

### ارزیابی توان داده‌های ژنوتیپی در پیش‌بینی صفات کمی فلفل با روش‌های آماری و یادگیری ماشین

سید مجید حسینی \*

دانشجوی دکتری مهندسی کامپیوتر، دانشگاه فردوسی مشهد، ایران

رایانامه: hosseini.seyedmajid@mail.um.ac.ir

#### چکیده

پیش‌بینی عملکرد فنوتیپی بر پایه اطلاعات ژنوتیپی می‌تواند با کاهش وابستگی به ارزیابی‌های گسترده فنوتیپی، به افزایش کارایی انتخاب در برنامه‌های اصلاح نباتات کمک کند. پژوهش حاضر با هدف ارزیابی توان داده‌های ژنوتیپی در پیش‌بینی صفات کمی فلفل (*Capsicum annuum* L.) و مقایسه عملکرد روش‌های آماری و یادگیری ماشین انجام شد. برای این منظور، داده‌های ژنوتیپی ۳۵۰ نمونه فلفل که نماینده طیف گسترده‌ای از تنوع ژنتیکی این گونه بودند، مورد استفاده قرار گرفت. داده‌های ژنوتیپی شامل ۱۰۱۹۵ نشانگر ژنتیکی بود و برای هر یک از ۱۵ صفت کمی، مقادیر فنوتیپی تعدیل‌شده حاصل از تحلیل چندمحیطی به‌عنوان متغیر هدف در نظر گرفته شد. به‌منظور مقایسه توان پیش‌بینی روش‌های مختلف، مدل‌های PLS، Ridge، Elastic Net، GBLUP و Extra Trees ارزیابی شدند. عملکرد مدل‌ها با استفاده از اعتبارسنجی متقاطع پنج‌بخشی بررسی شد، به‌طوری‌که در هر مرحله بخشی از نمونه‌ها به‌طور کامل از فرایند آموزش کنار گذاشته شدند و تنها برای ارزیابی نهایی مدل مورد استفاده قرار گرفتند. همچنین، تنظیم پارامترهای هر مدل صرفاً بر اساس داده‌های آموزشی انجام شد تا از ورود اطلاعات بخش آزمون به فرایند آموزش جلوگیری شود. ضریب همبستگی پیرسون بین مقادیر مشاهده‌شده و پیش‌بینی‌شده به‌عنوان معیار اصلی دقت در نظر گرفته شد. نتایج نشان داد توان پیش‌بینی صفات بر اساس داده‌های ژنوتیپی به‌طور قابل‌توجهی به نوع صفت وابسته بود و بهترین ضریب همبستگی در میان صفات از حدود ۰.۴۶ تا ۰.۹۳ متغیر بود. بیشترین دقت پیش‌بینی برای حداکثر عرض میوه با ضریب همبستگی ۰.۹۲۷ مشاهده شد. در مجموع، GBLUP بالاترین میانگین دقت پیش‌بینی را نشان داد؛ با این حال، روش‌های PLS، Ridge و Extra Trees نیز در بسیاری از صفات عملکردی نزدیک به GBLUP و در برخی موارد عملکرد بهتری نشان دادند. این نتایج نشان می‌دهد داده‌های ژنوتیپی از توان مناسبی برای پیش‌بینی تعدادی از صفات کمی فلفل برخوردارند، اما افزایش پیچیدگی الگوریتم به‌طور سیستماتیک به افزایش دقت پیش‌بینی منجر نمی‌شود و انتخاب مدل مناسب می‌تواند به ویژگی صفت مورد بررسی وابسته باشد.

