**דו'’ח בנושא Missing Values**

**מטרה**

בחינת דרכים להתמודדות עם Missing Values במידע.

**תקציר**

בניסוי נבדקו3 דרכים שונות להתמודדות עם דאטה חסר – מחיקת שורות, השלמה בעזרת median, השלמה בעזרת median עבור אותו הקלאס. התקבל כי השלמה בעזרת ה-median חזקה בהרבה מאשר מחיקת שורות, שכן יש איבוד מידע שימושי לאימון באופן זה. השלמה עבור קלאס לא שיפרה רבות את התוצאה.

**רקע תיאורטי**

בעיית הערכים החסרים בדאטה הינה נפוצה מאוד. נדרש תחילה להבין מדוע הערכים החסרים – האם בגלל טעות של ה-sensor (טעות הסתברותית)? אם כן נוכל להתמודד במס' דרכים – למשל הכנסת ערכי הממוצע, החציון ועוד. שיטה נפוצה נוספת היא המילוי על סמך שכנים קרובים. אם הטעות אינה נובעת מטעות הסתברותית אלא מגמתית (מי שמדרג סרט למשל אלו אנשים שבד'’כ או מאוד אהבו או שנאו אותו), ייכנס לנו bias למודל ולא יהיה אפשר לתקנו.

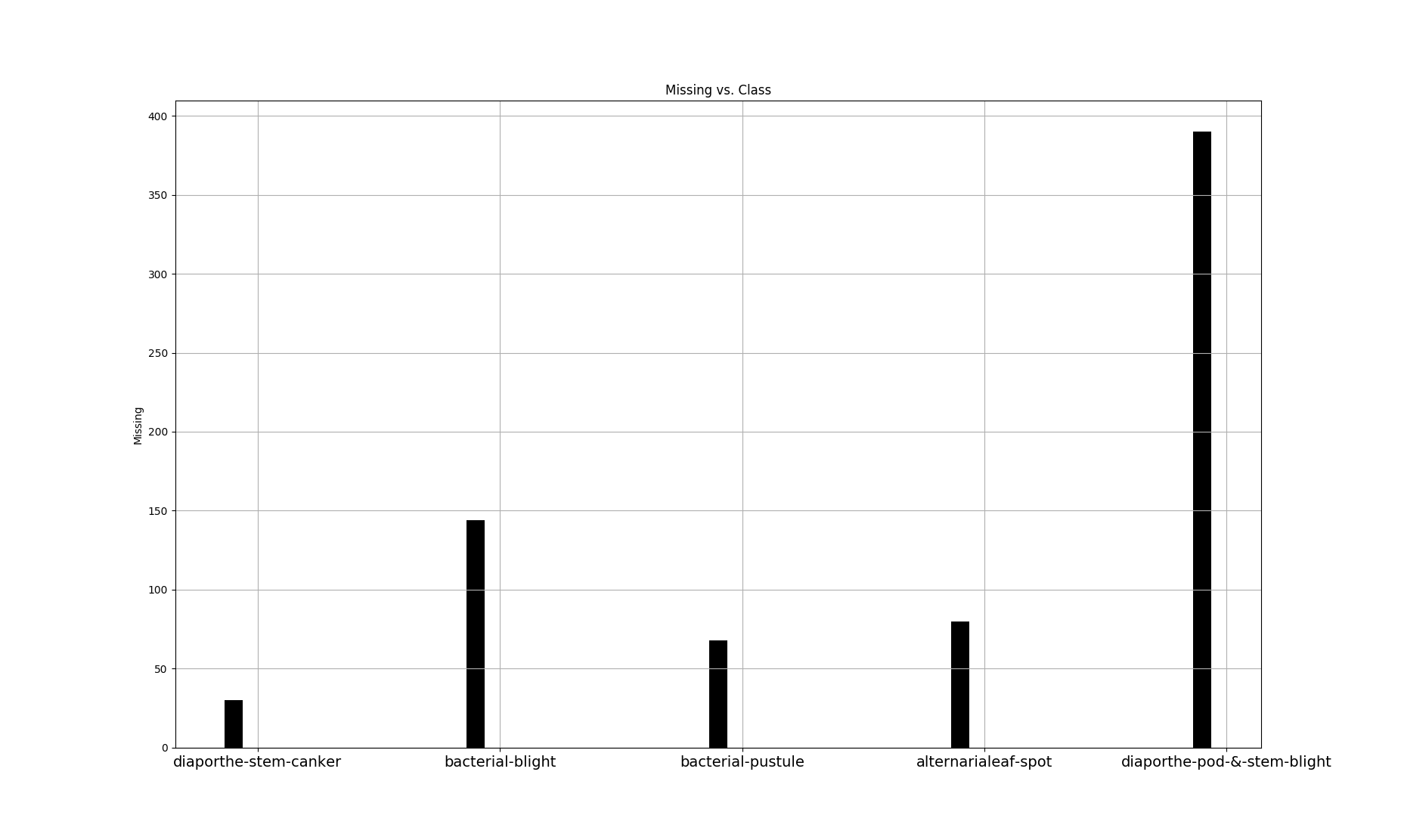
בדו'’ח זה נבחן טעות הסתברותית ולא מגמתית – דאטה חסר בדגימות מסוימות.

**מהלך הניסוי**

הדאטה שנבחן בניסוי הוא Soybean large data, עבורו צריך לסווגsoybean לקלאסים על סמך 35 פיצ'רים. בדאטה זה יש מס' לא מועט של ערכים חסרים.

השיטה הראשונה (והגרועה ביותר) להתמודדות איתם היא מחיקת השורות הרלוונטיות מה-train. לאחר מכן נבדקה אופציה מתקדמת יותר של imputation של הערכים החסרים בעזרת ערכי ה-median עבור כל משתנה.

נדרש לשפר את המודל אף יותר על נבדק הקשר בין הערכים החסרים לקלאס:



אפשר לראות כי ישנם 5 קלאסים מרכזיים עבורם חסרים הערכים הנ'’ל. לכן נפעל בדרך חכמה יותר – נבצע imputation עבור כל משתנה חסר **באותו הקלאס**.

**תוצאות**

האופציה הראשונה הייתה מחיקת שורות בהן ערכים חסרים. לאחר סיווג באמצעות מודל ה-Random Forest נמצא כי Accuracy = 0.843 – תוצאה נמוכה יחסית.

האופציה השנייה שיפרה את התוצאה רבות - עבורה התקבל Accuracy=0.939.

האופציה האחרונה לא הוסיפה רבות - לאחר התאמת מודל דומה התקבלAccuracy=0.940.

**דיון ומסקנות**

בניסוי נבחנו מס' דרכים בסיסיות שונות להתמודדות עם דאטה חסר – ונבחן האם בחינת התפלגות הדאטה הספציפי הזה יכולה לתרום למילוי ערך ה-Missing values בערך חכם יותר שיתרום לשיפור רב במודל. במקרה זה היו שתי מחלקות בהן כלל לא היו ערכים למשתנים שונים – ולכן impute לפי median באותה המחלקה לא עבד, ונדרש היה למלא ערכים אלו ב-median כללי. ייתכן ובחינת הערכים החסרים לפי מחלקות אינה רלוונטים פה – ועדיף להשתמש בתובנה אחרת אודות התפלגות המידע על מנת למלא את הערכים החסרים.