 3756}  
20  
{' Northeast': 1321, ' West': 2045, ' Not in universe': 92067, ' Midwest': 1697, ' Abroad': 202, ' South': 2424}  
21  
{' Connecticut': 64, ' South Dakota': 74, ' Iowa': 64, ' Wisconsin': 61, ' New Jersey': 37, ' Kansas': 69, ' Michigan': 198, ' Arizona': 136, ' Utah': 533, ' North Carolina': 397, ' District of Columbia': 55, ' California': 883, ' Massachusetts': 72, ' Montana': 106, ' Florida': 449, ' Abroad': 264, ' Alaska': 123, ' Vermont': 98, ' Kentucky': 113, ' Wyoming': 102, ' Nebraska': 101, ' New Mexico': 227, ' Alabama': 107, ' Missouri': 69, ' Ohio': 99, ' Virginia': 57, ' Colorado': 113, ' Illinois': 84, ' Tennessee': 94, ' Not in universe': 92397, ' Georgia': 110, ' West Virginia': 102, ' New Hampshire': 115, ' South Carolina': 48, ' Nevada': 79, ' Oregon': 109, ' Maine': 65, ' North Dakota': 225, ' Mississippi': 77, ' Delaware': 47, ' Oklahoma': 295, ' New York': 106, ' Minnesota': 292, ' Idaho': 13, ' Arkansas': 99, ' Indiana': 296, ' Pennsylvania': 98, ' Maryland': 78, ' Texas': 94, ' Louisiana': 118}  
22  
{' Child under 18 of RP of unrel subfamily': 352, ' Spouse of householder': 20654, ' Other Rel <18 never marr not in subfamily': 302, ' Grandchild 18+ spouse of subfamily RP': 3, ' Child 18+ ever marr RP of subfamily': 353, ' Child 18+ never marr RP of subfamily': 308, ' Spouse of RP of unrelated subfamily': 23, ' Secondary individual': 3155, ' Other Rel <18 never married RP of subfamily': 1, ' Child <18 ever marr not in subfamily': 14, ' Other Rel <18 ever marr not in subfamily': 1, ' Other Rel 18+ ever marr not in subfamily': 1020, ' Grandchild <18 never marr child of subfamily RP': 888, ' Other Rel <18 never marr child of subfamily RP': 302, ' Child <18 spouse of subfamily RP': 0, ' Grandchild 18+ ever marr not in subfamily': 19, ' Grandchild <18 never marr not in subfamily': 501, ' Householder': 26690, ' Grandchild 18+ never marr not in subfamily': 206, ' Child 18+ spouse of subfamily RP': 76, ' Child 18+ never marr Not in a subfamily': 6133, ' Other Rel 18+ ever marr RP of subfamily': 309, ' Grandchild 18+ never marr RP of subfamily': 0, ' Child <18 never marr not in subfamily': 24989, ' Other Rel 18+ never marr RP of subfamily': 62, ' Grandchild <18 never marr RP of subfamily': 0, ' Other Rel 18+ spouse of subfamily RP': 353, ' Child <18 never marr RP of subfamily': 38, ' Child 18+ ever marr Not in a subfamily': 464, ' Grandchild 18+ ever marr RP of subfamily': 5, ' Other Rel <18 spouse of subfamily RP': 1, ' In group quarters': 82, ' RP of unrelated subfamily': 339, ' Child <18 ever marr RP of subfamily': 5, ' Other Rel 18+ never marr not in subfamily': 833, ' Nonfamily householder': 11242, ' Other Rel <18 ever marr RP of subfamily': 2}  
23  
{' Child under 18 ever married': 21, ' Child under 18 never married': 25038, ' Spouse of householder': 20656, ' Child 18 or older': 7338, ' Nonrelative of householder': 3876, ' Group Quarters- Secondary individual': 54, ' Householder': 37942, ' Other relative of householder': 4829}  
25  
{' Abroad to nonMSA': 35, ' Not in universe': 722, ' MSA to MSA': 5288, ' MSA to nonMSA': 363, ' Nonmover': 91344, ' NonMSA to MSA': 294, ' NonMSA to nonMSA': 1343, ' Not identifiable': 199, ' Abroad to MSA': 165}  
26  
{' Different division same region': 225, ' Same county': 4878, ' Not in universe': 722, ' Different region': 582, ' Nonmover': 91344, ' Different county same state': 1328, ' Abroad': 202, ' Different state same division': 473}