**کلمات کلیدی:** فلفل، داده‌های ژنوتیپی، پیش‌بینی ژنومی، صفات کمی، یادگیری ماشین، انتخاب ژنومی

#### ۱. مقدمه

اصلاح صفات کمی در گیاهان به دلیل ماهیت پیچیده ژنتیکی این صفات و تأثیر عوامل محیطی، نیازمند ارزیابی تعداد زیادی از مواد ژنتیکی در مراحل مختلف برنامه‌های اصلاحی است. ارزیابی فنوتیپی گسترده، به‌ویژه هنگامی که تعداد ژنوتیپ‌ها

و صفات مورد بررسی افزایش می‌یابد، می‌تواند زمان‌بر و پرهزینه باشد. در چنین شرایطی، پیش‌بینی ژنومی رویکردی برای استفاده هم‌زمان از اطلاعات نشانگرهای گسترده در سراسر ژنوم به‌منظور پیش‌بینی عملکرد مواد ژنتیکی فراهم می‌کند. با گسترش داده‌های ژنومی، علاوه بر روش‌های آماری متداول، استفاده از روش‌های یادگیری ماشین برای مدل‌سازی رابطه میان اطلاعات ژنوتیپی و صفات کمی نیز افزایش یافته است [۱]. مطالعات مقایسه‌ای اخیر نشان داده‌اند که عملکرد این روش‌ها می‌تواند به ویژگی داده‌ها و صفت مورد بررسی وابسته باشد و افزایش پیچیدگی مدل لزوماً با افزایش دقت پیش‌بینی همراه نیست [۱].

فلفل (*Capsicum annuum* L.) از محصولات مهم خانواده Solanaceae است و تنوع قابل توجهی در صفات مرتبط با ساختار گیاه، زمان گلدهی، عملکرد، شکل، اندازه و ویژگی‌های کیفی میوه دارد. در سال‌های اخیر، استفاده از اطلاعات ژنومی برای پیش‌بینی صفات فلفل مورد توجه قرار گرفته است. لوزادا و همکاران [۲] با بررسی ۲۰۴ ژنوتیپ فلفل و ۱۴۹۲۲ نشانگر SNP، عملکرد چند روش آماری و یادگیری ماشین را برای صفات زراعی، فنولوژیک و مرتبط با عملکرد مقایسه کردند. نتایج آن‌ها نشان داد هیچ مدلی در تمام صفات برتری نداشت و بهترین روش بسته به صفت متفاوت بود؛ به‌طوری‌که برخی روش‌های خطی و GBLUP در تعدادی از صفات با روش‌های پیچیده‌تر رقابت کردند.

کاربرد پیش‌بینی ژنومی در فلفل در مطالعات جدیدتر نیز توسعه یافته است. کوندو و همکاران [۳] در سال ۲۰۲۵ توان پیش‌بینی چند صفت میوه شامل طول، عرض، وزن، شاخص شکل و ضخامت پریکارپ را در نتاج F1 فلفل بررسی کردند و برای صفات مورد مطالعه همبستگی ۰٫۸۳۳ تا ۰٫۹۰۸ بین مقادیر مشاهده‌شده و پیش‌بینی‌شده گزارش کردند. همچنین احمد و همکاران [۴] در مطالعه‌ای دیگر در سال ۲۰۲۵، یازده مدل پیش‌بینی ژنومی را برای صفات مرتبط با تندی فلفل ارزیابی کردند و نشان دادند که دقت پیش‌بینی علاوه بر مدل مورد استفاده، به صفت نیز وابسته است؛ در این مطالعه یکی از انواع GBLUP عملکرد مناسبی در میان صفات مختلف نشان داد. این نتایج در مجموع بیانگر ظرفیت بالای داده‌های ژنومی برای پیش‌بینی صفات فلفل و در عین حال اهمیت ارزیابی مدل‌ها در طیفی از صفات با ویژگی‌های متفاوت است.

