"بسمه تعالی"

گزارش کار پروژه نهایی

"پیش بینی بیماری ها از روی داده های میکرو آرایه های ژنتیک"

درس : داده کاوی

استاد : جناب آقای دکتر احمدی

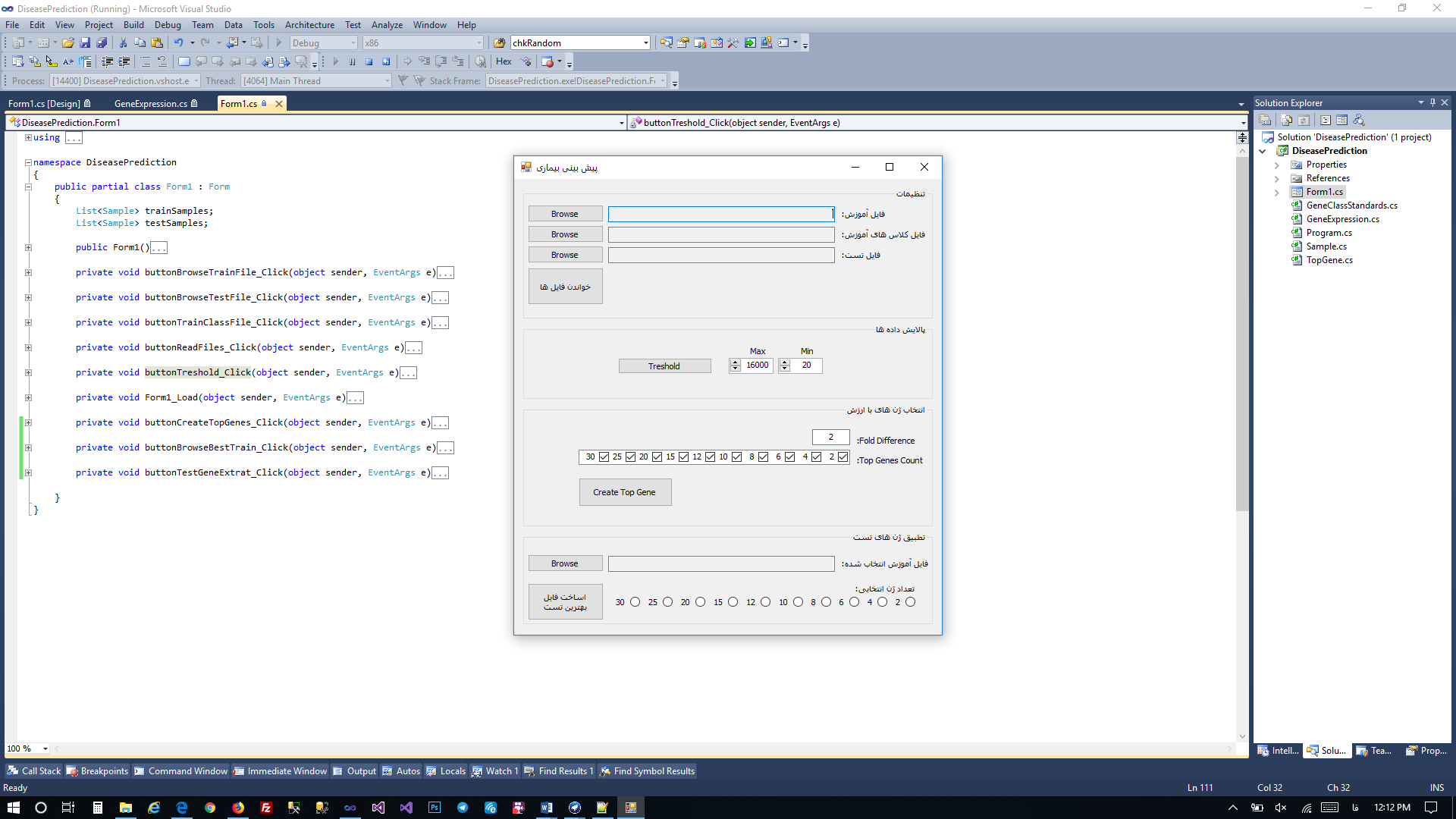
تهیه کننده : فرخنده زینالی آق قلعه

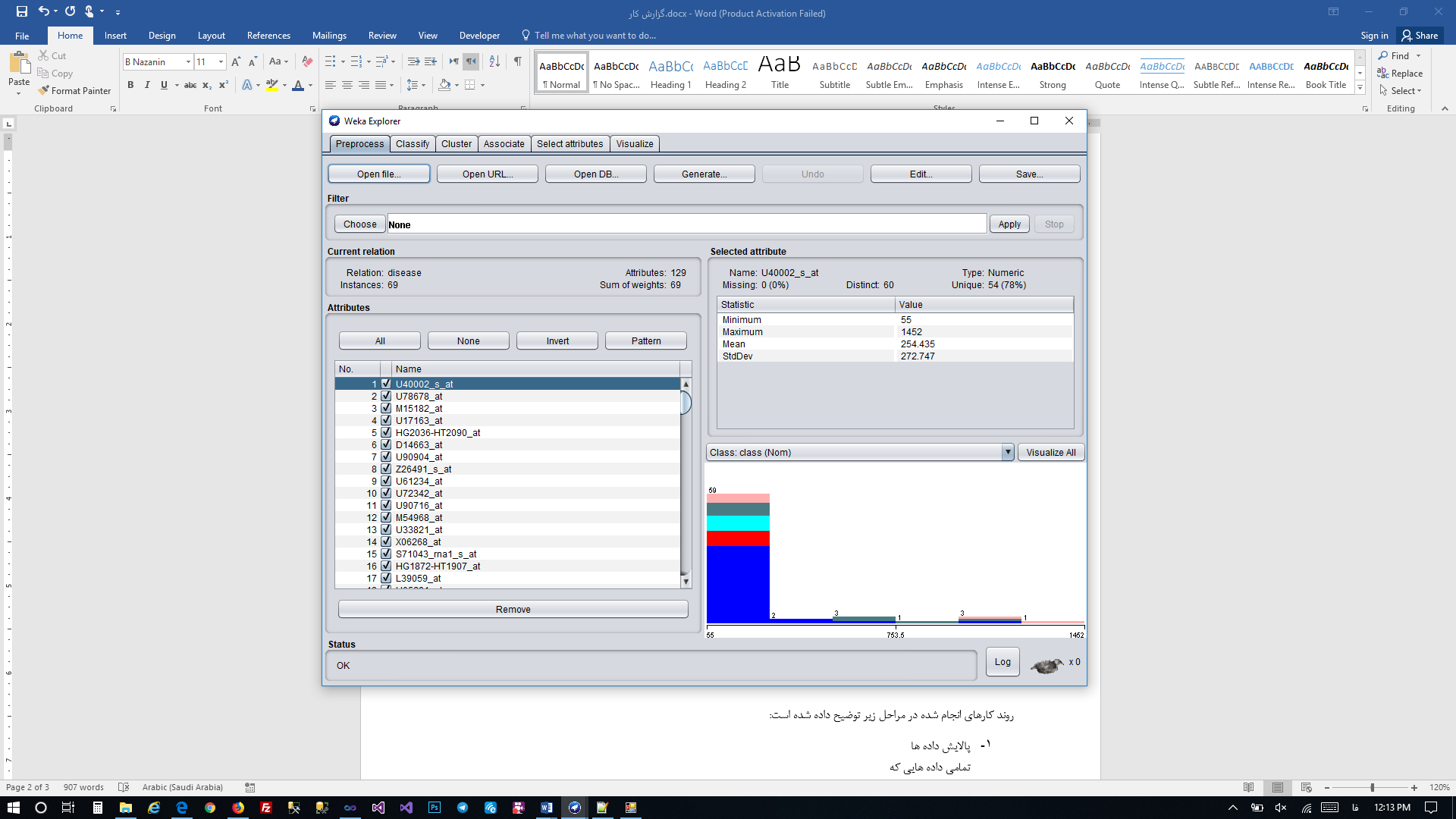
شماره دانشجویی : 9611274

بهار 98

در این پروژه قصد داریم از روی داده های میکرو آرایه های ژنتیک که برای آموزش در اختیار قرار گرفته و نوع بیماری نمونه های آن از قبل مشخص شده است استفاده شود تا نوع بيماري نمونه هاي ديگري از ميكرو آرايه هاي ژنتيك كه به عنوان فايل تست در اختيار قرار داده شده است با استفاده از يك كلاسيفاير منتخب پيش بيني گردد. تعداد نمونه هاي فايل آموزشي 69 و تعداد نمونه هاي فايل تست 23 مورد مي باشد.