27  
{' Different state in South': 474, ' Same county': 4878, ' Not in universe': 722, ' Different state in Northeast': 207, ' Different state in Midwest': 241, ' Nonmover': 91344, ' Different county same state': 1328, ' Different state in West': 357, ' Abroad': 202}  
28  
{' Not in universe under 1 year old': 50668, ' Yes': 41398, ' No': 7693}  
29  
{' Not in universe': 92067, ' Yes': 2882, ' No': 4810}  
31  
{' Not in universe': 72393, ' Mother only present': 6256, ' Neither parent present': 807, ' Both parents present': 19358, ' Father only present': 943}  
32  
{' Columbia': 289, ' Cuba': 587, ' Trinadad&Tobago': 68, ' El-Salvador': 501, ' Guatemala': 231, ' Outlying-U S (Guam USVI etc)': 75, ' Holand-Netherlands': 20, ' United-States': 82796, ' China': 395, ' Thailand': 54, ' Haiti': 156, ' Germany': 647, ' South Korea': 255, ' Iran': 102, ' Hungary': 167, ' Dominican-Republic': 680, ' Poland': 628, ' Philippines': 590, ' Puerto-Rico': 1287, ' Hong Kong': 50, ' Vietnam': 217, ' Honduras': 105, ' Mexico': 5047, ' Portugal': 196, ' England': 380, ' Jamaica': 250, ' India': 306, ' Yugoslavia': 122, ' Greece': 205, ' Japan': 208, ' Taiwan': 83, ' Nicaragua': 176, ' Canada': 659, ' Italy': 1118, ' Panama': 15, ' Scotland': 118, ' France': 84, ' Cambodia': 116, ' Ecuador': 193, ' Laos': 81, ' Peru': 157, ' Ireland': 306}  
33  
{' Columbia': 288, ' Cuba': 595, ' Trinadad&Tobago': 57, ' El-Salvador': 533, ' Guatemala': 229, ' Outlying-U S (Guam USVI etc)': 62, ' Holand-Netherlands': 19, ' United-States': 83084, ' China': 367, ' Thailand': 70, ' Haiti': 161, ' Germany': 675, ' South Korea': 296, ' Iran': 91, ' Greece': 154, ' Dominican- Republic': 537, ' Poland': 584, ' Philippines': 644, ' Puerto-Rico': 1192, ' Hong Kong': 48, ' Vietnam': 236, ' Honduras': 112, ' Mexico': 4986, ' Portugal': 179, ' England': 426, ' Jamaica': 250, ' India': 312, ' Yugoslavia': 101, ' Hungary': 170, ' Japan': 264, ' Taiwan': 103, ' Nicaragua': 184, ' Canada': 701, ' Italy': 916, ' Panama': 17, ' Scotland': 114, ' France': 88, ' Cambodia': 100, ' Ecuador': 200, ' Laos': 72, ' Peru': 166, ' Ireland': 337}  
34  
{' Columbia': 210, ' Cuba': 427, ' Trinadad&Tobago': 47, ' El-Salvador': 342, ' Guatemala': 167, ' Outlying-U S (Guam USVI etc)': 45, ' Holand-Netherlands': 10, ' United-States': 90116, ' China': 236, ' Thailand': 64, ' Haiti': 98, ' Germany': 419, ' South Korea': 235, ' Iran': 65, ' Hungary': 29, ' Dominican- Republic': 326, ' Poland': 170, ' Philippines': 453, ' Puerto-Rico': 691, ' Nicaragua': 129, ' Vietnam': 181, ' Honduras': 77, ' Mexico': 2935, ' Portugal': 93, ' England': 224, ' Jamaica': 175, ' India': 217, ' Yugoslavia': 47, ' Greece': 73, ' Japan': 177, ' Taiwan': 81, ' Hong Kong': 51, ' Canada': 345, ' Italy': 225, ' Panama': 10, ' Scotland': 34, ' France': 58, ' Cambodia': 62, ' Ecuador': 133, ' Laos': 53, ' Peru': 120, ' Ireland': 70}  
35  
{' Native- Born in Puerto Rico or U S Outlying': 737, ' Native- Born abroad of American Parent(s)': 950, ' Foreign born- U S citizen by naturalization': 3011, ' Native- Born in the United States': 88354, ' Foreign born- Not a citizen of U S ': 6705}  
36  
{' 2': 8233, ' 0': 90187, ' 1': 1339}  
37  
{' Not in universe': 98734, ' Yes': 198, ' No': 827}  
38  
{' 2': 75296, ' 0': 23437, ' 1': 1026}  
40

{' 94': 49815, ' 95': 49945} Part 2:

Chose mean of data to replace missing Continuous data because replacing by ean will not change overall mean and will not impact much on standard deviation too . This will keep the gaussian probabilty of variables almost same before and after replacement.

Chose mode to replace missing Categorical attributes because missing attributes are most likely to have the value of mode of attribute as probability of mode of value is maximum for that particular attribute.

Part 3:  
Ran the 10-fold cross validation 30 times and then noted the average cross validation accuracy and its standard deviation.