در مطالعه مک‌لنود و همکاران [۵]، مجموعه‌ای متنوع از ۳۵۰ نمونه *C. annuum* با استفاده از ۱۰۱۹۵ نشانگر ژنتیکی و داده‌های فنوتیپی حاصل از آزمایش‌های چندمحیطی مورد بررسی قرار گرفت. این مجموعه تنوع گسترده‌ای از صفات کمی را در بر می‌گیرد و بستر مناسبی برای بررسی رابطه میان اطلاعات ژنوتیپی و تغییرات فنوتیپی فراهم می‌کند. همچنین همان مطالعه به ظرفیت داده‌های حاصل برای توسعه رویکردهای پیش‌بینی ژنومی اشاره شده است. بر این اساس، پژوهش حاضر با هدف ارزیابی توان داده‌های ژنوتیپی در پیش‌بینی ۱۵ صفت کمی فلفل و مقایسه عملکرد روش‌های آماری و یادگیری ماشین انجام شد. در این پژوهش، توان مدل‌ها در پیش‌بینی نمونه‌هایی که در فرایند آموزش حضور نداشتند مورد ارزیابی قرار گرفت و به‌طور ویژه بررسی شد که آیا افزایش پیچیدگی روش‌های پیش‌بینی می‌تواند به‌صورت پایدار موجب افزایش دقت در صفات مختلف شود یا خیر.

## ۲. مواد و روش‌ها

داده‌های مورد استفاده در این پژوهش از مجموعه داده منتشرشده توسط مک‌لنود و همکاران [۵] تهیه شد. این داده‌ها شامل ۳۵۰ نمونه از گونه *Capsicum annuum* L. و ۱۰۱۹۵ نشانگر ژنتیکی بود. مجموعه مورد مطالعه بخش گسترده‌ای از تنوع ژنتیکی فلفل را پوشش می‌دهد و نمونه‌ها در آزمایش‌های مختلف فنوتیپی ارزیابی شده‌اند. در پژوهش حاضر، ۱۵ صفت کمی مرتبط با رشد گیاه، گلدهی، عملکرد و ویژگی‌های میوه بررسی شد. این صفات شامل طول محور گیاه (AL)، ارتفاع گیاه (PHe) و زمان گلدهی (FIT)؛ ویژگی‌های اندازه و شکل میوه شامل حداکثر طول میوه (FLe)، حداکثر عرض

میوه (FWi)، وزن میوه (FWe)، شاخص شکل میوه (FShI)، ضخامت پریکارپ (PTh) و تعداد حفره‌های داخلی میوه (LN)؛ صفات عملکردی شامل تعداد کل میوه (TFN) و وزن کل میوه (TFWe)؛ و ویژگی‌های کیفی شامل مقدار بریکس به‌عنوان شاخص مواد جامد محلول میوه (BX) و مؤلفه‌های رنگ خارجی میوه (FCa, FCL و FCB) بود. برای هر صفت، مقادیر فنوتیپی تعدیل‌شده حاصل از تحلیل چندمحیطی به‌عنوان متغیر هدف مدل‌های پیش‌بینی در نظر گرفته شد. در آماده‌سازی داده‌های ژنوتیپی، مقادیر گمشده هر نشانگر بر اساس اطلاعات مجموعه آموزشی جایگزین شدند. همچنین برای مدل‌هایی که به مقیاس متغیرها حساس بودند، استانداردسازی نشانگرها تنها با استفاده از پارامترهای محاسبه‌شده در داده‌های آموزشی انجام گرفت. در GBLUP\* نیز فراوانی آللی و ماتریس روابط ژنومی بر اساس نمونه‌های آموزشی محاسبه شد. به این ترتیب، اطلاعات نمونه‌های آزمون در مراحل آماده‌سازی داده‌های ژنوتیپی و برازش مدل وارد نشد.

برای مقایسه توان پیش‌بینی، پنج روش شامل Ridge, GBLUP, Elastic Net, PLS<sup>†</sup> و Extra Trees مورد ارزیابی قرار گرفتند. علاوه بر این، یک مدل ساده مبتنی بر میانگین داده‌های آموزشی به‌عنوان مدل مبنا در نظر گرفته شد تا عملکرد مدل‌های ژنومی با یک پیش‌بینی فاقد اطلاعات ژنوتیپی مقایسه شود. این مجموعه از مدل‌ها امکان مقایسه روش کلاسیک پیش‌بینی ژنومی، روش‌های خطی منظم‌شده، کاهش ابعاد و یک روش غیرخطی مبتنی بر درخت را فراهم کرد. عملکرد مدل‌ها با استفاده از اعتبارسنجی متقاطع تودرتو ارزیابی شد. در مرحله بیرونی، نمونه‌ها به پنج بخش تقسیم شدند و در هر مرحله چهار بخش برای آموزش و بخش باقی‌مانده برای آزمون استفاده شد؛ به‌طوری‌که هیچ نمونه‌ای به‌طور هم‌زمان در مجموعه آموزشی و آزمون حضور نداشت. تنظیم پارامترهای مدل‌ها نیز با اعتبارسنجی متقاطع سه‌بخشی و صرفاً درون داده‌های آموزشی انجام شد. پس از تکمیل پنج مرحله، پیش‌بینی‌های نمونه‌های آزمون جمع‌شدند و دقت مدل‌ها با استفاده از ضریب همبستگی پیرسون بین مقادیر مشاهده‌شده و پیش‌بینی‌شده به‌عنوان معیار اصلی ارزیابی شد.