براي انجام اين پروژه از دو برنامه استفاده شده است كه يكي براي محاسبات، اصلاح و ساخت فايل هاي مورد نياز در نرم افزار Weka كه يك برنامه تحت ويندوز با زبان برنامه نويسي C# شخصا پياده سازي شده است و ديگري خود نرم افزار Weka كه براي بكارگيري الگوريتم هاي دسته بندي و دريافت خروجي مورد استفاده قرار گرفته اند. تصوير اين دو نرم افزار را در زير مشاهده مي كنيد:





كد برنامه تحت ويندوز نوشته شده در فولدر موجود SourceCode مي باشد.

روند كارهاي انجام شده در مراحل زير توضيح داده شده است:

1. پالايش داده ها

اين مرحله توسط برنامه نوشته شده انجام مي شود. در ابتدا سه فايل نمونه هاي آموزشي، فايل كلاس هاي نمونه هاي آموزشي و فايل نمونه هاي تستي به برنامه معرفي مي شود. در ادامه با زدن دكمه خوتندن فايل ها برنامه داده ها را براي محاسبات پيش رو با فرمت خاص (جهت سهولت در انجام عمليات) در حافظه نگهداري مي كند (متد buttonReadFiles\_Click)

سپس با زدن دكمه Treshold در بخش پالايش داده ها ژن هاي نمونه آموزشي كه از مقادير ماكزيموم و مينيموم انتخابي (در اينجا مقادير 20 و 16000) حذف مي شوند و ديگر در محاسبات مورد استفاده قرار نمي گيرند. (متد buttonTreshold\_Click)

1. انتخاب ژن هاي منتخب براي هر كلاس

اين مرحله توسط برنامه نوشته شده انجام مي شود. در اين مرحله ابتدا بر اساس مقدار Fold Difference وارد شده (در اينجا مقدار 2) در نمونه هاي آموزشي ژن ها حذف مي شوند و در محاسبات بعدي مورد استفاده قرار نمي گيرند. سپس در هر كلاس براي ژن هاي باقي مانده مقدار T-Value محاسبه مي شود و براي هر نمونه آموزشي تعداد 2، 4، 6، 8، 10، 12، 15، 20، 25 و 30 ژن با بيشترين T-Value انتخاب شده و با هم تركيب مي شوند(در محاسبات از مقدار قدر مطلق مقادير T-Value استفاده شده است). سپس براي هر تعداد ژن هاي منتخب فايل هاي مربوطه شامل فايل هاي زير ساخته مي شود.

(متد buttonCreateTopGenes\_Click)

pp5i\_train.top2.gr.csv

pp5i\_train.top4.gr.csv

pp5i\_train.top6.gr.csv

pp5i\_train.top8.gr.csv

pp5i\_train.top10.gr.csv

pp5i\_train.top12.gr.csv

pp5i\_train.top15.gr.csv

pp5i\_train.top20.gr.csv

pp5i\_train.top25.gr.csv

pp5i\_train.top30.gr.csv

فرمت فايل هاي ژن هاي منتخب همانند فايل هاي آموزشي و تست با فرمت csv و به صورت ژن در سطر ساخته شده اند كه براي استفاده شدن در برنامه Weka مي بايست به فرمت arff با هدر فايل مخصوص شامل @Relation، @Attributes و @Data و به صورت ژن در ستون تبديل شوند. كه همزمان در برنامه اين كار نيز انجام مي شود و فايل هاي زير نيز ساخته مي شوند.