### ۳. نتایج و بحث

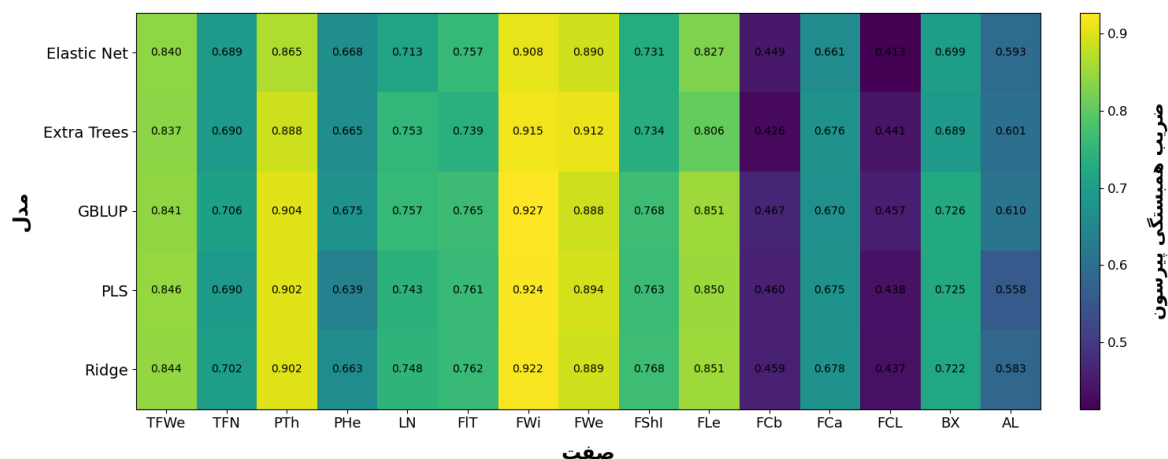
#### ۳-۱. مقایسه قابلیت پیش‌بینی صفات مختلف

نتایج نشان داد که توان پیش‌بینی بر اساس داده‌های ژنوتیپی به‌طور محسوسی بین صفات مختلف متفاوت بود. در میان مدل‌های مورد بررسی، بیشترین ضریب همبستگی برای هر صفت از ۰٫۴۵۷ تا ۰٫۹۲۷ متغیر بود. بالاترین دقت پیش‌بینی برای حداکثر عرض میوه (FWi) با ضریب همبستگی ۰٫۹۲۷ مشاهده شد. پس از آن، صفات وزن میوه (FWe) و ضخامت پریکارپ (PTh) به‌ترتیب با ضرایب همبستگی ۰٫۹۱۲ و ۰٫۹۰۴ از قابلیت پیش‌بینی بالایی برخوردار بودند. برای حداکثر طول میوه (FLe) نیز مقدار ۰٫۸۵۱ به دست آمد. در مقابل، کمترین دقت مربوط به مؤلفه‌های رنگ خارجی میوه FCL و FCB بود که بیشترین ضرایب همبستگی آن‌ها به‌ترتیب ۰٫۴۵۷ و ۰٫۴۶۷ بود.

این تفاوت نشان می‌دهد که میزان موفقیت پیش‌بینی ژنومی تنها تابع روش مورد استفاده نیست و ویژگی‌های هر صفت نیز نقش مهمی در دقت حاصل دارند. تفاوت در معماری ژنتیکی صفات، میزان تأثیر عوامل محیطی و کیفیت ارزیابی فنوتیپی می‌تواند از عوامل مؤثر بر این تفاوت باشد. به‌طور کلی، صفات مرتبط با اندازه و برخی ویژگی‌های میوه در مجموعه مورد بررسی قابلیت پیش‌بینی بیشتری نسبت به برخی صفات رنگی نشان دادند.

\* Genomic Best Linear Unbiased Prediction

† Partial Least Squares



شکل ۱- مقایسه ضریب همبستگی پیرسون بین مقادیر مشاهده شده و پیش‌بینی شده برای پنج مدل و ۱۵ صفت کمی فلفل.

### ۳-۲. مقایسه عملکرد مدل‌های پیش‌بینی

مقایسه میانگین عملکرد مدل‌ها در ۱۵ صفت نشان داد که GBLUP با میانگین ضریب همبستگی ۰٫۷۳۴ بالاترین دقت کلی را داشت. پس از آن Ridge و PLS به ترتیب با میانگین‌های ۰٫۷۲۹ و ۰٫۷۲۵ قرار گرفتند. میانگین ضریب همبستگی برای Elastic Net و Extra Trees نیز به ترتیب ۰٫۷۱۸ و ۰٫۷۱۴ بود. این اختلاف نسبتاً اندک نشان می‌دهد که چند روش مختلف توانسته‌اند بخش قابل توجهی از اطلاعات پیش‌بینی‌کننده موجود در داده‌های ژنوتیپی را استخراج کنند. این میانگین‌ها از مقادیر حاصل برای ۱۵ صفت در خروجی اعتبارسنجی محاسبه شدند. خلاصه عملکرد مدل‌های مورد بررسی بر اساس ضریب همبستگی پیرسون در جدول ۱ ارائه شده است.

جدول ۱- خلاصه عملکرد مدل‌های پیش‌بینی بر اساس ضریب همبستگی پیرسون

| مدل         | میانگین ضریب همبستگی | کمترین مقدار | بیشترین مقدار | تعداد بهترین عملکرد* |
|-------------|----------------------|--------------|---------------|----------------------|
| GBLUP       | ۰٫۷۳۴                | ۰٫۴۵۷        | ۰٫۹۲۷         | ۱۲                   |
| Ridge       | ۰٫۷۲۹                | ۰٫۴۳۷        | ۰٫۹۲۲         | ۳                    |
| PLS         | ۰٫۷۲۵                | ۰٫۴۳۸        | ۰٫۹۲۴         | ۱                    |
| Extra trees | ۰٫۷۱۸                | ۰٫۴۲۶        | ۰٫۹۱۵         | ۱                    |
| Elastic Net | ۰٫۷۱۴                | ۰٫۴۱۳        | ۰٫۹۰۸         | ۰                    |

\*تعداد صفاتی که مدل در آن‌ها بالاترین ضریب همبستگی را به دست آورده است و موارد تساوی نیز برای هر دو مدل محاسبه شده‌اند.

با وجود برتری کلی GBLUP، بهترین مدل برای همه صفات یکسان نبود. Ridge برای Fca بالاترین مقدار را به دست آورد و در صفات FShI و FLe عملکرد آن با GBLUP برابر بود. Extra Trees در پیش‌بینی وزن میوه (FWe)

با ضریب همبستگی ۰٫۹۱۲ بهترین عملکرد را نشان داد، در حالی که PLS برای وزن کل میوه (TFWe) با مقدار ۰٫۸۴۶ بالاترین دقت را داشت. بنابراین، اگرچه GBLUP در مجموع پایدارترین عملکرد را نشان داد، هیچ روش واحدی برای تمام صفات بهترین نبود.