pp5i\_train.top2.gr.arff

pp5i\_train.top4.gr.arff

pp5i\_train.top6.gr.arff

pp5i\_train.top8.gr.arff

pp5i\_train.top10.gr.arff

pp5i\_train.top12.gr.arff

pp5i\_train.top15.gr.arff

pp5i\_train.top20.gr.arff

pp5i\_train.top25.gr.arff

pp5i\_train.top30.gr.arff

1. تعيين بهترين كلاسيفاير و بهترين تعداد ژن منتخب

اين مرحله در برنامه Weka انجام مي شود. پس از ساختن فايل هاي با فرمت arff در مرحله قبل مي توان اين فايل ها را در برنامه Weka مورد استفاده قرار داد. براي اين منظور به ازاي متدهاي درخواستي شامل NaiveBayes، J48، IB1، IBK=2,3,4 و همينطور يك الگوريتم به انتخاب شخصي كه در اينجا RandomForest مي باشد هر يك از فايل هاي ژن هاي منتخب را مورد ارزيابي قرار داده و ميزان دقت و خطاي هر يك را بدست مي آوريم. نتايج اين عمليات را در جدول زير مشاهده مي كنيد:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **NaiveBayes** | **J48** | **IB1** | **IBK2** | **IBK3** | **IBK4** | **RandomForest** |
| **pp5i\_train.top2** | **Correct : 69**  **Incorrect : 0**  **RMSE : 0** | Correct : 65  Incorrect : 4  RMSE : 0.1446 | Correct : 69  Incorrect : 0  RMSE : 0.027 | Correct : 66  Incorrect : 3  RMSE : 0.1211 | Correct : 67  Incorrect : 2  RMSE : 0.1174 | Correct : 65  Incorrect : 4  RMSE : 0.1394 | Correct : 69  Incorrect : 0  RMSE : 0.076 |
| **pp5i\_train.top4** | **Correct : 69**  **Incorrect : 0**  **RMSE : 0** | Correct : 67  Incorrect : 2  RMSE : 0.103 | Correct : 69  Incorrect : 0  RMSE : 0.027 | Correct : 69  Incorrect : 0  RMSE : 0.1016 | Correct : 65  Incorrect : 4  RMSE : 0.1346 | Correct : 67  Incorrect : 2  RMSE : 0.1444 | Correct : 69  Incorrect : 0  RMSE : 0.0743 |
| **pp5i\_train.top6** | **Correct : 69**  **Incorrect : 0**  **RMSE : 0** | Correct : 67  Incorrect : 2  RMSE : 0.103 | Correct : 69  Incorrect : 0  RMSE : 0.027 | Correct : 69  Incorrect : 0  RMSE : 0.0942 | Correct : 67  Incorrect : 2  RMSE : 0.1198 | Correct : 66  Incorrect : 3  RMSE : 0.1242 | Correct : 69  Incorrect : 0  RMSE : 0.0753 |
| **pp5i\_train.top8** | **Correct : 69**  **Incorrect : 0**  **RMSE : 0** | Correct : 67  Incorrect : 2  RMSE : 0.1031 | Correct : 69  Incorrect : 0  RMSE : 0.027 | Correct : 69  Incorrect : 0  RMSE : 0.0862 | Correct : 67  Incorrect : 2  RMSE : 0.1085 | Correct : 67  Incorrect : 2  RMSE : 0.1226 | Correct : 69  Incorrect : 0  RMSE : 0.0719 |
| **pp5i\_train.top10** | **Correct : 69**  **Incorrect : 0**  **RMSE : 0** | Correct : 67  Incorrect : 2  RMSE : 0.1031 | Correct : 69  Incorrect : 0  RMSE : 0.027 | Correct : 69  Incorrect : 0  RMSE : 0.0942 | Correct : 67  Incorrect : 2  RMSE : 0.1054 | Correct : 66  Incorrect : 3  RMSE : 0.1255 | Correct : 69  Incorrect : 0  RMSE : 0.0742 |
| **pp5i\_train.top12** | **Correct : 69**  **Incorrect : 0**  **RMSE : 0** | Correct : 68  Incorrect : 1  RMSE : 0.0726 | Correct : 69  Incorrect : 0  RMSE : 0.027 | Correct : 69  Incorrect : 0  RMSE : 0.0942 | Correct : 67  Incorrect : 2  RMSE : 0.0989 | Correct : 67  Incorrect : 2  RMSE : 0.1151 | Correct : 69  Incorrect : 0  RMSE : 0.0704 |
| **pp5i\_train.top15** | **Correct : 69**  **Incorrect : 0**  **RMSE : 0** | Correct : 68  Incorrect : 1  RMSE : 0.0726 | Correct : 69  Incorrect : 0  RMSE : 0.027 | Correct : 69  Incorrect : 0  RMSE : 0.0862 | Correct : 67  Incorrect : 2  RMSE : 0.0989 | Correct : 67  Incorrect : 2  RMSE : 0.1118 | Correct : 69  Incorrect : 0  RMSE : 0.0696 |
| **pp5i\_train.top20** | **Correct : 69**  **Incorrect : 0**  **RMSE : 0** | Correct : 68  Incorrect : 1  RMSE : 0.0726 | Correct : 69  Incorrect : 0  RMSE : 0.027 | Correct : 69  Incorrect : 0  RMSE : 0.0674 | Correct : 67  Incorrect : 2  RMSE : 0.0989 | Correct : 67  Incorrect : 2  RMSE : 0.1051 | Correct : 69  Incorrect : 0  RMSE : 0.0702 |
| **pp5i\_train.top25** | **Correct : 69**  **Incorrect : 0**  **RMSE : 0** | Correct : 68  Incorrect : 1  RMSE : 0.0726 | Correct : 69  Incorrect : 0  RMSE : 0.027 | Correct : 69  Incorrect : 0  RMSE : 0.0774 | Correct : 67  Incorrect : 2  RMSE : 0.092 | Correct : 66  Incorrect : 3  RMSE : 0.1179 | Correct : 69  Incorrect : 0  RMSE : 0.0728 |
| **pp5i\_train.top30** | **Correct : 69**  **Incorrect : 0**  **RMSE : 0** | **Correct : 69**  **Incorrect : 0**  **RMSE : 0** | Correct : 69  Incorrect : 0  RMSE : 0.027 | Correct : 69  Incorrect : 0  RMSE : 0.0674 | Correct : 67  Incorrect : 2  RMSE : 0.0956 | Correct : 67  Incorrect : 2  RMSE : 0.1051 | Correct : 69  Incorrect : 0  RMSE : 0.071 |