### ۳-۳. بررسی اثر پیچیدگی مدل بر دقت پیش‌بینی

در یکی از یافته‌های مهم این پژوهش آن بود که افزایش پیچیدگی مدل به‌طور سیستماتیک با افزایش دقت پیش‌بینی همراه نبود. Extra Trees به‌عنوان روش غیرخطی مبتنی بر مجموعه درخت‌ها، تنها در برخی صفات مانند وزن میوه برتری نشان داد و میانگین عملکرد آن از GBLUP، Ridge و PLS کمتر بود. از سوی دیگر، GBLUP و Ridge که ساختار ساده‌تری دارند، در بخش عمده‌ای از صفات عملکردی مشابه یا بهتر از مدل‌های پیچیده‌تر ارائه کردند. فاصله میانگین ضریب همبستگی GBLUP و Ridge تنها حدود ۰٫۰۰۵ و فاصله GBLUP با Extra Trees حدود ۰٫۰۱۶ بود. بنابراین تفاوت میان الگوریتم‌ها در مقایسه با اختلاف مشاهده‌شده میان صفات نسبتاً محدود بود. این نتیجه با گزارش‌های پیشین در فلفل که بر وابستگی عملکرد مدل به نوع صفت و نبود برتری یک روش در تمام صفات تأکید کرده‌اند، همسو است [۲]. در مجموع، نتایج نشان می‌دهد انتخاب مدل در پیش‌بینی ژنومی بهتر است بر اساس ویژگی صفت و ارزیابی تجربی عملکرد صورت گیرد و استفاده از یک الگوریتم پیچیده‌تر به‌تنهایی تضمین‌کننده افزایش دقت نیست.

### ۴. نتیجه‌گیری

نتایج این پژوهش نشان داد که داده‌های ژنوتیپی می‌توانند برای پیش‌بینی بسیاری از صفات کمی فلفل با دقت قابل‌قبولی مورد استفاده قرار گیرند، هرچند میزان دقت پیش‌بینی به‌طور محسوسی بین صفات مختلف متفاوت بود. در میان مدل‌های بررسی‌شده، GBLUP در مجموع بالاترین میانگین دقت پیش‌بینی را به دست آورد، اما روش‌های Ridge، PLS و Extra Trees نیز در بسیاری از صفات عملکردی نزدیک و در برخی موارد بهتر نشان دادند. این یافته‌ها نشان می‌دهد که افزایش پیچیدگی الگوریتم لزوماً به بهبود سیستماتیک دقت پیش‌بینی منجر نمی‌شود و انتخاب مدل مناسب باید با توجه به ویژگی صفت و عملکرد تجربی مدل‌ها انجام گیرد. به‌طور کلی، نتایج حاصل بیانگر ظرفیت مناسب پیش‌بینی ژنومی برای استفاده از اطلاعات ژنوتیپی در ارزیابی و انتخاب مواد ژنتیکی فلفل و حمایت از برنامه‌های اصلاحی این محصول است.

### ۵. مراجع

1. Lourenço, V. M., Ogutu, J. O., Rodrigues, R. A. P., et al. (2024). "Genomic prediction using machine learning: a comparison of the performance of regularized regression, ensemble, instance-based and deep learning methods on synthetic and empirical data." *BMC Genomics*, 25, 152.
2. Lozada, D. N., Sandhu, K. S. and Bhatta, M. (2023). "Ridge regression and deep learning models for genome-wide selection of complex traits in New Mexican Chile peppers." *BMC Genomic Data*, 24, 80.
3. Kondo, F., Kumanomido, Y., D'Andrea, M., et al. (2025). "Phenotypic simulation for fruit-related traits in F1 progenies of chili peppers (*Capsicum annuum*) using genomic

prediction based solely on parental information.” *Molecular Genetics and Genomics*, 300, 15.

4. Ahmed, N., Matsushima, K., Kumanomido, Y., et al. (2025). “Inheritance characteristics and potential of genomic prediction for pungency levels in F1 progeny of chili pepper (*Capsicum annuum*).” *Breeding Science*, 75(4), 303–314.

5. McLeod, L., Barchi, L., Tumino, G., et al. (2023). “Multi-environment association study highlights candidate genes for robust agronomic quantitative trait loci in a novel worldwide *Capsicum* core collection.” *The Plant Journal*, 116, 1508–1528.