پس از بدست آوردن مقادير دقت و خطاي هر يك از الگوريتم ها روي فايل هاي ژن هاي منتخب شامل تعداد پيش بيني درست و غلط و همينطور مقدار خطاي Root Mean Square Error مشخص شد كه الگوريتم NiaveBayes براي كليه فايل هاي منتخب و الگوريتم J48 براي فايل منتخب 30 بهترين نتيجه را بدست مي دهند. لذا به دليل محاسبات كمتر فايل منتخب 2 كه در الگوريتم NaiveBayes نتيجه قابل قبولي را بدست داده است براي استفاده در پيش بيني نمونه هاي تستي انتخاب مي شود.

1. پيش بيني كلاس نمونه هاي تستي

در اين مرحله از هر دو برنامه نوشته شده و برنامه Weka استفاده می شود. در ابتدا از برنامه نوشته شده در بخش تطبیق ژن های تست فایل ژن های متخب و همین طور تعداد آن را انتخاب نموده و دکمه ساخت فایل بهترین تست را می زنیم (متد buttonTestGeneExtrat\_Click). در این عملیات بر اساس ژن های موجود در فایل منتخب، ژن های فایل نمونه های تستی نیز پالایش می شود یعنی تنها ژن هایی که در فایل منتخب موجود هستند باقی مانده و بقیه حذف می شوند كه فايل pp5i\_test.best2.csv بدست مي آيد و در ادامه برای اینکه بتوان اين اطلاعات را در برنامه Weka مورد استفاده قرار داد این فایل به فرمت arff تبدیل می شود و فايل pp5i\_test.best2.arff بدست مي آيد.

در مرحله بعدي فايل ژن هاي منتخب به عنوان فايل پيش پردازش معرفي شده و فايل بدست آمده در بخش قبلي براي نمونه هاي تستي نيز به عنوان فايل تست در برنامه Weka جهت تعيين كلاس نمونه هاي تستي مورد استفاده قرار مي گيرد. نتيجه اين عمليات در فايل FarkhondehZeinali\_Predicted.arff ذخيره شده است كه نتيجه كلاس هاي پيش بيني شده براي نمونه هاي تستي را در جدول زير مشاهده مي كنيد.

|  |  |
| --- | --- |
| **Sample** | **Class** |
| 101 | MGL |
| 102 | EPD |
| 103 | MED |
| 104 | MED |
| 105 | EPD |
| 106 | MED |
| 107 | MED |
| 108 | MED |
| 109 | EPD |
| 110 | JPA |
| 111 | JPA |
| 112 | MED |
| 113 | MED |
| 114 | MED |
| 115 | MED |
| 116 | MED |
| 117 | MED |
| 118 | MED |
| 119 | MED |
| 120 | MED |
| 121 | RHB |
| 122 | MED |
| 123 | RHB |